

Encyclopedia of Statistics in

BEHAVIORAL SCIENCE

Editors in Chief **Brian Everitt** **David Howell**

VOLUME

1

A-EXT

VOLUME 1

A Priori v Post Hoc Testing.	1-5	Ansari-Bradley Test.	93-94
ACE Model.	5-10	Arbuthnot, John.	94-95
Adaptive Random Assignment.	10-13	Area Sampling.	95-96
Adaptive Sampling.	13-16	Arithmetic Mean.	96-97
Additive Constant Problem.	16-18	Ascertainment Corrections.	97-99
Additive Genetic Variance.	18-22	Assortative Mating.	100-102
Additive Models.	22-24	Asymptotic Relative Efficiency.	102
Additive Tree.	24-25	Attitude Scaling.	102-110
Additivity Tests.	25-29	Attrition.	110-111
Adoption Studies.	29-33	Average Deviation.	111-112
Age-Period-Cohort Analysis.	33-38	Axes in Multivariate Analysis.	112-114
Akaike's Criterion.	38-39	Bagging.	115-117
Allelic Association.	40-43	Balanced Incomplete Block Designs.	118-125
All-X Models.	43-44	Bar Chart.	125-126
All-Y Models.	44	Battery Reduction.	126-129
Alternating Treatment Designs.	44-46	Bayes, Thomas.	129-130
Analysis of Covariance.	46-49	Bayesian Belief Networks.	130-134
Analysis of Covariance: Nonparametric.	50-52	Bayesian Item Response Theory Estimation.	134-139
Analysis of Variance.	52-56	Bayesian Methods for Categorical Data.	139-146
Analysis of Variance: Cell Means Approach.	56-66	Bayesian Statistics.	146-150
Analysis of Variance: Classification.	66-83	Bernoulli Family.	150-153
Analysis of Variance and Multiple Regression Approaches.	83-93	Binomial Confidence Interval.	153-154

Binomial Distribution: Estimating and Testing parameters.	155-157	Catastrophe Theory.	234-239
Binomial Effect Size Display.	157-158	Categorizing Data.	239-242
Binomial Test.	158-163	Cattell, Raymond Bernard.	242-243
Biplot.	163-164	Censored Observations.	243-244
Block Random Assignment.	165-167	Census.	245-247
Boosting.	168-169	Centering in Multivariate Linear Models.	247-249
Bootstrap Inference.	169-176	Central Limit Theory.	249-255
Box Plots.	176-178	Children of Twins Design.	256-258
Bradley-Terry Model.	178-184	Chi-Square Decomposition.	258-262
Breslow-Day Statistic.	184-186	Cholesky Decomposition.	262-263
Brown, William.	186-187	Classical Statistical Inference Extended: Split-Tailed Tests.	263-268
Bubble Plot.	187	Classical Statistical Inference: Practice versus Presentation.	268-278
Burt, Cyril Lodowic.	187-189	Classical Test Models.	278-282
Bush, Robert R.	189-190	Classical Test Score Equating.	282-287
Calculating Covariance.	191	Classification and Regression Trees.	287-290
Campbell, Donald T.	191-192	Clinical Psychology.	290-301
Canonical Correlation Analysis.	192-196	Clinical Trials and Intervention Studies.	301-305
Carroll-Arabie Taxonomy.	196-197	Cluster Analysis: Overview.	305-315
Carryover and Sequence Effects.	197-201	Clustered Data.	315
Case Studies.	201-204	Cochran, William Gemmell.	315-316
Case-Cohort Studies.	204-206	Cochran's C Test.	316-317
Case-Control Studies.	206-207	Coefficient of Variation.	317-318
Catalogue of Parametric Tests.	207-227	Cohen, Jacob.	318-319
Catalogue of Probability Density Functions.	228-234		

Cohort Sequential Design.	319-322	Confounding Variable.	391-392
Cohort Studies.	322-326	Contingency Tables.	393-397
Coincidences.	326-327	Coombs, Clyde Hamilton.	397-398
Collinearity.	327-328	Correlation.	398-400
Combinatorics for Categorical Variables.	328-330	Correlation and Covariance Matrices.	400-402
Common Pathway Model.	330-331	Correlation Issues in Genetics Research.	402-403
Community Intervention Studies.	331-333	Correlation Studies.	403-404
Comorbidity.	333-337	Correspondence Analysis.	404-415
Compensatory Equalization.	337-338	Co-Twin Control Methods.	415-418
Compensatory Rivalry.	338-339	Counter Null Value of an Effect Size.	422-423
Completely Randomized Design.	340-341	Counterbalancing.	418-420
Computational Models.	341-343	Counterfactual Reasoning.	420-422
Computer-Adaptive Testing.	343-350	Covariance.	423-424
Computer-Based Test Designs.	350-354	Covariance Matrices: Testing Equality of.	424-426
Computer-Based Testing.	354-359	Covariance Structure Models.	426-430
Concordance Rates.	359	Covariance/variance/correlation.	431-432
Conditional Independence.	359-361	Cox, Gertrude Mary.	432-433
Conditional Standard Errors of Measurement.	361-366	Cramer-Von Mises Test.	434
Confidence Intervals.	366-375	Criterion-Referenced Assessment.	435-440
Confidence Intervals: Nonparametric.	375-381	Critical Region.	440-441
Configural Frequency Analysis.	381-388	Cross Sectional Design.	453-454
Confounding in the Analysis of Variance.	389-391	Cross-Classified and Multiple Membership Models.	441-450

Cross-Lagged Panel Design.	450-451	Dropouts in Longitudinal Data.	515-518
Crossover Design.	451-452	Dropouts in Longitudinal Studies: Methods of Analysis.	518-522
Cross-validation.	454-457	Dummy Variables.	522-523
Cultural Transmission.	457-458		
Data Mining.	461-465		
de Finetti, Bruno.	465-466		
de Moivre, Abraham.	466		
Decision Making Strategies.	466-471		
Deductive Reasoning and Statistical Inference.	472-475		
DeFries-Fulker Analysis.	475-477		
Demand Characteristics.	477-478		
Deming, Edwards William.	478-479		
Design Effects.	479-483		
Development of Statistical Theory in the 20th Century.	483-485		
Differential Item Functioning.	485-490		
Direct and Indirect Effects.	490-492		
Direct Maximum Likelihood Estimation.	492-494		
Directed Alternatives in Testing.	495-496		
Direction of Causation Models.	496-499		
Discriminant Analysis.	499-505		
Distribution Free Inference, an Overview.	506-513		
Dominance.	513-514		
Dot chart.	514-515		

A Priori v Post Hoc Testing

VANCE W. BERGER

Volume 1, pp. 1–5

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

A Priori v Post Hoc Testing

Macdonald [11] points out some of the problems with post hoc analyses, and offers as an example the P value one would ascribe to drawing a particular card from a standard deck of 52 playing cards. If the null hypothesis is that all 52 cards have the same chance ($1/52$) to be selected, and the alternative hypothesis is that the ace of spades will be selected with probability one, then observing the ace of spades would yield a P value of $1/52$. For a Bayesian perspective (see **Bayesian Statistics**) on a similar situation involving the order in which songs are played on a CD, see Sections 4.2 and 4.4 of [13]. Now then, with either cards or songs on a CD, if no alternative hypothesis is specified, then there is the problem of inherent multiplicity. Consider that regardless of what card is selected, or what song is played first, one could call it the target (alternative hypothesis) after-the-fact (post hoc), and then draw the proverbial bull's eye around it, quoting a P value of $1/52$ (or $1/12$ if there are 12 songs on the CD). We would have, then, a guarantee of a low P value (at least in the case of cards, or more so for a lottery), thereby violating the probabilistic interpretation that under the null hypothesis a P value should, in the continuous case, have a uniform distribution on the unit interval $[0,1]$. In any case, the P value should be less than any number k in the unit interval $[0,1]$, with probability no greater than k [8].

The same problem occurs when somebody finds that a given baseball team always wins on Tuesdays when they have a left-handed starting pitcher. What is the probability of such an occurrence? This question cannot even be properly formulated, let alone answered, without first specifying an appropriate probability model within which to embed this event [6]. Again, we have inherent multiplicity. How many other outcomes should we take to be as statistically significant as or more statistically significant than this one? To compute a valid P value, we need the null probability of all of these outcomes in the extreme region, and so we need both an enumeration of all of these outcomes and their ranking, based on the extent to which they contradict the null hypothesis [3, 10].

Inherent multiplicity is also at the heart of a potential controversy when an interim analysis is used, the null hypothesis is not rejected, the study continues to the final analysis, and the final P value is greater than the adjusted alpha level yet less than the overall alpha level (see **Sequential Testing**). For example, suppose that a maximum of five analyses are planned, and the overall alpha level is 0.05 two-sided, so that 1.96 would be used as the critical value for a single analysis. But with five analyses, the critical values might instead be $\{2.41, 2.41, 2.41, 2.41, 2.41\}$ if the Pocock sequential boundaries are used or $\{4.56, 3.23, 2.63, 2.28, 2.04\}$ if the O'Brien–Fleming sequential boundaries are used [9]. Now suppose that none of the first four tests result in early stopping, and the test statistic for the fifth analysis is 2.01. In fact, the test statistic might even assume the value 2.01 for each of the five analyses, and there would be no early stopping.

In such a case, one can lament that if only no penalty had been applied for the interim analysis, then the final results, or, indeed, the results of any of the other four analyses, would have attained statistical significance. And this is true, of course, but it represents a shift in the ranking of all possible outcomes. Prior to the study, it was decided that a highly significant early difference would have been treated as more important than a small difference at the end of the study. That is, an initial test statistic greater than 2.41 if the Pocock sequential boundaries are used, or an initial test statistic greater than 4.56 if the O'Brien–Fleming sequential boundaries are used, would carry more weight than a final test statistic of 1.96. Hence, the bet (for statistical significance) was placed on the large early difference, in the form of the interim analysis, but it turned out to be a losing bet, and, to make matters worse, the standard bet of 1.96 with one analysis would have been a winning bet. Yet, lamenting this regret is tantamount to requesting a refund on a losing lottery ticket. In fact, almost any time there is a choice of analyses, or test statistics, the P value will depend on this choice [4]. It is clear that again inherent multiplicity is at the heart of this issue.

Clearly, rejecting a prespecified hypotheses is more convincing than rejecting a post hoc hypotheses, even at the same alpha level. This suggests that the timing of the statement of the hypothesis could have implications for how much alpha is applied to the resulting analysis. In fact, it is difficult to

answer the questions ‘Where does alpha come from?’ and ‘How much alpha should be applied?’, but in trying to answer these questions, one may well suggest that the process of generating alpha requires a prespecified hypothesis [5]. Yet, this is not very satisfying because sometimes unexpected findings need to be explored. In fact, discarding these findings may be quite problematic itself [1]. For example, a confounder may present itself only after the data are in, or a key assumption underlying the validity of the planned analysis may be found to be violated. In theory, it would always be better to test the hypothesis on new data, rather than on the same data that suggested the hypothesis, but this is not always feasible, or always possible [1]. Fortunately, there are a variety of approaches to controlling the overall Type I error rate while allowing for flexibility in testing hypotheses that were suggested by the data. Two such approaches have already been mentioned, specifically the Pocock sequential boundaries and the O’Brien–Fleming sequential boundaries, which allow one to avoid having to select just one analysis time [9].

In the context of the **analysis of variance**, Fisher’s least significant difference (LSD) can be used to control the overall Type I error rate when arbitrary pairwise comparisons are desired (*see Multiple Comparison Procedures*). The approach is based on operating in protected mode, so that these pairwise comparisons occur only if an overall equality null hypothesis is first rejected (*see Multiple Testing*). Of course, the overall Type I error rate that is being protected is the one that applies to the global null hypothesis that all means are the same. This may offer little consolation if one mean is very large, another is very small, and, because of these two, all other means can be compared without adjustment (*see Multiple Testing*). The Scheffe method offers simultaneous inference, as in any linear combination of means can be tested. Clearly, this generalizes the pairwise comparisons that correspond to pairwise comparisons of means.

Another area in which post hoc issues arise is the selection of the primary outcome measure. Sometimes, there are various outcome measures, or end points, to be considered. For example, an intervention may be used in hopes of reducing childhood smoking, as well as drug use and crime. It may not be clear at the beginning of the study which of these outcome measures will give the best chance to

demonstrate statistical significance. In such a case, it can be difficult to select one outcome measure to serve as the primary outcome measure. Sometimes, however, the outcome measures are fusible [4], and, in this case, this decision becomes much easier. To clarify, suppose that there are two candidate outcome measures, say response and complete response (however these are defined in the context in question). Furthermore, suppose that a complete response also implies a response, so that each subject can be classified as a nonresponder, a partial responder, or a complete responder.

In this case, the two outcome measures are fusible, and actually represent different cut points of the same underlying ordinal outcome measure [4]. By specifying neither component outcome measure, but rather the information-preserving composite end point (IPCE), as the primary outcome measure, one avoids having to select one or the other, and can find legitimate significance if either outcome measure shows significance. The IPCE is simply the underlying ordinal outcome measure that contains each component outcome measure as a binary sub-endpoint. Clearly, using the IPCE can be cast as a method for allowing post hoc testing, because it obviates the need to prospectively select one outcome measure or the other as the primary one. Suppose, for example, that two key outcome measures are response (defined as a certain magnitude of benefit) and complete response (defined as a somewhat higher magnitude of benefit, but on the same scale). If one outcome measure needs to be selected as the primary one, then it may be unclear which one to select. Yet, because both outcome measures are measured on the same scale, this decision need not be addressed, because one could fuse the two outcome measures together into a single trichotomous outcome measure, as in Table 1.

Even when one recognizes that an outcome measure is ordinal, and not binary, there may still be a desire to analyze this outcome measure as if it were binary by dichotomizing it. Of course, there is a different binary sub-endpoint for each cut point of

Table 1 Hypothetical data set #1

	No response	Partial response	Complete response
Control	10	10	10
Active	10	0	20

the original ordinal outcome measure. In the previous paragraph, for example, one could analyze the binary response outcome measure (20/30 in the control group vs 20/30 in the active group in the fictitious data in Table 1), or one could analyze the binary complete response outcome measure (10/30 in the control group vs 20/30 in the active group in the fictitious data in Table 1). With k ordered categories, there are $k - 1$ binary sub-endpoints, together comprising the Lancaster decomposition [12].

In Table 1, the overall response rate would not differentiate the two treatment groups, whereas the complete response rate would. If one knew this ahead of time, then one might select the overall response rate. But the data could also turn out as in Table 2.

Now the situation is reversed, and it is the overall response rate that distinguishes the two treatment groups (30/30 or 100% in the active group vs 20/30 or 67% in the control group), whereas the complete response rate does not (10/30 or 33% in the active group vs 10/30 or 33% in the control group). If either pattern is possible, then it might not be clear, prior to collecting the data, which of the two outcome measures, complete response or overall response, would be preferred. The Smirnov test (see **Kolmogorov-Smirnov Tests**) can help, as it allows one to avoid having to prespecify the particular sub-endpoint to analyze. That is, it allows for the simultaneous testing of both outcome measures in the cases presented above, or of all $k - 1$ outcome measures more generally, while still preserving the overall Type I error rate. This is achieved by letting the data dictate the outcome measure (i.e., selecting that outcome measure that maximizes the test statistic), and then comparing the resulting test statistic not to its own null sampling distribution, but rather to the null sampling distribution of the maximally chosen test statistic.

Adaptive tests are more general than the Smirnov test, as they allow for an optimally chosen set of scores for use with a linear rank test, with the scores essentially being selected by the data [7]. That is, the Smirnov test allows for a data-dependent choice of

Table 3 Hypothetical data set #3

	No response	Partial response	Complete response
Control	10	10	10
Active	5	10	15

the cut point for a subsequent application on of an analogue of Fisher's exact test (see **Exact Methods for Categorical Data**), whereas adaptive tests allow the data to determine the numerical scores to be assigned to the columns for a subsequent linear rank test. Only if those scores are zero to the left of a given column and one to the right of it will the linear rank test reduce to Fisher's exact test. For the fictitious data in Tables 1 and 2, for example, the Smirnov test would allow for the data-dependent selection of the analysis of either the overall response rate or the complete response rate, but the Smirnov test would not allow for an analysis that exploits reinforcing effects. To see why this can be a problem, consider Table 3.

Now both of the aforementioned measures can distinguish the two treatment groups, and in the same direction, as the complete response rates are 50% and 33%, whereas the overall response rates are 83% and 67%. The problem is that neither one of these measures by itself is as large as the effect seen in Table 1 or Table 2. Yet, overall, the effect in Table 3 is as large as that seen in the previous two tables, but only if the reinforcing effects of both measures are considered. After seeing the data, one might wish to use a linear rank test by which numerical scores are assigned to the three columns and then the mean scores across treatment groups are compared. One might wish to use equally spaced scores, such as 1, 2, and 3, for the three columns. Adaptive tests would allow for this choice of scores to be used for Table 3 while preserving the Type I error rate by making the appropriate adjustment for the inherent multiplicity.

The basic idea behind adaptive tests is to subject the data to every conceivable set of scores for use with a linear rank test, and then compute the minimum of all the resulting P values. This minimum P value is artificially small because the data were allowed to select the test statistic (that is, the scores for use with the linear rank test). However, this minimum P value can be used not as a (valid) P value, but rather as a test statistic to be compared to the null sampling distribution of the minimal P value so

Table 2 Hypothetical data set #2

	No response	Partial response	Complete response
Control	10	10	10
Active	0	20	10

computed. As a result, the sample space can be partitioned into regions on which a common test statistic is used, and it is in this sense that the adaptive test allows the data to determine the test statistic, in a post hoc fashion. Yet, because of the manner in which the reference distribution is computed (on the basis of the exact design-based permutation null distribution of the test statistic [8] factoring in how it was selected on the basis of the data), the resulting test is exact. This adaptive testing approach was first proposed by Berger [2], but later generalized by Berger and Ivanova [7] to accommodate preferred alternative hypotheses and to allow for greater or lesser belief in these preferred alternatives.

Post hoc comparisons can and should be explored, but with some caveats. First, the criteria for selecting such comparisons to be made should be specified prospectively [1], when this is possible. Of course, it may not always be possible. Second, plausibility and subject area knowledge should be considered (as opposed to being based exclusively on statistical considerations) [1]. Third, if at all possible, these comparisons should be considered as hypothesis-generating, and should lead to additional studies to produce new data to test these hypotheses, which would have been post hoc for the initial experiments, but are now prespecified for the additional ones.

References

- [1] Adams, K.F. (1998). Post hoc subgroup analysis and the truth of a clinical trial, *American Heart Journal* **136**, 753–758.
- [2] Berger, V.W. (1998). Admissibility of exact conditional tests of stochastic order, *Journal of Statistical Planning and Inference* **66**, 39–50.
- [3] Berger, V.W. (2001). The p-value interval as an inferential tool, *The Statistician* **50**(1), 79–85.
- [4] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [5] Berger, V.W. (2004). On the generation and ownership of alpha in medical studies, *Controlled Clinical Trials* **25**, 613–619.
- [6] Berger, V.W. & Bears, J. (2003). When can a clinical trial be called ‘randomized’? *Vaccine* **21**, 468–472.
- [7] Berger, V.W. & Ivanova, A. (2002). Adaptive tests for ordered categorical data, *Journal of Modern Applied Statistical Methods* **1**, 269–280.
- [8] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**(1), 74–82.
- [9] Demets, D.L. & Lan, K.K.G. (1994). Interim analysis: the alpha spending function approach, *Statistics in Medicine* **13**, 1341–1352.
- [10] Hacking, I. (1965). *The Logic of Statistical Inference*, Cambridge University Press, Cambridge.
- [11] Macdonald, R.R. (2002). The incompleteness of probability models and the resultant implications for theories of statistical inference, *Understanding Statistics* **1**(3), 167–189.
- [12] Permutt, T. & Berger, V.W. (2000). A new look at rank tests in ordered $2 \times k$ contingency tables, *Communications in Statistics – Theory and Methods* **29**, 989–1003.
- [13] Senn, S. (1997). *Statistical Issues in Drug Development*, Wiley, Chichester.

VANCE W. BERGER

ACE Model

HERMINE H. MAES

Volume 1, pp. 5–10

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

ACE Model

Introduction

The ACE model refers to a genetic epidemiological model that postulates that additive genetic factors (A) (*see Additive Genetic Variance*), common environmental factors (C), and specific environmental factors (E) account for individual differences in a phenotype (P) (*see Genotype*) of interest. This model is used to quantify the contributions of genetic and environmental influences to variation and is one of the fundamental models of basic genetic epidemiology [6]. Its name is therefore a simple acronym that allows researchers to communicate the fundamentals of a genetic model quickly, which makes it a useful piece of jargon for the genetic epidemiologist. The focus is thus the causes of variation between individuals. In mathematical terms, the total variance of a trait (V_P) is predicted to be the sum of the variance components: $V_P = V_A + V_C + V_E$, where V_A is the **additive genetic variance**, V_C the shared environmental variance (*see Shared Environment*), and V_E the specific environmental variance. The aim of fitting the ACE model is to answer questions about the importance of nature and nurture on individual differences such as ‘How much of the variation in a trait is accounted for by genetic factors?’ and ‘Do shared environmental factors contribute significantly to the trait variation?’. The first of these questions addresses **heritability**, defined as the proportion of the total variance explained by genetic factors ($h^2 = V_A / V_P$). The nature-nurture question is quite old. It was Sir **Francis Galton** [5] who first recognized that comparing the similarity of identical and fraternal twins yields information about the relative importance of heredity versus environment on individual differences. At the time, these observations seemed to conflict with Gregor Mendel’s classical experiments that demonstrated that the inheritance of model traits in carefully bred material agreed with a simple theory of particulate inheritance. **Ronald Fisher** [4] synthesized the views of Galton and Mendel by providing the first coherent account of how the ‘correlations between relatives’ could be explained ‘on the supposition of Mendelian inheritance’. In this chapter, we will first explain each of the sources of variation in quantitative traits in more detail. Second, we briefly discuss the utility of

the classical **twin design** and the tool of **path analysis** to represent the twin model. Finally, we introduce the concepts of model fitting and apply them by fitting models to actual data. We end by discussing the limitations and assumptions, as well as extensions of the ACE model.

Quantitative Genetics

Fisher assumed that the variation observed for a trait was caused by a large number of individual genes, each of which was inherited in a strict conformity to Mendel’s laws, the so-called polygenic model. If the model includes many environmental factors also of small and equal effect, it is known as the multifactorial model. When the effects of many small factors are combined, the distribution of trait values approximates the normal (Gaussian) distribution, according to the **central limit theorem**. Such a distribution is often observed for quantitative traits that are measured on a continuous scale and show individual variation around a mean trait value, but may also be assumed for qualitative or categorical traits, which represent an imprecise measurement of an underlying continuum of liability to a trait (*see Liability Threshold Models*), with superimposed thresholds [3]. The factors contributing to this variation can thus be broken down in two broad categories, genetic and environmental factors. Genetic factors refer to effects of loci on the genome that contain variants (or alleles). Using quantitative genetic theory, we can distinguish between additive and nonadditive genetic factors. Additive genetic factors (A) are the sum of all the effects of individual loci. Nonadditive genetic factors are the result of interactions between alleles on the same locus (dominance, D) or between alleles on different loci (**epistasis**). Environmental factors are those contributions that are nongenetic in origin and can be divided into shared and nonshared environmental factors. Shared environmental factors (C) are aspects of the environment that are shared by members of the same family or people who live together, and contribute to similarity between relatives. These are also called *common or between-family environmental factors*. Nonshared environmental factors (E), also called *specific, unique, or within-family environmental factors*, are factors unique to an individual. These E factors contribute to variation within family members, but not to their covariation. Various study designs exist to quantify the contributions

of these four sources of variation. Typically, these designs include individuals with different degrees of genetic relatedness and environmental similarity. One such design is the family study (*see Family History Versus Family Study Methods in Genetics*), which studies the correlations between parents and offspring, and/or siblings (in a nuclear family). While this design is very useful to test for familial resemblance, it does not allow us to separate additive genetic from shared environmental factors. The most popular design that does allow the separation of genetic and environmental (shared and unshared) factors is the classical twin study.

The Classical Twin Study

The classical twin study consists of a design in which data are collected from identical or monozygotic (MZ) and fraternal or dizygotic (DZ) twins reared together in the same home. MZ twins have identical genotypes, and thus share all their genes. DZ twins, on the other hand, share on average half their genes, as do regular siblings. Comparing the degree of similarity in a trait (or their correlation) provides an indication of the importance of genetic factors to the trait variability. Greater similarity for MZ versus DZ twins suggests that genes account for at least part of the trait. The recognition of this fact led to the development of heritability indices, based on the MZ and DZ correlations. Although these indices may provide a quick indication of the heritability, they may result in nonsensical estimates. Furthermore, in addition to genes, environmental factors that are shared by family members (or twins in this case) also contribute to familial similarity. Thus,

if environmental factors contribute to a trait and they are shared by twins, they will increase correlations equally between MZ and DZ twins. The relative magnitude of the MZ and DZ correlations thus tells us about the contribution of additive genetic (a^2) and shared environmental (c^2) factors. Given that MZ twins share their genotype and shared environmental factors (if reared together), the degree to which they differ informs us of the importance of specific environmental (e^2) factors.

If the twin similarity is expressed as correlations, one minus the MZ correlation is the proportion due to specific environment (Figure 1). Using the raw scale of measurement, this proportion can be estimated from the difference between the MZ covariance and the variance of the trait. With the trait variance and the MZ and DZ covariance as unique observed statistics, we can estimate the contributions of additive genes (A), shared (C), and specific (E) environmental factors, according to the genetic model. A useful tool to generate the expectations for the variances and covariances under a model is path analysis [11].

Path Analysis

A path diagram is a graphical representation of the model, and is mathematically complete. Such a path diagram for a genetic model, by convention, consists of boxes for the observed variables (the traits under study) and circles for the **latent variables** (the genetic and environmental factors that are not measured but inferred from data on relatives, and are standardized). The contribution of the latent variables to the variances of the observed variables is specified in the path

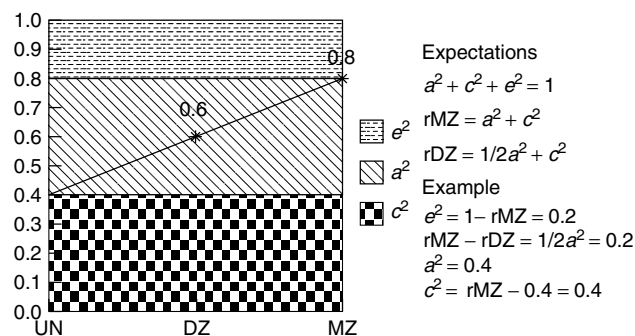


Figure 1 Derivation of variance components from twin correlations

coefficients, which are regression coefficients (represented by single-headed arrows from the latent to the observed variables). We further add two kinds of double-headed arrows to the path coefficients model. First, each of the latent variables has a double-headed arrow pointing to itself, which is fixed to 1.0. Note that we can either estimate the contribution of the latent variables through the path coefficients and standardize the latent variables or we can estimate the variances of the latent variables directly while fixing the paths to the observed variables. We prefer the path coefficients approach to the variance components model, as it generalizes much more easily to advanced models. Second, on the basis of quantitative genetic theory, we model the covariance between twins by adding double-headed arrows between the additive genetic and shared environmental latent variables. The correlation between the additive genetic latent variables is fixed to 1.0 for MZ twins, because they share all their genes. The corresponding value for DZ twins is 0.5, derived from biometrical principles [7]. The correlation between shared environmental latent variables is fixed to 1.0 for MZ and DZ twins, reflecting the equal environments assumption. Specific environmental factors do not contribute to covariance between twins, which is implied by omitting a double-headed arrow. The full path diagrams for MZ and DZ twins are presented in Figure 2.

The expected covariance between two variables in a path diagram may be derived by tracing all connecting routes (or ‘chains’) between the variables while following the rules of path analysis, which are: (a) trace backward along an arrow, change direction in a double-headed arrow and then trace forward, or simply forward from one variable to the other; this implies to trace through at most one two-way arrow in each chain of paths; (b) pass through each variable only once in each chain of paths. The expected covariance between two variables, or the expected variance of a variable, is computed by multiplying

together all the coefficients in a chain, and then summing over all legitimate chains. Using these rules, the expected covariance between the phenotypes of twin 1 and twin 2 for MZ twins and DZ twins can be shown to be:

$$\text{MZ cov} = \begin{bmatrix} a^2 + c^2 + e^2 & a^2 + c^2 \\ a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}$$

$$\text{DZ cov} = \begin{bmatrix} a^2 + c^2 + e^2 & 0.5a^2 + c^2 \\ 0.5a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}$$

This translation of the ideas of the theory into mathematical form comprises the stage of model building. Then, it is necessary to choose the appropriate study design, in this case the classical twin study, to generate critical data to test the model.

Model Fitting

The stage of model fitting allows us to compare the predictions with actual observations in order to evaluate how well the model fits the data using goodness-of-fit statistics. Depending on whether the model fits the data or not, it is accepted or rejected, in which case an alternative model may be chosen. In addition to the goodness-of-fit of the model, estimates for the genetic and environmental parameters are obtained. If a model fits the data, we can further test the significance of these parameters of the model by adding or dropping parameters and evaluate the improvement or decrease in model fit using likelihood-ratio tests. This is equivalent to estimating **confidence intervals**. For example, if the ACE model fits the data, we may drop the additive genetic (a) parameter and refit the model (now a CE model). The difference in the goodness-of-fit statistics for the two models, the ACE and the CE models, provides a likelihood-ratio test with one degree of freedom for the significance of a . If this test is significant, additive genetic factors contribute significantly to the variation

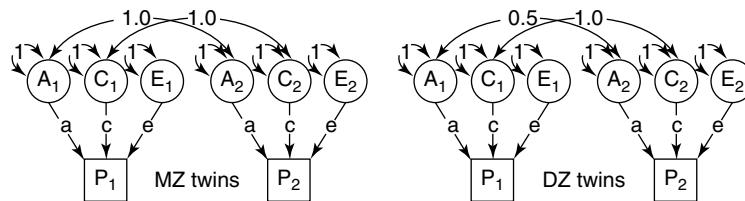


Figure 2 Path diagram for the ACE model applied to data from MZ and DZ twins

in the trait. If it is not, a could be dropped from the model, according to the principle of parsimony. Alternatively, we could calculate the confidence intervals around the parameters. If these include zero for a particular parameter, it indicates that the parameter is not significantly different from zero and could be dropped from the model. Given that significance of parameters is related to power of the study, confidence intervals provide useful information around the precision with which the point estimates are known. The main advantages of the model fitting approach are thus (a) assessing the overall model fit, (b) incorporating sample size and precision, and (c) providing sensible heritability estimates. Other advantages include that it (d) generalizes to the multivariate case and to extended pedigrees, (e) allows the addition of covariates, (f) makes use of all the available data, and (g) is suitable for selected samples. If we are interested in testing the ACE model and quantifying the degree to which genetic and environmental factors contribute to the variability of a trait, data need to be collected on relatively large samples of genetically informative relatives, for example, MZ and DZ twins. The ACE model can then be fitted either directly to the raw data or to summary statistics (covariance matrices) and decisions made about the model on the basis of the goodness-of-fit. There are several statistical modeling packages available capable of fitting the model, for example, EQS, SAS, Lisrel, and Mx (*see Structural Equation Modeling: Software*). The last program was designed specifically with genetic epidemiologic models in mind, and provides great flexibility in specifying both basic and advanced models [10]. Mx models are specified in terms of matrices, and matrix algebra is used to generate the expected covariance matrices or other statistics of the model to be fitted.

Example

We illustrate the ACE model, with data collected in the Virginia Twin Study of Adolescent Behavior

Development (VTSABD) [2]. One focus of the study is conduct disorder, which is characterized by a set of disruptive and destructive behaviors. Here we use a summed symptom score, normalized and standardized within age and sex, and limit the example to the data on 8–16-year old boys, rated by their mothers. Using the sum score data on 295 MZ and 176 DZ pairs of twins, we first estimated the means, variances, and covariances by maximum likelihood in Mx [10], separately for the two twins and the two zygosity groups (MZ and DZ, see Table 1). This model provides the overall likelihood of the data and serves as the ‘saturated’ model against which other models may be compared. It has 10 estimated parameters and yields a -2 times log-likelihood of 2418.575 for 930 degrees of freedom, calculated as the number of observed statistics (940 nonmissing data points) minus the number of estimated parameters. First, we tested the equality of means and variances by twin order and zygosity by imposing equality constraints on the respective parameters. Neither means nor variances were significantly different for the two members of a twin pair, nor did they differ across zygosity ($\chi^2_6 = 5.368$, $p = .498$).

Then we fitted the ACE model, thus partitioning the variance into additive genetic, shared, and specific environmental factors. We estimated the means freely as our primary interest is in the causes of individual differences. The likelihood ratio test – obtained by subtracting the -2 log-likelihood of the saturated model from that of the ACE model (2421.478) for the difference in degrees of freedom of the two models (933–930) – indicates that the ACE model gives an adequate fit to the data ($\chi^2_3 = 2.903$, $p = .407$). We can evaluate the significance of each of the parameters by estimating confidence intervals, or by fitting submodels in which we fix one or more parameters to zero. The series of models typically tested includes the ACE, AE, CE, E, and ADE models. Alternative models can be compared by several fit indices,

Table 1 Means and variances estimated from the raw data on conduct disorder in VTSABD twins

	Monozygotic male twins (MZM)		Dizygotic male twins (DZM)	
	T1	T2	T1	T2
Expected means	–0.0173	–0.0228	0.0590	–0.0688
Expected covariance matrix	T1	0.9342	T1	1.0908
	T2	0.5930	T2	0.3898
		0.8877		0.9030

Table 2 Goodness-of-fit statistics and parameter estimates for conduct disorder in VTSABD twins

Model	χ^2	df	p	AIC	$\Delta\chi^2$	df	a^2	c^2	e^2
ACE	2.903	3	.407	-3.097			.57 (.33 – .72)	.09 (.00 – .31)	.34 (.28 – .40)
AE	3.455	4	.485	-4.545	0.552	1	.66 (.60 – .72)		.34 (.28 – .40)
CE	26.377	4	.000	18.377	23.475	1		.55 (.48 – .61)	.45 (.39 – .52)
E	194.534	5	.000	184.534	191.63	2			1.00
ADE	3.455	3	.327	-2.545			.66 (.30 – .72)	d^2 .00 (.00 – .36)	.34 (.28 – .40)

AIC: Akaike's information criterion; a^2 : additive genetic variance component; c^2 : shared environmental variance component; e^2 : specific environmental variance component.

for example, the **Akaike's Information Criterion** (AIC; [1]), which takes into account both goodness-of-fit and parsimony and favors the model with the lowest value for AIC. Results from fitting these models are presented in Table 2. Dropping the shared environmental parameter c did not deteriorate the fit of the model. However, dropping the a path resulted in a significant decrease in model fit, suggesting that additive genetic factors account for part of the variation observed in conduct disorder symptoms, in addition to specific environmental factors. The latter are always included in the models for two main reasons. First, almost all variables are subject to error. Second, the likelihood is generally not defined when twins are predicted to correlate perfectly. The same conclusions would be obtained from judging the confidence intervals around the parameters a^2 (which do not include zero) and c^2 (which do include zero). Not surprisingly, the E model fits very badly, indicating highly significant family resemblance.

Typically, the ADE model (with dominance instead of common environmental influences) is also fitted, predicting a DZ correlation less than half the MZ correlation. This is the opposite expectation of the ACE model that predicts a DZ correlation greater than half the MZ correlation. Given that dominance (d) and shared environment (c) are confounded in the classical twin design and that the ACE and ADE models are not nested, both are fitted and preference is given to the one with the best absolute goodness-of-fit, in this case the ACE model. Alternative designs, for example, twins reared apart, provide additional unique information to identify and simultaneously estimate c and d separately. In this example, we conclude that the AE model is the best fitting and most parsimonious model to explain variability in conduct disorder symptoms in adolescent boys rated by their mothers in the VTSABD. Additive genetic factors account for

two-thirds of the variation, with the remaining one-third explained by specific environmental factors. A more detailed description of these methods may be found in [8].

Limitations and Assumptions

Although the classical twin study is a powerful design to infer the causes of variation in a trait of interest, it is important to reflect on the limitations when interpreting results from fitting the ACE model to twin data. The power of the study depends on a number of factors, including among others the study design, the sample size, the effect sizes of the components of variance, and the significance level [9]. Further, several assumptions are made when fitting the ACE model. First, it is assumed that the effects of A, C, and E are linear and additive (i.e., no genotype by environment interaction) and mutually independent (i.e., no genotype-environment covariance). Second, the effects are assumed to be equal across twin order and zygosity. Third, we assume that the contribution of environmental factors to twins' similarity for a trait is equal for MZ and DZ twins (equal environments assumption). Fourth, no direct influence exists from a twin on his/her co-twin (no reciprocal sibling environmental effect). Finally, the parental phenotypes are assumed to be independent (random mating). Some of these assumptions may be tested by extending the twin design.

Extensions

Although it is important to answer the basic questions about the importance of genetic and environmental factors to variation in a trait, the information obtained remains descriptive. However, it forms the basis for

more advanced questions that may inform us about the nature and kind of the genetic and environmental factors. Some examples of these questions include: Is the contribution of genetic and/or environmental factors the same in males and females? Is the heritability equal in children, adolescents, and adults? Do the same genes account for variation in more than one phenotype, or thus explain some or all of the covariation between the phenotypes? Does the impact of genes and environment change over time? How much parent-child similarity is due to shared genes versus shared environmental factors?

This basic model can be extended in a variety of ways to account for sex limitation, genotype $\times$ environment interaction, sibling interaction, and to deal with multiple variables measured simultaneously (multivariate genetic analysis) or longitudinally (developmental genetic analysis). Other relatives can also be included, such as siblings, parents, spouses, and children of twins, which may allow better separation of genetic and cultural transmission and estimation of assortative mating and twin and sibling environment. The addition of measured genes (genotypic data) or measured environments may further refine the partitioning of the variation, if these measured variables are linked or associated with the phenotype of interest. The ACE model is thus the cornerstone of modeling the causes of variation.

References

- [1] Akaike, H. (1987). Factor analysis and AIC, *Psychometrika* **52**, 317–332.
- [2] Eaves, L.J., Silberg, J.L., Meyer, J.M., Maes, H.H., Simonoff, E., Pickles, A., Rutter, M., Neale, M.C., Reynolds, C.A., Erickson, M., Heath, A.C., Loeber, R., Truett, K.R. & Hewitt, J.K. (1997). Genetics and developmental psychopathology: 2. The main effects of genes and environment on behavioral problems in the Virginia Twin study of adolescent behavioral development, *Journal of Child Psychology and Psychiatry* **38**, 965–980.
- [3] Falconer, D.S. (1989). *Introduction to Quantitative Genetics*, Longman Scientific & Technical, New York.
- [4] Fisher, R.A. (1918). The correlations between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
- [5] Galton, F. (1865). Hereditary talent and character, *MacMillan's Magazine* **12**, 157–166.
- [6] Kendler, K.S. & Eaves, L.J. (2004). *Advances in Psychiatric Genetics*, American Psychiatric Association Press.
- [7] Mather, K. & Jinks, J.L. (1971). *Biometrical Genetics*, Chapman and Hall, London.
- [8] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers BV, Dordrecht.
- [9] Neale, M.C., Eaves, L.J. & Kendler, K.S. (1994). The power of the classical twin study to resolve variation in threshold traits, *Behavior Genetics* **24**, 239–225.
- [10] Neale, M.C., Boker, S.M., Xie, G. & Maes, H.H. (2003). *Mx: Statistical Modeling*, 6th Edition, VCU Box 900126, Department of Psychiatry, Richmond, 23298.
- [11] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.

HERMINE H. MAES

Adaptive Random Assignment

VANCE W. BERGER AND YANYAN ZHOU

Volume 1, pp. 10–13

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Adaptive Random Assignment

Adaptive Allocation

The primary objective of a comparative trial is to provide a precise and valid treatment comparison (*see Clinical Trials and Intervention Studies*). Another objective may be to minimize exposure to the inferior treatment, the identity of which may be revealed during the course of the study. The two objectives together are often referred to as *bandit problems* [5], an essential feature of which is to balance the conflict between information gathering (benefit to society) and the immediate payoff that results from using what is thought to be best at the time (benefit to the individual). Because **randomization** promotes (but does not guarantee [3]) comparability among the study groups in both known and unknown covariates, randomization is rightfully accepted as the ‘gold standard’ solution for the first objective, valid comparisons. There are four major classes of randomization procedures, including unrestricted randomization, restricted randomization, covariate-adaptive randomization, and response-adaptive randomization [6]. As the names would suggest, the last two classes are adaptive designs.

Unrestricted randomization is not generally used in practice because it is susceptible to chronological bias, and this would interfere with the first objective, the valid treatment comparison. Specifically, the lack of restrictions allows for long runs of one treatment or another, and hence the possibility that at some point during the study, even at the end, the treatment group sizes could differ substantially. If this is the case, so that more ‘early’ subjects are in one treatment group and more ‘late’ subjects are in another, then any apparent treatment effects would be confounded with time effects. Restrictions on the randomization are required to ensure that at no point during the study are the treatment group sizes too different. Yet, too many restrictions lead to a predictable allocation sequence, which can also compromise validity. It can be a challenge to find the right balance of restrictions on the randomization [4], and sometimes a adaptive design is used. Perhaps the most common covariate-adaptive design is minimization [7], which minimizes a covariate imbalance function.

Covariate-adaptive Randomization Procedures

Covariate-adaptive (also referred to as *baseline-adaptive*) randomization is similar in intention to **stratification**, but takes the further step of balancing baseline covariate distributions dynamically, on the basis of the existing baseline composition of the treatment groups at the time of allocation. This procedure is usually used when there are too many important prognostic factors for stratification to handle reasonably (there is a limit to the number of strata that can be used [8]). For example, consider a study of a behavioral intervention with only 50 subjects, and 6 strong predictors. Even if each of these 6 predictors is binary, that still leads to 64 strata, and on average less than one subject per stratum. This situation would defeat the purpose of stratification, in that most strata would then not have both treatment groups represented, and hence no matching would occur. The treatment comparisons could then not be considered within the strata.

Unlike stratified randomization, in which an allocation schedule is generated separately for each stratum prior to the start of study, covariate-adaptive procedures are dynamic. The treatment assignment of a subject is dependent on the subject’s vector of covariates, which will not be determined until his or her arrival. Minimization [7] is the most commonly used covariate-adaptive procedure. It ensures excellent balance between the intervention groups for specified prognostic factors by assigning the next participant to whichever group minimizes the imbalance between groups on specified prognostic factors. The balance can be with respect to main effects only, say gender and smoking status, or it can mimic stratification and balance with respect to joint distributions, as in the cross classification of smoking status and gender. In the former case, each treatment group would be fairly equally well represented among smokers, nonsmokers, males, and females, but not necessarily among female smokers, for example.

As a simple example, suppose that the trial is underway, and 32 subjects have already been enrolled, 16 to each group. Suppose further that currently Treatment Group A has four male smokers, five female smokers, four male nonsmokers, and three female nonsmokers, while Treatment Group B has five male smokers, six female smokers, two male nonsmokers, and three female nonsmokers. The 33rd subject to be enrolled is a male smoker. Provisionally

2 Adaptive Random Assignment

place this subject in Treatment Group A, and compute the marginal male imbalance to be $(4 + 4 + 1 - 5 - 2) = 2$, the marginal smoker imbalance to be $(4 + 5 + 1 - 5 - 6) = -1$, and the joint male smoker imbalance to be $(4 + 1 - 5) = 0$. Now provisionally place this subject in Treatment Group B and compute the marginal male imbalance to be $(4 + 4 - 5 - 2 - 1) = 0$, the marginal smoker imbalance to be $(4 + 5 - 5 - 6 - 1) = -2$, and the joint male smoker imbalance to be $(4 - 5 - 1) = -2$. Using the joint balancing, Treatment Group A would be preferred. The actual allocation may be deterministic, as in simply assigning the subject to the group that leads to better balance, A in this case, or it may be stochastic, as in making this assignment with high probability. For example, one might add one to the absolute value of each imbalance, and then use the ratios as probabilities.

So here the probability of assignment to A would be $(2 + 1)/[(0 + 1) + (2 + 1)] = 3/4$ and the probability of assignment to B would be $(0 + 1)/[(2 + 1) + (0 + 1)] = 1/4$. If we were using the marginal balancing technique, then a weight function could be used to weigh either gender or smoking status more heavily than the other or they could each have the same weight. Either way, the decision would again be based, either deterministically or stochastically, on which treatment group minimizes the imbalance, and possibly by how much.

Response-adaptive Randomization Procedures

In response-adaptive randomization, the treatment allocations depend on the previous subject outcomes, so that the subjects are more likely to be assigned to the ‘superior’ treatment, or at least to the one that is found to be superior so far. This is a good way to address the objective of minimizing exposure to an inferior treatment, and possibly the only way to address both objectives discussed above [5]. Response-adaptive randomization procedures may determine the allocation ratios so as to optimize certain criteria, including minimizing the expected number of treatment failures, minimizing the expected number of patients assigned to the inferior treatment, minimizing the total sample size, or minimizing the total cost. They may also follow intuition, often as urn models. A typical urn model starts with k balls of each color, with each color representing a distinct treatment group (that is, there is a

one-to-one correspondence between the colors of the balls in the urn and the treatment groups to which a subject could be assigned). A ball is drawn at random from the urn to determine the treatment assignment. Then the ball is replaced, possibly along with other balls of the same color or another color, depending on the response of the subject to the initial treatment [10].

With this design, the allocation probabilities depend not only on the previous treatment assignments but also on the responses to those treatment assignments; this is the basis for calling such designs ‘response adaptive’, so as to distinguish them from covariate-adaptive designs. Perhaps the most well-known actual trial that used a response-adaptive randomization procedure was the Extra Corporeal Membrane Oxygenation (ECMO) Trial [1]. ECMO is a surgical procedure that had been used for infants with respiratory failure who were dying and were unresponsive to conventional treatment of ventilation and drug. Data existed to suggest that the ECMO treatment was safe and effective, but no randomized controlled trials had confirmed this. Owing to prior data and beliefs, the ECMO investigators were reluctant to use equal allocation. In this case, response-adaptive randomization is a practical procedure, and so it was used.

The investigators chose the randomized play-the-winner RPW(1,1) rule for the trial. This means that after a ball is chosen from the urn and replaced, one additional ball is added to the urn. This additional ball is of the same color as the previously chosen ball if the outcome is a response (survival, in this case). Otherwise, it is of the opposite color. As it turns out, the first patient was randomized to the ECMO treatment and survived, so now ECMO had two balls to only one conventional ball. The second patient was randomized to conventional therapy, and he died. The urn composition then had three ECMO balls and one control ball. The remaining 10 patients were all randomized to ECMO, and all survived. The trial then stopped with 12 total patients, in accordance with a prespecified stopping rule.

At this point, there was quite a bit of controversy regarding the validity of the trial, and whether it was truly a controlled trial (since only one patient received conventional therapy). Comparisons between the two treatments were questioned because they were based on a sample of size 12, again, with only one subject in one of the treatment groups. In fact, depending

on how the data were analyzed, the P value could range from 0.001 (an analysis that assumes complete randomization and ignores the response-adaptive randomization; [9]) to 0.620 (a permutation test that conditions on the observed sequences of responses; [2]) (see **Permutation Based Inference**).

Two important lessons can be learned from the ECMO Trial. First, it is important to start with more than one ball corresponding to each treatment in the urn. It can be shown that starting out with only one ball of each treatment in the urn leads to instability with the randomized play-the-winner rule. Second, a minimum sample size should be specified to avoid the small sample size found in ECMO. It is also possible to build in this requirement by starting the trial as a nonadaptively randomized trial, until a minimum number of patients are recruited to each treatment group. The results of an interim analysis at this point can determine the initial constitution of the urn, which can be used for subsequent allocations, and updated accordingly. The allocation probability will then eventually favor the treatment with fewer failures or more success, and the proportion of allocations to the better arm will converge to one.

References

- [1] Bartlett, R.H., Roloff, D.W., Cornell, R.G., Andrews, A.F., Dillon, P.W. & Zwischenberger, J.B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study, *Pediatrics* **76**, 479–487.
- [2] Begg, C.B. (1990). On inferences from Wei's biased coin design for clinical trials, *Biometrika* **77**, 467–484.
- [3] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias, *Journal of Modern Applied Statistical Methods* **2**(1), 80–86.
- [4] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.
- [5] Berry, D.A. & Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*, Chapman & Hall, London.
- [6] Rosenberger, W.F. & Lachin, J.M. (2002). *Randomization in Clinical Trials*, John Wiley & Sons, New York.
- [7] Taves, D.R. (1974). Minimization: a new method of assigning patients to treatment and control groups, *Clinical Pharmacology Therapeutics* **15**, 443–453.
- [8] Therneau, T.M. (1993). How many stratification factors are “too many” to use in a randomization plan? *Controlled Clinical Trials* **14**(2), 98–108.
- [9] Wei, L.J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule, *Biometrika* **75**, 603–606.
- [10] Wei, L.J. & Durham, S.D. (1978). The randomized play-the-winner rule in medical trials, *Journal of the American Statistical Association* **73**, 840–843.

VANCE W. BERGER AND YANYAN ZHOU

Adaptive Sampling

ZHEN LI AND VANCE W. BERGER

Volume 1, pp. 13–16

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Adaptive Sampling

Traditional sampling methods do not allow the selection for a sampling unit to depend on the previous observations made during an initial survey; that is, sampling decisions are made and fixed prior to the survey. In contrast, adaptive sampling refers to a sampling technique in which the procedure for selecting sites or units to be included in the sample may depend on the values of the variable of interest already observed during the study [10]. Compared to the traditional fixed sampling procedure, adaptive sampling techniques often lead to more effective results.

To motivate the development of adaptive sampling procedures, consider, for example, a population clustered over a large area that is generally sparse or empty between clusters. If a simple random sample (see **Simple Random Sampling**) is used to select geographical subsections of the large area, then many of the units selected may be empty, and many clusters will be missed. It would, of course, be possible to oversample the clusters if it were known where they are located. If this is not the case, however, then adaptive sampling might be a reasonable procedure. An initial sample of locations would be considered. Once individuals are detected in one of the selected units, the neighbors of that unit might also be added to the sample. This process would be iterated until a cluster sample is built.

This adaptive approach would seem preferable in environmental pollution surveys, drug use epidemiology studies, market surveys, studies of rare animal species, and studies of contagious diseases [12]. In fact, an adaptive approach was used in some important surveys. For example, moose surveys were conducted in interior Alaska by using an adaptive sampling design [3]. Because the locations of highest moose abundance was not known prior to the survey, the spatial location of the next day's survey was based on the current results [3]. Likewise, Roesch [4] estimated the prevalence of insect infestation in some hardwood tree species in Northeastern North America. The species of interest were apt to be rare and highly clustered in their distribution, and therefore it was difficult to use traditional sampling procedures. Instead, an adaptive sampling was used. Once a tree of the species was found, an area of specified radius around it would be searched for additional individuals of the species [4].

Adaptive sampling offers the following advantages [10, 12]:

1. Adaptive sampling takes advantage of population characteristics to obtain more precise estimates of population abundance or density. For example, populations of plants, animals, minerals, and fossil fuels tend to exhibit aggregation patterns because of schooling, flocking, and environmental patchiness. Because the location and shape of the aggregation cannot be predicted before a survey, adaptive sampling can provide a dynamic way to increase the effectiveness of the sampling effort.
2. Adaptive sampling reduces unit costs and time, and improves the precision of the results for a given sample size. Adaptive sampling increases the number of observations, so that more endangered species are observed, and more individuals are monitored. This can result in good estimators of interesting parameters. For example, in spatial sampling, adaptive cluster sampling can provide unbiased efficient estimators of the abundance of rare, clustered populations.
3. Some theoretical results show that adaptive procedures are optimal in the sense of giving the most precise estimates with a given amount of sampling effort.

There are also problems related to adaptive sampling [5]:

1. The final sample size is random and unknown, so the appropriate theories need to be developed for a sampling survey with a given precision of estimation.
2. An inappropriate criterion for adding neighborhoods will affect sample units and compromise the effectiveness of the sampling effort.
3. Great effort must be expended in locating initial units.

Although the idea of adaptive sampling was proposed for some time, some of the practical methods have been developed only recently. For example, adaptive cluster sampling was introduced by Thompson in 1990 [6]. Other new developments include two-stage adaptive cluster sampling [5], adaptive cluster double sampling [2], and inverse adaptive cluster sampling [1]. The basic idea behind adaptive cluster sampling is illustrated in

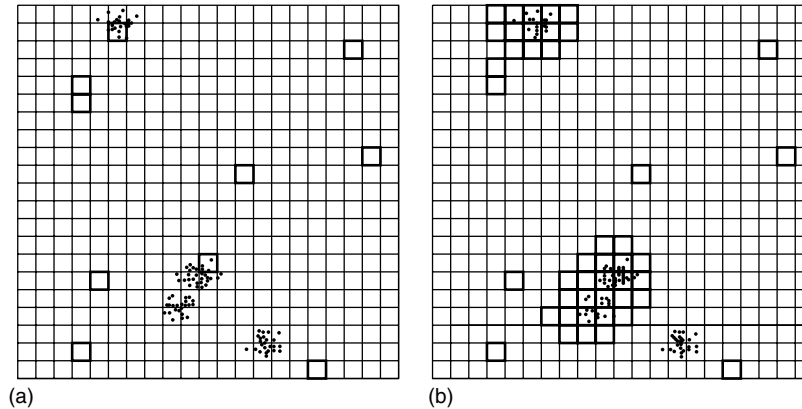


Figure 1 Adaptive cluster sampling and its result (From Thompson, S.K. (1990). Adaptive cluster sampling, *Journal of the American Statistical Association* **85**, 1050–1059 [6])

Figure 1 [6]. There are 400 square units. The following steps are carried out in the sampling procedure.

1. An initial random sample of 10 units is shown in Figure 1(a).
2. In adaptive sampling, we need to define a neighborhood for a sampling unit. A neighborhood can be decided by a prespecified and nonadaptive rule. In this case, the neighborhood of a unit is its set of adjacent units (left, right, top, and bottom).
3. We need to specify a criterion for searching a neighbor. In this case, once one or more objects are observed in a selected unit, its neighborhood is added to the sample.
4. Repeat step 3 for each neighbor unit until no object is observed. In this case, the sample consists of 45 units. See Figure 1(b).

Stratified adaptive cluster sampling (*see Stratification*) is an extension of the adaptive cluster approach. On the basis of prior information about the population or simple proximity of the units, units that are thought to be similar to each other are grouped into strata. Following an initial stratified sample, additional units are added to the sample from the neighborhood of any selected unit when it satisfies the criterion. If additional units are added to the sample, where the high positive identifications are observed, then the sample mean will overestimate the population mean. Unbiased estimators can be obtained by making use of new observations in addition to the observations initially selected.

Thompson [8] proposed several types of estimators that are unbiased for the population mean or total. Some examples are estimators based on expected numbers of initial intersections, estimators based on initial intersection probabilities, and modified estimators based on the Rao–Blackwell method.

Another type of adaptive sampling is the design with primary and secondary units. Systematic adaptive cluster sampling and strip adaptive cluster sampling belong to this type. For both sampling schemes, the initial design could be systematic sampling or strip sampling. That is, the initial design is selected in terms of primary units, while subsequent sampling is in terms of secondary units. Conventional estimators of the population mean or total are biased with such a procedure, so Thompson [7] developed unbiased estimators, such as estimators based on partial selection probabilities and estimators based on partial inclusion probabilities. Thompson [7] has shown that by using a point pattern representing locations of individuals or objects in a spatially aggregated population, the adaptive design can be substantially more efficient than its conventional counterparts.

Commonly, the criterion for additional sampling is a fixed and prespecified rule. In some surveys, however, it is difficult to decide on the fixed criterion ahead of time. In such cases, the criterion could be based on the observed sample values. Adaptive cluster sampling based on order statistics is particularly appropriate for some situations, in which the investigator wishes to search for high values of the variable of interest in addition to estimating the overall mean

or total. For example, the investigator may want to find the pollution ‘hot spots’. Adaptive cluster sampling based on order statistics is apt to increase the probability of observing units with high values, while at the same time allowing for unbiased estimation of the population mean or total. Thompson has shown that these estimators can be improved by using the Rao–Blackwell method [9].

Thompson and Seber [11] proposed the idea of detectability in adaptive sampling. Imperfect detectability is a source of nonsampling error in the natural survey and human population survey. This is because even if a unit is included in the survey, it is possible that not all of the objects can be observed. Examples are a vessel survey of whales and a survey of homeless people. To estimate the population total in a survey with imperfect detectability, both the sampling design and the detection probabilities must be taken into account. If imperfect detectability is not taken into account, then it will lead to underestimates of the population total. In the most general case, the values of the variable of interest are divided by the detection probability for the observed object, and then estimation methods without detectability problems are used.

Finally, regardless of the design on which the sampling is obtained, optimal sampling strategies should be considered. Bias and mean-square errors are usually measured, which lead to reliable results.

References

- [1] Christman, M.C. & Lan, F. (2001). Inverse adaptive cluster sampling, *Biometrics* **57**, 1096–1105.
- [2] Félix Medina, M.H. & Thompson S.K. (1999). Adaptive cluster double sampling, in *Proceedings of the Survey Research Section, American Statistical Association*, Alexandria, VA.
- [3] Gasaway, W.C., DuBois, S.D., Reed, D.J. & Harbo, S.J. (1986). Estimating moose population parameters from aerial surveys, *Biological Papers of the University of Alaska (Institute of Arctic Biology) Number 22*, University of Alaska, Fairbanks.
- [4] Roesch Jr, F.A. (1993). Adaptive cluster sampling for forest inventories, *Forest Science* **39**, 655–669.
- [5] Salehi, M.M. & Seber, G.A.F. (1997). Two-stage adaptive cluster sampling, *Biometrics* **53**(3), 959–970.
- [6] Thompson, S.K. (1990). Adaptive cluster sampling, *Journal of the American Statistical Association* **85**, 1050–1059.
- [7] Thompson, S.K. (1991a). Adaptive cluster sampling: designs with primary and secondary units, *Biometrics* **47**(3), 1103–1115.
- [8] Thompson, S.K. (1991b). Stratified adaptive cluster sampling, *Biometrika* **78**(2), 389–397.
- [9] Thompson, S.K. (1996). Adaptive cluster sampling based on order statistics, *Environmetrics* **7**, 123–133.
- [10] Thompson S.K. (2002). *Sampling*, 2nd Edition, John Wiley & Sons, New York.
- [11] Thompson, S.K. & Seber, G.A.F. (1994). Detectability in conventional and adaptive sampling, *Biometrics* **50**(3), 712–724.
- [12] Thompson, S.K. & Seber, G.A.F. (2002). *Adaptive Sampling*, Wiley, New York.

(See also **Survey Sampling Procedures**)

ZHEN LI AND VANCE W. BERGER

Additive Constant Problem

MICHAEL W. TROSSET

Volume 1, pp. 16–18

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Additive Constant Problem

Introduction

Consider a set of objects or stimuli, for example, a set of colors, and an experiment that produces information about the pairwise dissimilarities of the objects. From such information, two-way **multidimensional scaling** (MDS) constructs a graphical representation of the objects. Typically, the representation consists of a set of points in a low-dimensional Euclidean space. Each point corresponds to one object. Metric two-way MDS constructs the representation in such a way that the pairwise distances between the points approximate the pairwise dissimilarities of the objects.

In certain types of experiments, for example, **Fechner's** method of paired comparisons, Richardson's [7] method of triadic combinations, Klingberg's [4] method of multidimensional rank order, and Torgerson's [9] complete method of triads, the observed dissimilarities represent comparative distances, that is, distances from which an unknown scalar constant has been subtracted. The additive constant problem is the problem of estimating this constant.

The additive constant problem has been formulated in different ways, most notably by Torgerson [9], Messick and Abelson [6], Cooper [2], Saito [8], and Cailliez [1]. In assessing these formulations, it is essential to distinguish between the cases of errorless and fallible data. The former is the province of distance geometry, for example, determining whether or not adding any constant converts the set of dissimilarities to a set of Euclidean distances. The latter is the province of computational and graphical statistics, namely, finding an effective low-dimensional representation of the data.

Classical Formulation

The additive constant problem was of fundamental importance to Torgerson [9], who conceived MDS as comprising three steps: (1) obtain a scale of comparative distances between all pairs of objects; (2) convert the comparative distances to absolute

(Euclidean) distances by adding a constant; and (3) construct a configuration of points from the absolute distances. Here, the comparative distances are given and (1) need not be considered. Discussion of (2) is facilitated by first considering (3).

Suppose that we want to represent a set of objects in p -dimensional Euclidean space. First, we let δ_{ij} denote the dissimilarity of objects i and j . Notice that $\delta_{ii} = 0$, that is, an object is not dissimilar from itself, and that $\delta_{ij} = \delta_{ji}$. It is convenient to organize these dissimilarities into a matrix, Δ . Next, we let X denote a configuration of points. Again, it is convenient to think of X as a matrix in which row i stores the p coordinates of point i . Finally, let $d_{ij}(X)$ denote the Euclidean distances between points i and j in configuration X . As with the dissimilarities, it is convenient to organize the distances into a matrix, $D(X)$. Our immediate goal is to find a configuration whose interpoint distances approximate the specified dissimilarities, that is, to find an X for which $D(X) \approx \Delta$.

The embedding problem of classical distance geometry inquires if there is a configuration whose interpoint distances *equal* the specified dissimilarities. Torgerson [9] relied on the following solution. First, one forms the matrix of squared dissimilarities, $\Delta * \Delta = (\delta_{ij}^2)$. Next, one transforms the squared dissimilarities by double centering (from each δ_{ij}^2 , subtract the averages of the squared dissimilarities in row i and column j , then add the overall average of all squared dissimilarities), then multiplying by $-1/2$. In Torgerson's honor, this transformation is often denoted τ . The resulting matrix is $B^* = \tau(\Delta * \Delta)$. There exists an X for which $D(X) = \Delta$ if and only if all of the **eigenvalues** (latent roots) of B^* are nonnegative and at most p of them are strictly positive. If this condition is satisfied, then the number of strictly positive eigenvalues is called the *embedding dimension* of Δ . Furthermore, if $XX^t = B^*$, then $D(X) = \Delta$.

For Torgerson [9], Δ was a matrix of comparative distances. The dissimilarity matrix to be embedded was $\Delta(c)$, obtained by adding c to each δ_{ij} for which $i \neq j$. The scalar quantity c is the additive constant. In the case of errorless data, Torgerson proposed choosing c to minimize the embedding dimension of $\Delta(c)$. His procedure was criticized and modified by Messick and Abelson [6], who argued that Torgerson underestimated c . Alternatively, one can always choose c sufficiently large that $\Delta(c)$

2 Additive Constant Problem

can be embedded in $(n - 2)$ -dimensional Euclidean space, where n is the number of objects. Cailliez [1] derived a formula for the smallest c for which this embedding is possible.

In the case of fallible data, a different formulation is required. Torgerson argued:

‘This means that with fallible data the condition that B^* be positive semidefinite as a criterion for the points’ existence in real space is not to be taken too seriously. What we would like to obtain is a B^* -matrix whose latent roots consist of

1. A few large positive values (the “true” dimensions of the system), and
2. The remaining values small and distributed about zero (the “error” dimensions).

It may be that for fallible data we are asking the wrong question. Consider the question, “For what value of c will the points be most nearly (in a least-squares sense) in a space of a given dimensionality?”’

Torgerson’s [9] question was posed by de Leeuw and Heiser [3] as the problem of finding the symmetric positive semidefinite matrix of rank $\leq p$ that best approximates $\tau(\Delta(c) * \Delta(c))$ in a least-squares sense. This problem is equivalent to minimizing

$$\zeta(c) = \sum_{i=1}^p [\max(\lambda_i(c), 0) - \lambda_i(c)]^2 + \sum_{i=p+1}^n \lambda_i(c)^2, \quad (1)$$

where $\lambda_1(c) \geq \dots \geq \lambda_n(c)$ are the eigenvalues of $\tau(\Delta(c) * \Delta(c))$. The objective function ζ may have nonglobal minimizers. However, unless n is very large, modern computers can quickly graph $\zeta(\cdot)$, so that the basin containing the global minimizer can be identified by visual inspection. The global minimizer can then be found by a unidimensional search algorithm.

Other Formulations

In a widely cited article, Saito [8] proposed choosing c to maximize a ‘normalized index of fit,’

$$P(c) = \frac{\sum_{i=1}^p \lambda_i^2(c)}{\sum_{i=1}^n \lambda_i^2(c)}. \quad (2)$$

Saito assumed that $\lambda_p(c) > 0$, which implies that $[\max(\lambda_i(c), 0) - \lambda_i(c)]^2 = 0$ for $i = 1, \dots, p$. One can then write

$$P(c) = 1 - \frac{\sum_{i=p+1}^n \lambda_i^2(c)}{\sum_{i=1}^n \lambda_i^2(c)} = 1 - \frac{\zeta(c)}{\eta(c)}. \quad (3)$$

Hence, Saito’s formulation is equivalent to minimizing $\zeta(c)/\eta(c)$, and it is evident that his formulation encourages choices of c for which $\eta(c)$ is large. Why one should prefer such choices is not so clear. Trosset, Baggerly, and Pearl [10] concluded that Saito’s criterion typically results in a larger additive constant than would be obtained using the classical formulation of Torgerson [9] and de Leeuw and Heiser [3].

A comprehensive formulation of the additive constant problem is obtained by introducing a loss function, σ , that measures the discrepancy between a set of p -dimensional Euclidean distances and a set of dissimilarities. One then determines both the additive constant and the graphical representation of the data by finding a pair (c, D) that minimizes $\sigma(D, \Delta(c))$. The classical formulation’s loss function is the squared error that results from approximating $\tau(\Delta(c) * \Delta(c))$ with $\tau(D * D)$. This loss function is sometimes called the *strain criterion*. In contrast, Cooper’s [2] loss function was Kruskal’s [5] raw stress criterion, the squared error that results from approximating $\Delta(c)$ with D . Although the raw stress criterion is arguably more intuitive than the strain criterion, Cooper’s formulation cannot be reduced to a unidimensional optimization problem.

References

- [1] Cailliez, F. (1983). The analytical solution of the additive constant problem, *Psychometrika* **48**, 305–308.
- [2] Cooper, L.G. (1972). A new solution to the additive constant problem in metric multidimensional scaling, *Psychometrika* **37**, 311–322.
- [3] de Leeuw, J. & Heiser, W. (1982). Theory of multidimensional scaling, in *Handbook of Statistics*, Vol. 2, P.R. Krishnaiah & I.N. Kanal, eds, North Holland, Amsterdam, pp. 285–316, Chapter 13.
- [4] Klingberg, F.L. (1941). Studies in measurement of the relations among sovereign states, *Psychometrika* **6**, 335–352.
- [5] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.

-
- [6] Messick, S.J. & Abelson, R.P. (1956). The additive constant problem in multidimensional scaling, *Psychometrika* **21**, 1–15.
- [7] Richardson, M.W. (1938). Multidimensional psychophysics, *Psychological Bulletin* **35**, 659–660; Abstract of presentation at the forty-sixth annual meeting of the American Psychological Association, American Psychological Association (APA), Washington, D.C. September 7–10, 1938.
- [8] Saito, T. (1978). The problem of the additive constant and eigenvalues in metric multidimensional scaling, *Psychometrika* **43**, 193–201.
- [9] Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method, *Psychometrika* **17**, 401–419.
- [10] Trosset, M.W., Baggerly, K.A. & Pearl, K. (1996). Another look at the additive constant problem in multidimensional scaling, Technical Report 96–7, Department of Statistics-MS 138, Rice University, Houston.
- (See also **Bradley–Terry Model**; **Multidimensional Unfolding**)

MICHAEL W. TROSSET

Additive Genetic Variance

DANIELLE POSTHUMA

Volume 1, pp. 18–22

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Additive Genetic Variance

The starting point for gene finding is the observation of population variation in a certain trait. This ‘observed’, or phenotypic, variation may be attributed to genetic and environmental causes. Although environmental causes of phenotypic variation should not be ignored and are highly interesting, in the following section we will focus on the biometric model underlying genetic causes of variation, specifically *additive* genetic causes of variation.

Within a population, one, two, or many different alleles may exist for a gene (*see Allelic Association*). Uniallelic systems will not contribute to population variation. For simplicity, we assume in this treatment one gene with two possible alleles, alleles A1 and A2. By convention, allele A1 has frequency p , while allele A2 has frequency q , and $p + q = 1$. With two alleles, there are three possible **genotypes**: A1A1, A1A2, and A2A2, with corresponding genotypic frequencies p^2 , $2pq$, and q^2 (assuming random mating, equal viability of alleles, no selection, no migration and no mutation, *see* [3]). The genotypic effect on a phenotypic trait (i.e., *the genotypic value*) of genotype A1A1, is by convention called ‘ a ’ and the effect of genotype A2A2 ‘ $-a$ ’. The effect of the heterozygous genotype A1A2 is called ‘ d ’. If the genotypic value of the heterozygote lies exactly at the midpoint of the genotypic values of the two homozygotes ($d = 0$), there is said to be no genetic dominance. If allele A1 is completely dominant over allele A2, effect d equals effect a . If d is larger than a , there is overdominance. If d is unequal to zero and the two alleles produce three discernable phenotypes of the trait, d is unequal to a . This model is also known as the *classical biometrical model* [3, 6] (*see* Figure 1 for a worked example).

The genotypic contribution of a gene to the population mean of a trait (i.e., the mean effect of a gene, or μ) is the sum of the products of the frequencies and the genotypic values of the different genotypes:

$$\begin{aligned} \text{Mean effect} &= (ap^2) + (2pqd) + (-aq^2) \\ &= a(p - q) + 2pqd. \end{aligned} \quad (1)$$

This mean effect of a gene consists of two components: the contribution of the homozygotes [$a(p - q)$] and the contribution of the heterozygotes

[$2pqd$]. If there is no dominance, that is d equals zero, there is no contribution of the heterozygotes and the mean is a simple function of the allele frequencies. If d equals a , which is defined as complete dominance, the population mean becomes a function of the square of the allele frequencies; substituting d for a gives $a(p - q) + 2pqa$, which simplifies to $a(1 - 2q^2)$.

Complex traits such as height or weight are not very likely influenced by a single gene, but are assumed to be influenced by many genes. Assuming only additive and independent effects of all of these genes, the expectation for the population mean (μ) is the sum of the mean effects of all the separate genes, and can formally be expressed as $\mu = \sum a(p - q) + 2 \sum dpq$ (*see* also Figure 2).

Average Effects and Breeding Values

Let us consider a relatively simple trait that seems to be mainly determined by genetics, for example eye color. As can be widely observed, when a brown-eyed parent mates with a blue-eyed parent, their offspring will not be either brown eyed or blue eyed, but may also have green eyes. At present, three genes are known to be involved in human eye color. Two of these genes lie on chromosome 15: the EYCL2 and EYCL3 genes (also known as the BEY1 and BEY2 gene respectively) and one gene lies on chromosome 19; the EYCL1 gene (or GEY gene) [1, 2]. For simplicity, we ignore one gene (BEY1), and assume that only GEY and BEY2 determine eye color. The BEY2 gene has two alleles: a blue allele and a brown allele. The brown allele is completely dominant over the blue allele. The GEY gene also has two alleles: a green allele and a blue allele. The green allele is dominant over the blue allele of GEY but also over the blue allele of BEY2. The brown allele of BEY2 is dominant over the green allele of GEY. Let us assume that the brown-eyed parent has genotype *brown–blue* for the BEY2 gene and *green–blue* for the GEY gene, and that the blue-eyed parent has genotype *blue–blue* for both the BEY2 gene and the GEY gene. Their children can be (a) brown eyed: *brown–blue* for the BEY2 gene and either *blue–blue* or *green–blue* for the GEY gene; (b) green eyed: *blue–blue* for the BEY2 gene and *green–blue* for the GEY gene; (c) blue eyed: *blue–blue* for the BEY2 gene and *blue–blue* for the GEY gene. The possibility of having green-eyed children from a brown-eyed

2 Additive Genetic Variance

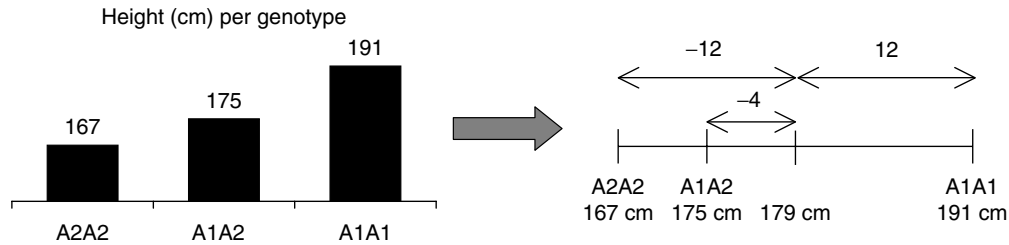


Figure 1 Worked example of genotypic effects, average effects, breeding values, and genetic variation. Assume body height is determined by a single gene with two alleles A1 and A2, and frequencies $p = 0.6$, $q = 0.4$. Body height differs per genotype: A2A2 carriers are 167 cm tall, A1A2 carriers are 175 cm tall, and A1A1 carriers are 191 cm tall. Half the difference between the heights of the two homozygotes is a , which is 12 cm. The midpoint of the two homozygotes is 179 cm, which is also the intercept of body height within the population, that is, subtracting 179 from the three genotypic means scales the midpoint to zero. The deviation of the heterozygote from the midpoint (d) = -4 cm. The mean effect of this gene to the population mean is thus $12(0.6 - 0.4) + 2 * 0.6 * 0.4 * -4 = 0.48$ cm. To calculate the average effect of allele A1 (α_1) c, we sum the product of the conditional frequencies and genotypic values of the two possible genotypes, including the A1 allele. The two genotypes are A1A1 and A1A2, with genotypic values 12 and -4. Given one A1 allele, the frequency of A1A1 is 0.6 and of A1A2 is 0.4. Thus, $12 * 0.6 - 4 * 0.4 = 5.6$. We need to subtract the mean effect of this gene (0.48) from 5.6 to get the average effect of the A1 allele (α_1): $5.6 - 0.48 = 5.12$. Similarly, the average effect of the A2 allele (α_2) can be shown to equal -7.68. The breeding value of A1A1 carriers is the sum of the average effects of the two A1 alleles, which is $5.12 + 5.12 = 10.24$. Similarly, for A1A2 carriers this is $5.12 - 7.68 = 2.56$ and for A2A2 carriers this is $-7.68 - 7.68 = -15.36$. The genetic variance (V_G) related to this gene is 82.33, where V_A is 78.64 and V_D is 3.69

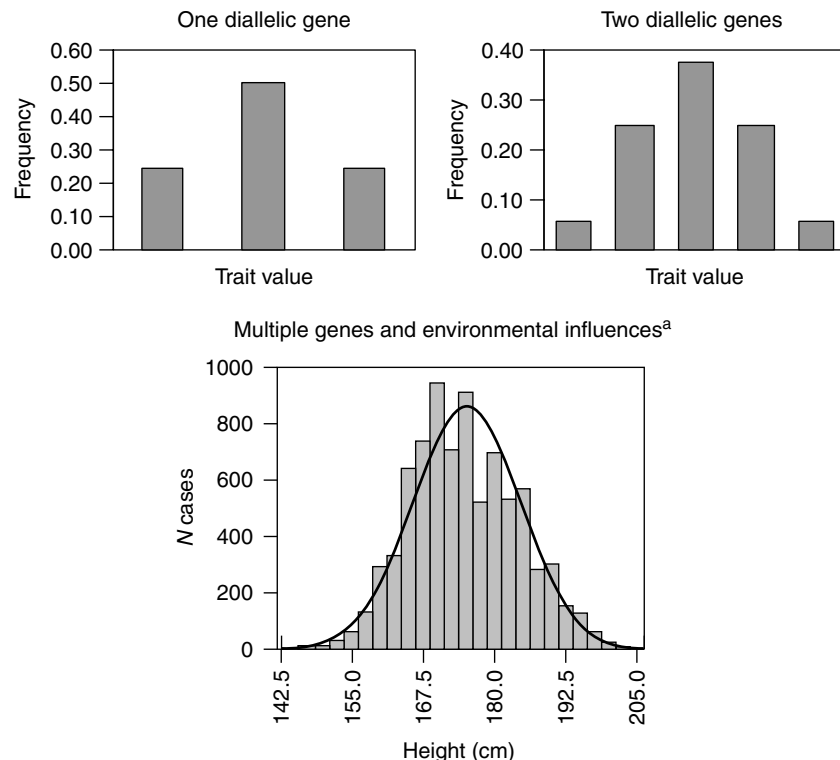


Figure 2 The combined discrete effects of many single genes result in continuous variation in the population. ^aBased on 8087 adult subjects from the Dutch Twin Registry (<http://www.tweelingenregister.org>)

parent and a blue-eyed parent is of course a consequence of the fact that parents transmit alleles to their offspring and not their genotypes. Therefore, parents cannot directly transmit their genotypic values a , d , and $-a$ to their offspring. To quantify the transmission of genetic effects from parents to offspring, and ultimately to decompose the observed variance in the offspring generation into genetic and environmental components, the concepts *average effect* and *breeding value* have been introduced [3].

Average effects are a function of genotypic values and allele frequencies within a population. The average effect of an allele is defined as ‘.. the mean deviation from the population mean of individuals which received that allele from one parent, the allele received from the other parent having come at random from the population’ [3]. To calculate the average effects denoted by α_1 and α_2 of alleles A1 and A2 respectively, we need to determine the frequency of the A1 (or A2) alleles in the genotypes of the offspring coming from a single parent. Again, we assume a single locus system with two alleles. If there is random mating between gametes carrying the A1 allele and gametes from the population, the frequency with which the A1 gamete unites with another gamete containing A1 (producing an A1A1 genotype in the offspring) equals p , and the frequency with which the gamete containing the A1 gamete unites with a gamete carrying A2 (producing an A1A2 genotype in the offspring) is q . The genotypic value of the genotype A1A1 in the offspring is a and the genotypic value of A1A2 in the offspring is d , as defined earlier. The mean value of the genotypes that can be produced by a gamete carrying the A1 allele equals the sum of the products of the frequency and the genotypic value. Or, in other terms, it is $pa + qd$. The average genetic effect of allele A1 (α_1) equals the deviation of the mean value of all possible genotypes that can be produced by gametes carrying the A1 allele from the population mean. The population mean has been derived earlier as $a(p - q) + 2pqd$ (1). The average effect of allele A1 is thus: $\alpha_1 = pa + qd - [a(p - q) + 2pqd] = q[a + d(q - p)]$. Similarly, the average effect of the A2 allele is $\alpha_2 = pd - qa - [a(p - q) + 2pqd] = -p[a + d(q - p)]$. $\alpha_1 - \alpha_2$ is known as α or the average effect of gene substitution. If there is no dominance, $\alpha_1 = qa$ and $\alpha_2 = -pa$, and the average effect of gene substitution α thus equals the genotypic value a ($\alpha = \alpha_1 - \alpha_2 = qa + pa = (q + p)a = a$).

The *breeding value* of an individual equals the sum of the average effects of gene substitution of an individual's alleles, and is therefore directly related to the mean genetic value of its offspring. Thus, the breeding value for an individual with genotype A1A1 is $2\alpha_1$ (or $2q\alpha$), for individuals with genotype A1A2 it is $\alpha_1 + \alpha_2$ (or $(q - p)\alpha$), and for individuals with genotype A2A2 it is $2\alpha_2$ (or $-2p\alpha$).

The breeding value is usually referred to as the *additive effect* of an allele (note that it includes both the values a and d), and differences between the genotypic effects (in terms of a , d , and $-a$, for genotypes A1A1, A1A2, A2A2 respectively) and the breeding values ($2q\alpha$, $(q - p)\alpha$, $-2p\alpha$, for genotypes A1A1, A1A2, A2A2 respectively) reflect the presence of dominance. Obviously, breeding values are of utmost importance to animal and crop breeders in determining which crossing will produce offspring with the highest milk yield, the fastest race horse, or the largest tomatoes.

Genetic Variance

Although until now we have ignored environmental effects, quantitative geneticists assume that populationwise the phenotype (P) is a function of both genetic (G) and environmental effects (E): $P = G + E$, where E refers to the environmental deviations, which have an expected average value of zero. By excluding the term $G \times E$, we assume no interaction between the genetic effects and the environmental effects (see **Gene-Environment Interaction**). If we also assume there is no covariance between G and E , the variance of the phenotype is given by $V_P = V_G + V_E$, where V_G represents the variance of the genotypic values of all contributing loci including both additive and nonadditive components, and V_E represents the variance of the environmental deviations. Statistically, the total genetic variance (V_G) can be obtained by applying the standard formula for the variance: $\sigma^2 = \sum f_i(x_i - \mu)^2$, where f_i denotes the frequency of genotype i , x_i denotes the corresponding genotypic mean of that genotype, and μ denotes the population mean, as calculated in (1). Thus, $V_G = p^2[a - (a(p - q) + 2pqd)]^2 + 2pq[d - (a(p - q) + 2pqd)]^2 + q^2[-a - (a(p - q) + 2pqd)]^2$. This can be simplified to $V_G = p^2[2q(a - dp)]^2 + 2pq[a(q - p) + d(1 - 2pq)]^2 + q^2[-2p(a + dq)]^2$, and further simplified to $V_G = 2pq[a + d(q - p)]^2 + (2pqd)^2 = V_A + V_D$ [3].

4 Additive Genetic Variance

If the phenotypic value of the heterozygous genotype lies midway between A1A1 and A2A2, the total genetic variance simplifies to $2pqa^2$. If d is not equal to zero, the ‘additive’ genetic variance component contains the effect of d . Even if $a = 0$, V_A is usually greater than zero (except when $p = q$). Thus, although V_A represents the variance due to the additive influences, it is not only a function of p , q , and a but also of d . Formally, V_A represents the variance of the breeding values, when these are expressed in terms of deviations from the population mean. The consequences are that, except in the rare situation in which all contributing loci are diallelic with $p = q$ and $a = 0$, V_A is usually greater than zero. Models that decompose the phenotypic variance into components of V_D , without including V_A , are therefore biologically implausible. When more than one locus is involved and it is assumed that the effects of these loci are uncorrelated and there is no interaction (i.e., no **epistasis**), the V_G ’s of each individual locus may be summed to obtain the total genetic variances of all loci that influence a trait [4, 5].

In most human quantitative genetic models, the observed variance of a trait is not modeled directly as a function of p , q , a , d , and environmental deviations (as all of these are usually unknown), but instead is modeled by comparing the observed resemblance between pairs of differential, known genetic

relatedness, such as monozygotic and dizygotic twin pairs (*see ACE Model*). Ultimately, p , q , a , d , and environmental deviations are the parameters that quantitative geneticists hope to ‘quantify’.

Acknowledgments

The author wishes to thank Eco de Geus and Dorret Boomsma for reading draft versions of this chapter.

References

- [1] Eiberg, H. & Mohr, J. (1987). Major genes of eye color and hair color linked to LU and SE, *Clinical Genetics* **31**(3), 186–191.
- [2] Eiberg, H. & Mohr, J. (1996). Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q, *European Journal of Human Genetics* **4**(4), 237–241.
- [3] Falconer, D.S. & Mackay, T.F.C. (1996). Introduction to Quantitative Genetics, Longan Group Ltd, Fourth Edition.
- [4] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh: Earth Sciences* **52**, 399–433.
- [5] Mather, K. (1949). *Biometrical Genetics*, Methuen, London.
- [6] Mather, K. & Jinks, J.L. (1982). *Biometrical Genetics*, Chapman & Hall, New York.

DANIELLE POSTHUMA

Additive Models

ROBERT J. VANDENBERG

Volume 1, pp. 22–24

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Additive Models

Although it may be found in the context of **experimental design** or **analysis of variance** (ANOVA) models, additivity or additive models is most commonly found in discussions of results from **multiple linear regression** analyses. Figure 1 is a reproduction of Cohen, Cohen, West, and Aiken's [1] graphical illustration of an additive model versus the same model but with an interaction present between their fictitious independent variables, X and Z , within the context of regression. Simply stated, additive models are ones in which there is no interaction between the independent variables, and in the case of the present illustration, this is defined by the following equation:

$$\hat{Y} = b_1X + b_2Z + b_0, \quad (1)$$

where $\hat{Y}$ is the predicted value of the dependent variable, b_1 is the regression coefficient for estimating Y from X (i.e., the change in Y per unit change in X), and similarly b_2 is the regression coefficient for estimating Y from Z . The intercept, b_0 , is a constant value to make adjustments for differences between X and Y units, and Z and Y units. Cohen et al. [1] use the following values to illustrate additivity:

$$\hat{Y} = 0.2X + 0.6Z + 2. \quad (2)$$

The point is that the regression coefficient for each independent variable (predictor) is constant over all values of the other independent variables in the model. Cohen et al. [1] illustrated this constancy using the example in Figure 1(a). The darkened lines in Figure 1(a) represent the regression of Y on X at each of three values of Z , two, five, and eight. Substituting the values in (2) for X (2, 4, 6, 8 and

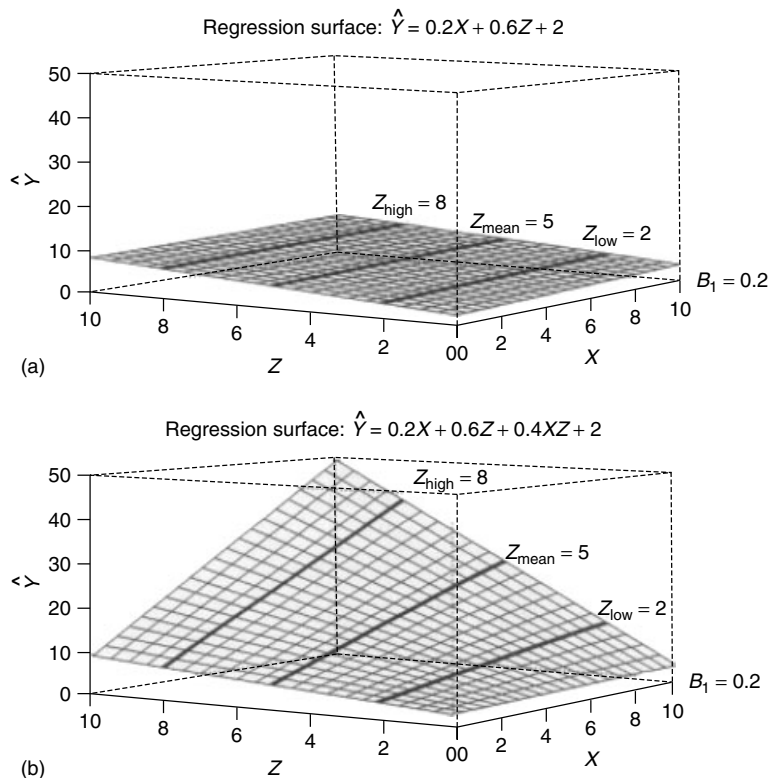


Figure 1 Additive versus interactive effects in regression contexts. Used with permission: Figure 7.1.1, p. 259 of Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah

2 Additive Models

10) along each of three values of Z will produce the darkened lines. These lines are parallel meaning that the regression of Y on X is constant over the values of Z . One may demonstrate this as well by holding values of X to two, five, and eight, and substituting all of the values of Z into (2). The only aspect of Figure 1(a) that varies is the height of the regression lines. There is a general upward displacement of the lines as Z increases.

Figure 1(b) is offered as a contrast. In this case, X and Z are presumed to have an interaction or joint effect that is above any additive effect of the variables. This is represented generally by

$$\hat{Y} = b_1X + b_2Z + b_3XZ + b_0 \quad (3)$$

and specifically for purposes of the illustration by

$$\hat{Y} = 0.2X + 0.6Z + 0.4XZ + 2. \quad (4)$$

Applying the same exercise used in (2) above would result in Figure 1(b). The point is that the regression of Y on X is not constant over the values of Z (and neither would the regression of Y on Z at values of X), but depends very much on the value of Z at which the regression of Y on X is calculated. This conditional effect is illustrated in Figure 1(b) by the angle of the plane representing the predicted values of Y at the joint of X and Z values.

As noted above, additive models are also considered in the context of experimental designs but much less frequently. The issue is exactly the same as in multiple regression, and is illustrated nicely by Charles Schmidt's graph which is reproduced in Figure 2. The major point of Figure 2 is that when there is no interaction between the independent variables (A and B in the figure), the main effects (additive effects) of each independent variable may be

Additivity assumption			
$R_{ij} = w_a a_i + w_b b_j$			
Example for a 2×2 design			
	A_1	A_2	$A_1 B_j - A_2 B_j$
B_1	$w_a a_1 + w_b b_1$	$w_a a_2 + w_b b_1$	$w_a(a_1 - a_2)$
B_2	$w_a a_1 + w_b b_2$	$w_a a_2 + w_b b_2$	$w_a(a_1 - a_2)$
	$w_b(b_1 - b_2)$	$w_b(b_1 - b_2)$	0
	$A_i B_1 - A_i B_2$		

Non-additivity assumption			
$R_{ij} = w_a a_i + w_b b_j + f(a_i, b_j)$			
Example for a 2×2 design			
	A_1	A_2	$A_1 B_j - A_2 B_j$
B_1	$w_a a_1 + w_b b_1 + f(a_1, b_1)$	$w_a a_2 + w_b b_1 + f(a_2, b_1)$	$w_a(a_1 - a_2) + [f(a_1, b_1) - f(a_2, b_1)]$
B_2	$w_a a_1 + w_b b_2 + f(a_1, b_2)$	$w_a a_2 + w_b b_2 + f(a_2, b_2)$	$w_a(a_1 - a_2) + [f(a_1, b_2) - f(a_2, b_2)]$
	$w_b(b_1 - b_2) + [f(a_1, b_1) - f(a_1, b_2)] - f(a_2, b_2)$	$w_b(b_1 - b_2) + [f(a_2, b_1) - f(a_2, b_2)]$	$[f(a_1, b_1) + f(a_2, b_2)] - [f(a_2, b_1) + f(a_1, b_2)]$
	$A_i B_1 - A_i B_2$		

Figure 2 Additive versus interactive effects in experimental designs. Used with permission: Professor Charles F. Schmidt, Rutgers University, http://www.rci.rutgers.edu/~cfs/305_html/MentalChron/MChronAdd.html

independently determined (shown in the top half of Figure 2). If, however, there is an interaction between the independent variables, then this joint effect needs to be accounted for in the analysis (illustrated by the gray components in the bottom half of Figure 2).

Reference

- [1] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the*

Behavioral Sciences, 3rd Edition, Lawrence Erlbaum, Mahwah.

Further Reading

Schmidt, C.F. (2003). <http://www.rci.rutgers.edu/~cfs/305.html/MentalChron/MChronAdd.html>

ROBERT J. VANDENBERG

Additive Tree

JAMES E. CORTER

Volume 1, pp. 24–25

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Additive Tree

Additive trees (also known as *path-length trees*) are often used to represent the proximities among a set of objects (see **Proximity Measures**). For example, Figure 1 shows an additive tree representing the similarities among seven Indo-European languages. The modeled proximities are the percentages of cognate terms between each pair of languages based on example data from Atkinson and Gray [1]. The additive tree gives a visual representation of the pattern of proximities, in which very similar languages are represented as neighbors in the tree.

Formally, an additive tree is a weighted tree graph, that is, a connected graph without cycles in which each arc is associated with a weight. In an additive tree, the weights represent the length of each arc. Additive trees are sometimes known as path-length trees, because the distance between any two points in an additive tree can be expressed as the sum of the lengths of the arcs in the (unique) path connecting the two points. For example, the tree distance between ‘English’ and ‘Swedish’ in Figure 1 is given by the sum of the lengths of the horizontal arcs in the path connecting them (the vertical lines in the diagram are merely to connect the tree arcs).

Distances in an additive tree satisfy the condition known as the *additive tree inequality*. This condition states that for any four objects a , b , c , and e ,

$$d(a, b) + d(c, e) \leq \max\{d(a, c) + d(b, e), d(a, e) + d(b, c)\}$$

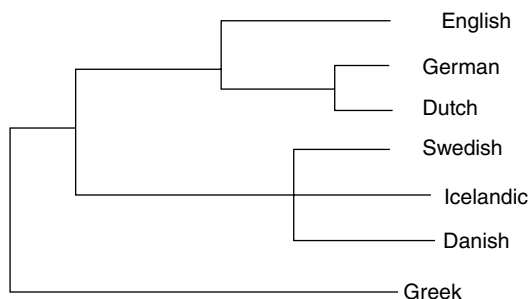


Figure 1 An additive tree representing the percentage of shared cognates between each pair of languages, for sample data on seven Indo-European languages

Alternatively, the condition may be stated as follows: if x and y , and u and v are relative neighbors in the tree (as in Figure 2(a)), then the six distances must satisfy the inequality

$$\begin{aligned} d(x, y) + d(u, v) &\leq d(x, u) + d(y, v) \\ &= d(x, v) + d(y, u). \end{aligned} \quad (1)$$

If the above inequality is restricted to be a double equality, the tree would have the degenerate structure shown in Figure 2(b). This structure is sometimes called a ‘bush’ or a ‘star’. The additive tree structure is very flexible and can represent even a one-dimensional structure (i.e., a line) as well as those in Figure 2 (as can be seen by imagining that the leaf arcs for objects x and v in Figure 2(a) shrank to zero length). The length of a leaf arc in an additive tree can represent how typical or atypical an object is within its cluster or within the entire set of objects. For example, objects x and v in Figure 2(a) are more typical (i.e., similar to other objects in the set) than are u and y .

The additive trees in Figure 2 are displayed in an unrooted form. In contrast, the additive tree in Figure 1 is displayed in a rooted form – that is, one point in the graph is picked, arbitrarily or otherwise, and that point is displayed as the leftmost point in the graph. Changing the root of an additive tree can change the apparent grouping of objects into clusters, hence the interpretation of the tree structure.

When additive trees are used to model behavioral data, which contains error as well as true structure, typically the best-fitting tree is sought. That is, a tree structure is sought such that distances in the tree approximate as closely as possible (usually in a least-squares sense) the observed dissimilarities among the modeled objects. Methods for fitting additive trees to errorful data

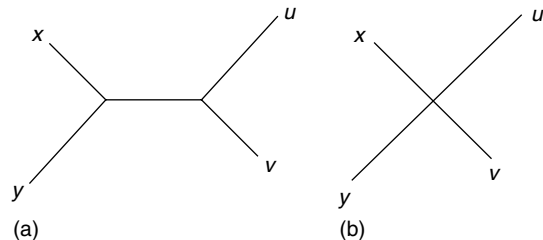


Figure 2 Two additive trees on four objects, displayed in unrooted form

2 Additive Tree

include those described in [2–6] and [7], the last method using a **maximum-likelihood** approach. A public-domain software program for fitting additive trees, GTREE [3], may be obtained at several sites, including <http://www.netlib.org/mds/> or <http://www.columbia.edu/~jec34/>. Routines implementing the algorithm of Hubert and Arabie [5] are also available (*see* http://ljhoff.psych.uiuc.edu/cda_toolbox/cda_toolbox_manual.pdf).

References

- [1] Atkinson, Q.D. & Gray, R.D. (2004). Are accurate dates an intractable problem for historical linguistics? in *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory*, C. Lipo, M. O'Brien, S. Shennan & M. Collard, eds, Aldine de Gruyter, New York.
- [2] Corter, J.E. (1982). ADDTREE/P: a PASCAL program for fitting additive trees based on Sattath and Tversky's ADDTREE algorithm, *Behavior Research Methods & Instrumentation* **14**, 353–354.
- [3] Corter, J.E. (1998). An efficient metric combinatorial algorithm for fitting additive trees, *Multivariate Behavioral Research* **33**, 249–271.
- [4] De Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data, *Psychometrika* **48**, 621–626.
- [5] Hubert, L. & Arabie, P. (1995). Iterative projection strategies for the least squares fitting of tree structures to proximity data, *British Journal of Mathematical & Statistical Psychology* **48**, 281–317.
- [6] Sattath, S. & Tversky, A. (1977). Additive similarity trees, *Psychometrika* **42**, 319–345.
- [7] Wedel, M. & Bijmolt, T.H.A. (2000). Mixed tree and spatial representations of dissimilarity judgments, *Journal of Classification* **17**, 243–271.

(See also **Hierarchical Clustering; Multidimensional Scaling**)

JAMES E. CORTER

Additivity Tests

GEORGE KARABATSOS

Volume 1, pp. 25–29

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Additivity Tests

When a test of additivity is performed on a set of data, the null hypothesis that a dependent variable is an additive, noninteractive function of two (or more) independent variables and the alternative hypothesis of nonadditivity are characterized by one or more interactions between the independent variables. If the dependent variable is on a quantitative (interval or ratio) scale (*see Measurement: Overview*), it is possible to perform a test of additivity in the context of **analysis of variance** (ANOVA). A more general test of additivity is achieved in the context of additive conjoint measurement theory. According to this theory, in order for additivity to hold on some monotonic transformation of the dependent variable, such that combinations of the independent variables are measurable on a common interval scale, it is necessary for data to be consistent with a hierarchy of (qualitative) cancellation axioms. The following two sections describe tests of additivity that are based on ANOVA and additive conjoint measurement, respectively.

Testing for Additivity, Assuming Quantitative Measurement

Suppose that IJ exchangeable sequences $\{Y_{1ij}, \dots, Y_{nij}, \dots, Y_{Nij}; i = 1, \dots, I, j = 1, \dots, J\}$ of data are observed from a two-factor experimental design, where Y_{nij} refers to the n th observation of a quantitative dependent variable in cell ij , corresponding to a level $i \in \{1, \dots, I\}$ of one independent variable and level $j \in \{1, \dots, J\}$ of a second independent variable. It is natural to model such data by a two-way ANOVA, given by

$$Y_{nij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{nij} \quad (1)$$

for all levels $i = 1, \dots, I$ and $j = 1, \dots, J$, where μ is the grand mean of the dependent variable, the population parameter α_i represents the main effect of level i , the parameter β_j is the main effect of level j , the parameter γ_{ij} is the interaction effect of levels i and j , and ε_{nij} is error assumed to be a random sample from a $N(0, \sigma_{\text{Error}}^2)$ normal distribution (*see Analysis of Variance; Multiple Linear Regression;*

Repeated Measures Analysis of Variance). In an ANOVA, the well-known F -statistic

$$F_{\text{Int}} = \frac{\sigma_{\text{Int}}^2}{\sigma_{\text{Error}}^2} = \frac{SS_{\text{Int}}/df_{\text{Int}}}{SS_{\text{Error}}/df_{\text{Error}}} \quad (2)$$

provides a test of the null hypothesis of additivity $H_0: \{\gamma_{ij} = 0, \forall i, j\}$ versus the alternative hypothesis of nonadditivity $H_1: \{\gamma_{ij} \neq 0, \text{ for some } i, j\}$. Under H_0 , statistic (2) follows an F distribution with $\{df_{\text{Int}}, df_{\text{Error}}\}$ degrees of freedom, where σ_{Int}^2 and SS_{Int} are the variance and sums-of-squares due to interaction, respectively, and SS_{Error} is the error sums-of-squares (e.g., [34]) (*see Catalogue of Probability Density Functions*). Under a chosen Type I error rate, the additive null hypothesis is rejected when the value of F_{Int} is unusually large. The F test (2) can be extended to test for interactions between three or more independent variables, and/or to test for interactions in two or more dependent variables (*see Multivariate Analysis of Variance*). Also, there are alternative tests of additivity, such as those [2, 3] based on the rank of the interaction matrix $\gamma = (\gamma_{ij})$, as well as distribution-free tests.

When there is exactly one observation per cell ij , the ANOVA model is saturated, with zero degrees of freedom left ($df_{\text{Error}} = 0$) to perform the F test of additivity. To circumvent a saturated model, several researchers have proposed testing the additivity hypothesis $H_0: \{\gamma_{ij} = 0, \forall i, j\}$, by restricting each of the interaction parameters (*see Interaction Effects*) by some specific function, under the nonadditive alternative hypothesis H_1 [7, 10, 11, 18, 22–25, 27, 28, 33, 35, 36]. For example, Tukey [36, 33] proposed testing $H_1: \{\gamma_{ij} = \lambda\alpha_i\beta_j \neq 0, \text{ some } i, j\}$, while Johnson and Graybill [11] proposed $H_1: \{\gamma_{ij} = \lambda\delta_i\xi_j \neq 0, \text{ some } i, j\}$, where the so-called ‘free’ parameters λ , δ_i , and ξ_j represent sources of interaction that are not due to main effects. Alternatively, Tusell [37] and Boik [4,5] proposed tests of additivity that do not require the data analyst to assume any particular functional form of interaction, and, in fact, they are sensitive in detecting many forms of nonadditivity [6].

Testing for Additivity, without Assuming Quantitative Measurement

Let an element of a (nonempty) product set $ax \in A_1 \times A_2$ denote the dependent variable that results

2 Additivity Tests

after combining the effect of level $a \in A_1 = \{a, b, c, \dots\}$ from one independent variable, and the effects of level $x \in A_2 = \{x, y, z, \dots\}$ from another independent variable. According to the theory of additive conjoint measurement [20], the effects of two independent variables are additive if and only if

$$ax \succsim by \text{ implies } f_1(a) + f_2(x) \geq f_1(b) + f_2(y) \quad (3)$$

holds for all $ax, by \in A_1 \times A_2$, where $\succsim$ denotes a weak order, and the functions f_1 and f_2 map the observed effects of the independent variables onto interval scales. In order for the additive representation (3) to hold for some monotonic transformation of the dependent variable, a hierarchy of cancellation axioms must be satisfied [17, 20, 26]. For example, single cancellation (often called order-independence) is satisfied when

$$ax \succsim bx \text{ if and only if } ay \succsim by \quad (4a)$$

$$ax \succsim ay \text{ if and only if } bx \succsim by \quad (4b)$$

hold for all $a, b \in A_1$ and all $x, y \in A_2$. Double cancellation is satisfied when

$$ay \succsim bx \text{ and } bz \succsim cy \text{ implies } az \succsim cx \quad (5)$$

holds for all $a, b, c \in A_1$ and all $x, y, z \in A_2$. The additive representation (3) and cancellation axioms can be extended to any number of independent variables [17], and, of course, an additive representation is unnecessary when all independent variables have zero effects.

In evaluating the fit of data to the cancellation axioms, many researchers have either counted the number of axiom violations or employed multiple nonparametric test statistics (e.g., [8, 9, 19, 26, 29–31, 38]) (see **Binomial Confidence Interval; Binomial Distribution; Estimating and Testing Parameters; Median; Kendall's Coefficient of Concordance; Kendall's Tau – τ**). Unfortunately, such approaches to testing additivity are not fully satisfactory. They assume that different tests of cancellation are statistically independent, which they are not. Also, as is well-known, the Type I error rate quickly increases with the number of statistical tests performed.

These statistical issues are addressed with a model-based approach to testing cancellation axioms. Suppose that $\{Y_{1k}, \dots, Y_{n_k}, \dots, Y_{N_k}; k = 1, \dots, m\}$

are m exchangeable sequences of N_k observations of a dependent variable, where Y is either a real-valued scalar or vector, and each sequence arises from some experimental condition $k \in \{1, \dots, m\}$. For example, $m = IJ$ conditions may be considered in a two-factor experimental design. According to **de Finetti's** representation theorem (e.g., [1]), the following Bayesian model describes the joint probability of m exchangeable sequences:

$$\begin{aligned} p(Y_{1k}, \dots, Y_{n_k}, \dots, Y_{N_k}; k = 1, \dots, m) \\ = \int_{\Theta \subseteq \Omega} \prod_{k=1}^m \prod_{n_k=1}^{N_k} p(Y_{n_k} | \Theta_k) p(\Theta_1, \dots, \Theta_k, \dots, \Theta_m) \\ \times d\Theta_1, \dots, \Theta_k, \dots, \Theta_m, \end{aligned} \quad (6)$$

where $p(Y_{n_k} | \Theta_k)$ is the sampling likelihood at data point Y_{n_k} given the k th population parameter Θ_k , and $p(\Theta_1, \dots, \Theta_k, \dots, \Theta_m)$ is the prior distribution over the parameter vector $\Theta = (\Theta_1, \dots, \Theta_k, \dots, \Theta_m)$ (see **Bayesian Statistics**). The notation $\Theta \subseteq \Omega$ refers to the fact that any set of cancellation axioms implies order-restrictions on the dependent variable (as shown in (4) and (5)), such that the parameter vector Θ is constrained to lie within a proper subset Ω of its total parameter space. The form of the constraint Ω depends on the set of cancellation axioms under consideration. A test of a set of cancellation axioms is achieved by testing the fit of a set of data $\{Y_{1k}, \dots, Y_{n_k}, \dots, Y_{N_k}; k = 1, \dots, m\}$ to the model in (6).

Karabatsos [12, 15, 16] implemented this approach for testing several cancellation axioms, in the case where, for $k = 1, \dots, m$, the dependent variable is dichotomous $Y_{n_k} \in \{0, 1\}$ and Θ_k is a binomial parameter. For example, Karabatsos [12] tested single cancellation (4) by evaluating the fit of dichotomous data $\{Y_{1k}, \dots, Y_{n_k}, \dots, Y_{N_k}; k = 1, \dots, m\}$ to the model in (6), where all $m = IJ$ binomial parameters were subject to the constraint Ω that $\Theta_{ij} \leq \Theta_{i+1,j}$ for all ordered levels $i = 1, \dots, I - 1$ of the first independent variable and $\Theta_{ij} \leq \Theta_{i,j+1}$ for all ordered levels $j = 1, \dots, J - 1$ of the second independent variable.

Karabatsos [14] later generalized this binomial approach, by considering a vector of multinomial parameters $\Theta_k = (\theta_{1k}, \dots, \theta_{rk}, \dots, \theta_{Rk})$ for each experimental condition $k = 1, \dots, m$, where θ_{rk} refers to the probability of the r th response pattern. Each response pattern is characterized by a particular

weak order defined over all elements of $A_1 \times A_2$. In this context, Ω refers to the sum-constraint $\sum_{r_k \in \sim V_k} \theta_{rk} \geq C$ for each experimental condition k , and some chosen threshold $C \in [1/2, 1]$, where $\sim V_k$ is the set of response patterns that do not violate a given cancellation axiom.

Karabatsos [13] proposed a slightly different multinomial model, as a basis for a Bayesian bootstrap [32] approach to isotonic (inequality-constrained) regression (see **Bootstrap Inference**). This procedure can be used to estimate the non-parametric posterior distribution of a discrete- or continuous-valued dependent variable Y , subject to the order-constraints of the set of all possible linear orders (for example, $Y_1 \leq Y_2 \leq \dots \leq Y_k \leq \dots \leq Y_m$) that satisfy the *entire hierarchy* of cancellation axioms. Here, a test of additivity is achieved by evaluating the fit of the observed data $\{Y_{1k}, \dots, Y_{nk}, \dots, Y_{Nk}; k = 1, \dots, m\}$ to the corresponding order-constrained posterior distribution of Y .

Earlier, as a non-Bayesian approach to additivity testing, Macdonald [21] proposed isotonic regression to determine the least-squares **maximum-likelihood estimate** (MLE) of the dependent variable $\{\hat{Y}_k; k = 1, \dots, m\}$, subject to a linear order-constraint (e.g., $Y_1 \leq Y_2 \leq \dots \leq Y_k \leq \dots \leq Y_m$) that satisfies a given cancellation axiom (see **Least Squares Estimation**). He advocated testing each cancellation axiom separately, by evaluating the fit of the observed data $\{Y_{1k}, \dots, Y_{nk}, \dots, Y_{Nk}; k = 1, \dots, m\}$ to the MLE $\{\hat{Y}_k; k = 1, \dots, m\}$ under the corresponding axiom.

Acknowledgments

Karabatsos's research is supported by National Science Foundation grant SES-0242030, program of Methodology, Measurement, and Statistics. Also, this work is supported in part by Spencer Foundation grant SG2001000020.

References

- [1] Bernardo, J.M. (2002). *Bayesian Theory* (second reprint), Wiley, New York.
- [2] Boik, R.J. (1986). Testing the rank of a matrix with applications to the analysis of interactions in ANOVA, *Journal of the American Statistical Association* **81**, 243–248.
- [3] Boik, R.J. (1989). Reduced-rank models for interaction in unequally-replicated two-way classifications, *Journal of Multivariate Analysis* **28**, 69–87.
- [4] Boik, R.J. (1990). Inference on covariance matrices under rank restrictions, *Journal of Multivariate Analysis* **33**, 230–246.
- [5] Boik, R.J. (1993a). Testing additivity in two-way classifications with no replications: the locally best invariant test, *Journal of Applied Statistics* **20**, 41–55.
- [6] Boik, R.J. (1993b). A comparison of three invariant tests of additivity in two-way classifications with no replications, *Computational Statistics and Data Analysis* **15**, 411–424.
- [7] Boik, R.J. & Marasinghe, M.G. (1989). Analysis of non-additive multiway classifications, *Journal of the American Statistical Association* **84**, 1059–1064.
- [8] Coombs, C.H. & Huang, L.C. (1970). Polynomial psychophysics of risk, *Journal of Mathematical Psychology* **7**, 317–338.
- [9] Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation, *Psychological Review* **83**, 65–79.
- [10] Harter, H.L. & Lum, M.D. (1962). An interpretation and extension of Tukey's one-degree of freedom for non-additivity, *Aeronautical Research Laboratory Technical Report, ARL*, pp. 62–313.
- [11] Johnson, D.E. & Graybill, F.A. (1972). An analysis of a two-way model with interaction and no replication, *Journal of the American Statistical Association* **67**, 862–868.
- [12] Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory, *Journal of Applied Measurement* **2**, 389–423.
- [13] Karabatsos, G. (2004a). *A Bayesian Bootstrap Approach To Testing The Axioms Of Additive Conjoint Measurement*. Manuscript under review.
- [14] Karabatsos, G. (2004b). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. To appear, *Journal of Mathematical Psychology*.
- [15] Karabatsos, G. & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of non-parametric item-response theory, *Applied Psychological Measurement* **2**, 110–125.
- [16] Karabatsos, G. & Ullrich, J.R. (2002). Enumerating and testing conjoint measurement models, *Mathematical Social Sciences* **43**, 485–504.
- [17] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement: Additive and Polynomial Representations*, Academic Press, New York.
- [18] Krishnaiah, P.R. & Yochmowitz, M.G. (1980). Inference on the structure of interaction in the two-way classification model, in *Handbook of Statistics*, Vol. 1, P.R. Krishnaiah, ed., North Holland, Amsterdam, pp. 973–984.
- [19] Levelt, W.J.M., Riemersma, J.B. & Bunt, A.A. (1972). Binaural additivity of loudness, *British Journal of Mathematical and Statistical Psychology* **25**, 51–68.
- [20] Luce, R.D. & Tukey, J.W. (1964). Additive conjoint measurement: a new type of fundamental measurement, *Journal of Mathematical Psychology* **1**, 1–27.

4 Additivity Tests

- [21] Macdonald, R.R. (1984). Isotonic regression analysis and additivity, in *Trends in Mathematical Psychology*, E. Degreef & J. Buggenhaut, eds, Elsevier Science Publishers, North Holland, pp. 239–255.
- [22] Mandel, J. (1961). Non-additivity in two-way analysis of variance, *Journal of the American Statistical Association* **56**, 878–888.
- [23] Mandel, J. (1969). The partitioning of interaction in analysis of variance, *Journal of Research, National Bureau of Standards* **B 73B**, 309–328.
- [24] Mandel, J. (1971). A new analysis of variance model for non-additive data, *Technometrics* **13**, 1–18.
- [25] Marasinghe, M.G. & Boik, R.J. (1993). A three-degree of freedom test of additivity in three-way classifications, *Computational Statistics and Data Analysis* **16**, 47–61.
- [26] Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*, Lawrence Earlbaum Associates, Hillsdale.
- [27] Milliken, G.A. & Graybill, F.A. (1970). Extensions of the general linear hypothesis model, *Journal of the American Statistical Association* **65**, 797–807.
- [28] Milliken, G.A. & Johnson, D.E. (1989). *Analysis of Messy Data*, Vol. 2, Van Nostrand Reinhold, New York.
- [29] Nygren, T.E. (1985). An examination of conditional violations of axioms for additive conjoint measurement, *Applied Psychological Measurement* **9**, 249–264.
- [30] Nygren, T.E. (1986). A two-stage algorithm for assessing violations of additivity via axiomatic and numerical conjoint analysis, *Psychometrika* **51**, 483–491.
- [31] Perline, R., Wright, B.D. & Wainer, H. (1979). The Rasch model as additive conjoint measurement, *Applied Psychological Measurement* **3**, 237–255.
- [32] Rubin, D.B. (1981). The Bayesian bootstrap, *Annals of Statistics* **9**, 130–134.
- [33] Scheffé, H. (1959). *The Analysis of Variance* (Sixth Printing), Wiley, New York.
- [34] Searle, S.R. (1971). *Linear Models*, John Wiley & Sons, New York.
- [35] Tukey, J.W. (1949). One degree of freedom for non-additivity, *Biometrics* **5**, 232–242.
- [36] Tukey, J.W. (1962). The future of data analysis, *Annals of Mathematical Statistics* **33**, 1–67.
- [37] Tusell, F. (1990). Testing for interaction in two-way ANOVA tables with no replication, *Computational Statistics and Data Analysis* **10**, 29–45.
- [38] Tversky, A. (1967). Additivity, utility, and subjective probability, *Journal of Mathematical Psychology* **4**, 175–201.

GEORGE KARABATSOS

Adoption Studies

MICHAEL C. NEALE

Volume 1, pp. 29–33

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Adoption Studies

Adoption usually refers to the rearing of a nonbiological child in a family. This practice is commonplace after wars, which leave many children orphaned, and is moderately frequent in peacetime. Approximately 2% of US citizens are adoptees.

Historically, adoption studies have played a prominent role in the assessment of genetic variation in human and animal traits [10]. Most early studies focused on cognitive abilities [9], but there is now greater emphasis on psychopathology [5] and physical characteristics, such as body mass [20]. Adoption studies have made major substantive contributions to these areas, identifying the effects of genetic factors where they were previously thought to be absent [3, 12, 15].

In recent years, the adoption study has been overshadowed by the much more popular twin study [17] (*see Twin Designs*). Part of this shift may be due to the convenience of twin studies and the complex ethical and legal issues involved in the ascertainment and sampling of adoptees. Certain Scandinavian countries – especially Denmark, Sweden, and Finland [8, 13, 14] – maintain centralized databases of adoptions, and, thus, have been able to mount more representative and larger adoption studies than elsewhere.

The adoption study is a ‘natural experiment’ that mirrors cross-fostering designs used in genetic studies of animals, and, therefore, has a high face validity as a method to resolve the effects of genes and environment on individual differences. Unfortunately, the adoption study also has many methodological difficulties. First is the need to maintain confidentiality, which can be a problem even at initial ascertainment, as some adoptees do not know that they are adopted. Recent legal battles for custody fought between biological and adoptive parents make this a more critical issue than ever. Secondly, in many substantive areas, for example, psychopathology, there are problems with sampling, in that neither the biological nor the adoptive parents can be assumed to be a random sample of parents in the population. For example, poverty and its sequel may be more common among biological parents who have their children adopted into other families than among parents who rear their children themselves. Conversely, prospective adoptive parents are, on average, and through self-selection, older and less fertile than biological parents. In addition, they

are often carefully screened by adoption agencies, and may be of higher socioeconomic status than nonadoptive parents. Statistical methods (see below) may be used to control for these sampling biases if a random sample of parents is available. Some studies indicate that adoptive and biological parents are quite representative of the general population for demographic characteristics and cognitive abilities [19], so this potential source of bias may not have substantially affected study results.

Thirdly, selective placement is a common methodological difficulty. For statistical purposes, the ideal adoption study would have randomly selected adoptees placed at random into randomly selected families in the population. Often, there is a partial **matching** of the characteristics of the adoptee (e.g., hair and eye color, religion and ethnicity) to those of the adopting family. This common practice may improve the chances of successful adoption. Statistically, it is necessary to control for the matching as far as possible. Ideally, the matching characteristics used should be recorded and modeled. Usually, such detailed information is not available, so matching is assumed to be based on the variables being studied and modeled accordingly (see below). In modern adoption studies, these methods are used often [18, 19].

Types of Adoption Study

Nuclear families in which at least one member is not biologically related to the others offer a number of potential comparisons that can be genetically informative (see Table 1). Of special note are monozygotic (MZ) twins reared apart (MZ_A). Placed into uncorrelated environments, the correlation between MZ twins directly estimates the proportion of variation due to all genetic sources of variance (‘broad **heritability**’). Estimation of heritability in this way is statistically much more powerful than, for example, the classical twin study that compares MZ and dizygotic (DZ) twins reared together (MZ_T and DZ_T). With MZ_A twins, the test for heritability is a test of the null hypothesis that the correlation is zero, whereas the comparison of MZ_T and DZ_T is a test of a difference between correlations (*see Correlation Issues in Genetics Research*). Environmental effects shared by members of a twin pair (known as ‘common’, ‘shared’ or ‘family’ environment or ‘C’) are excluded by design. If this source of variation is of

2 Adoption Studies

Table 1 Coefficients of genetic and environmental variance components quantifying resemblance between adopted and biological relatives, assuming random sampling, mating and placement

Relationship	Variance component						
	V_A	V_D	V_{AA}	V_{AD}	V_{DD}	E_S	E_P
BP-BC	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0	1
BP-AC	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0	0
AP-AC	0	0	0	0	0	0	1
AC-BC	0	0	0	0	0	1	0
BC-BC _T	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	1	0
BC-BC _A	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	0	0
MZ _T -MZ _T	1	1	1	1	1	1	0
MZ _A -MZ _A	1	1	1	1	1	0	0

Note: V_A – additive genetic; V_D – dominance genetic; V_{AA} – additive $\times$ additive interaction; V_{AD} – additive $\times$ dominance interaction; V_{DD} – dominance $\times$ dominance interaction; E_S – environment shared by siblings; E_P – environment shared or transmitted between parent and child. Relationships are: MZ – monozygotic twin; DZ – dizygotic twin; BP – biological parent; BC – biological child; AP – adoptive parent; AC – adopted child. The subscripts T and A refer to reared together and reared apart, respectively.

interest, then additional groups of relatives, such as unrelated individuals reared together, are needed to estimate it. Similar arguments may be made about across-generational sources of resemblance. Heath and Eaves [11] compared the power to detect genetic and environmental transmission across several twin-family (twins and their parents or twins and their children) adoption designs.

Methods of Analysis

Most modern adoption study data are analyzed with **Structural Equation Models** (SEM) [2, 17]. SEM is an extension of **multiple linear regression** analysis that involves two types of variable: *observed variables* that have been measured, and **latent variables** that have not. Two variables may be specified as causally related or simply correlated from unspecified effects. It is common practice to represent the variables and their relationships in a path diagram (see **Path Analysis and Path Diagrams**), where single-headed arrows indicate causal relationships, and double-headed arrows represent correlations. By convention, observed variables are shown as squares and latent variables are shown as circles.

Figure 1 shows the genetic and environmental transmission from biological and adoptive parents to

three children. Two of the children are offspring of the biological parents (siblings reared together), while the third is adopted. This diagram may also be considered as multivariate, allowing for the joint analysis of multiple traits. Each box and circle then represents a vector of observed variables. Multivariate analyses (see **Multivariate Analysis: Overview**) are particularly important when studying the relationship between parental attributes and outcomes in their offspring. For example, harsh parenting may lead to psychiatric disorders. Both variables should be studied in a multivariate genetically informative design such as an adoption or twin study to distinguish between the possible direct and indirect genetic and environmental pathways.

From the rules of path analysis [22, 23] we can derive predicted covariances among the relatives in terms of the parameters of the model in Figure 1. These expectations may, in turn, be used in a structural equation modeling program such as Mx [16] (see **Software for Behavioral Genetics**) to estimate the parameters using **maximum likelihood** or some other goodness-of-fit function. Often, simpler models than the one shown will be adequate to account for a particular set of data.

A special feature of the diagram in Figure 1 is the dotted lines representing delta-paths [21].

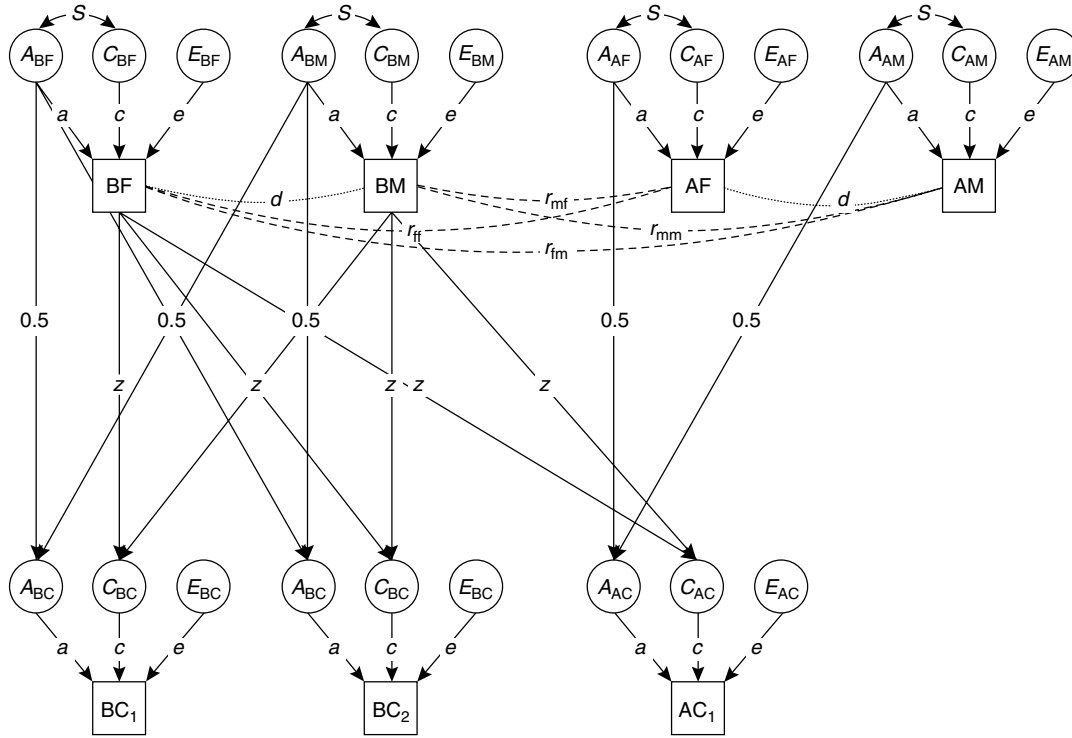


Figure 1 Path diagram showing sources of variation and covariance between: adoptive mother, AM; adoptive father, AF; their own biological children, BC_1 and BC_2 ; a child adopted into their family, AC_1 ; and the adopted child's biological parents, BF and BM

These represent the effects of two possible types of selection: assortative mating, in which husband and wife correlate; and selective placement, in which the adoptive and biological parents are not paired at random. The effects of these processes may be deduced from the Pearson–Aitken selection formulas [1]. These formulas are derived from linear regression under the assumptions of multivariate linearity and homoscedasticity. If we partition the variables into selected variables, X_S , and unselected variables X_N , then it can be shown that changes in the covariance of X_S lead to changes in covariances among X_N and the cross-covariances (X_S with X_N). Let the original (preselection) covariance matrix of X_S be $\mathbf{A}$, the original covariance matrix of X_N be $\mathbf{C}$, and the covariance between X_N and X_S be $\mathbf{B}$. The preselection matrix may be written

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B}' & \mathbf{C} \end{array} \right).$$

If selection transforms $\mathbf{A}$ to $\mathbf{D}$, then the new covariance matrix is given by

$$\left(\begin{array}{c|c} \mathbf{D} & \mathbf{DA}^{-1}\mathbf{B} \\ \hline \mathbf{B}'\mathbf{A}^{-1}\mathbf{D} & \mathbf{C} - \mathbf{B}'(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{DA}^{-1})\mathbf{B} \end{array} \right).$$

Similarly, if the original means are $(\mathbf{x}_S; \mathbf{x}_N)'$ and selection modifies $\mathbf{x}_S$ to $\tilde{\mathbf{x}}_S$, then the vector of means after selection is given by

$$[\mathbf{x}_S; \mathbf{x}_N + \mathbf{A}^{-1}\mathbf{B}(\mathbf{x}_S - \tilde{\mathbf{x}}_S)]'.$$

These formulas can be applied to the covariance structure of all the variables in Figure 1. First, the formulas are applied to derive the effects of assortative mating, and, secondly, they are applied to derive the effects of selective placement. In both cases, only the covariances are affected, not the means. An interesting third possibility would be to control for the effects of nonrandom selection of the

biological and adoptive relatives, which may well change both the means and the covariances.

Selected Samples

A common approach in adoption studies is to identify members of adoptive families who have a particular disorder, and then examine the rates of this disorder in their relatives. These rates are compared with those from control samples. Two common starting points for this type of study are (a) the adoptees (the adoptees' families method), and (b) the biological parents (the adoptees study method). For rare disorders, this use of selected samples may be the only practical way to assess the impact of genetic and environmental factors.

One limitation of this type of method is that it focuses on one disorder, and is of limited use for examining **comorbidity** between disorders. This limitation is in contrast to the population-based sampling approach, where many characteristics – and their covariances or comorbidity – can be explored simultaneously.

A second methodological difficulty is that ascertained samples of the disordered adoptees or parents may not be representative of the population. For example, those attending a clinic may be more severe or have different risk factors than those in the general population who also meet criteria for diagnosis, but do not attend the clinic.

Genotype × Environment Interaction

The natural experiment of an adoption study provides a straightforward way to test for **gene–environment interaction**. In the case of a continuous phenotype, interaction may be detected with linear regression on

1. the mean of the biological parents' phenotypes (which directly estimates heritability)
2. the mean of the adoptive parents' phenotypes
3. the product of points 1 and 2.

Significance of the third term would indicate significant $G \times E$ interaction. With binary data such as psychiatric diagnoses, the rate in adoptees may be compared between subjects with biological or adoptive parents affected, versus both affected. $G \times E$

interaction has been found for alcoholism [7] and substance abuse [6].

Logistic regression is a popular method to test for genetic and environmental effects and their interaction on binary outcomes such as psychiatric diagnoses. These analyses lack the precision that structural equation modeling can bring to testing and quantifying specific hypotheses, but offer a practical method of analysis for binary data. Analysis of binary data can be difficult within the framework of SEM, requiring either very large sample sizes for asymptotic weighted least squares [4], or integration of the multivariate normal distribution (*see Catalogue of Probability Density Functions*) over as many dimensions as there are relatives in the pedigree, which is numerically intensive.

References

- [1] Aitken, A.C. (1934). Note on selection from a multivariate normal population, *Proceedings of the Edinburgh Mathematical Society, Series B* **4**, 106–110.
- [2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [3] Bouchard Jr, T.J. & McGue, M. (1981). Familial studies of intelligence: a review, *Science* **212**, 1055–1059.
- [4] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.
- [5] Cadoret, R.J. (1978). Psychopathology in adopted-away offspring of biologic parents with antisocial behavior, *Archives of General Psychiatry* **35**, 176–184.
- [6] Cadoret, R.J., Troughton, E., O'Gorman, T.W. & Heywood, E. (1986). An adoption study of genetic and environmental factors in drug abuse, *Archives of General Psychiatry* **43**, 1131–1136.
- [7] Cloninger, C.R., Bohman, M. & Sigvardsson, S. (1981). Inheritance of alcohol abuse: cross-fostering analysis of adopted men, *Archives of General Psychiatry* **38**, 861–868.
- [8] Cloninger, C.R., Bohman, M., Sigvardsson, S. & von Knorring, A.L. (1985). Psychopathology in adopted-out children of alcoholics: the Stockholm adoption study, in *Recent Developments in Alcoholism*, Vol. 3, M. Galanter, ed., Plenum Press, New York, pp. 37–51.
- [9] DeFries, J.C. & Plomin, R. (1978). Behavioral genetics, *Annual Review of Psychology* **29**, 473–515.
- [10] Fuller, J.L. & Thompson, W.R. (1978). *Foundations of Behavior Genetics*, Mosby, St. Louis.
- [11] Heath, A.C. & Eaves, L.J. (1985). Resolving the effects of phenotype and social background on mate selection, *Behavior Genetics* **15**, 15–30.

-
- [12] Heston, L.L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers, *British Journal of Psychiatry* **112**, 819–825.
- [13] Kaprio, J., Koskenvuo, M. & Langinvainio, H. (1984). Finnish twins reared apart: smoking and drinking habits. Preliminary analysis of the effect of heredity and environment, *Acta Geneticae Medicae et Gemellologiae* **33**, 425–433.
- [14] Kety, S.S. (1987). The significance of genetic factors in the etiology of schizophrenia: results from the national study of adoptees in Denmark, *Journal of Psychiatric Research* **21**, 423–429.
- [15] Mendlewicz, J. & Rainer, J.D. (1977). Adoption study supporting genetic transmission in manic-depressive illness, *Nature* **268**, 327–329.
- [16] Neale, M.C. (1995). *Mx: Statistical Modeling*, 3rd Edition, Box 980126 MCV, Richmond, p. 23298.
- [17] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, Boston.
- [18] Phillips, K. & Fulker, D.W. (1989). Quantitative genetic analysis of longitudinal trends in adoption designs with application to IQ in the Colorado adoption project, *Behavior Genetics* **19**, 621–658.
- [19] Plomin, R. & DeFries, J.C. (1990). *Behavioral Genetics: A Primer*, 2nd Edition, Freeman, New York.
- [20] Sorensen, T.I. (1995). The genetics of obesity, *Metabolism* **44**, 4–6.
- [21] Van Eerdewegh, P. (1982). *Statistical selection in multivariate systems with applications in quantitative genetics*, Ph.D. thesis, Washington University.
- [22] Vogler, G.P. (1985). Multivariate path analysis of familial resemblance, *Genetic Epidemiology* **2**, 35–53.
- [23] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.

MICHAEL C. NEALE

Age–Period–Cohort Analysis

THEODORE HOLFORD

Volume 1, pp. 33–38

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Age–Period–Cohort Analysis

To determine the effect of time on a particular outcome for an individual, it is essential to understand the relevant temporal perspective. Age affects many aspects of life, including the risk of disease, so this is an essential component of any analysis of time trends. Period denotes the date of the outcome, and if the outcome varies with period it is likely due to some underlying factor that affects the outcome, and varies in the same way for the entire population regardless of age. Cohort, on the other hand, refers to generational effects caused by factors that only affect particular age groups when the level of the factor changes with time.

An example of a period effect would be the potential influence of an air contaminant that affected all age groups in the same way. If the level of exposure to that factor increased/decreased with time, exerting a change in the outcome in all age groups, then we would expect a related pattern across all age groups in the study. In studies that take place over long periods of time, the technology for measuring the outcome may change, giving rise to an artifactual effect that was not due to change in exposure to a causative agent. For example, intensive screening for disease can identify disease cases that would not previously have been identified, thus artificially increasing the disease rate in a population that has had no change in exposure over time (*see Cohort Studies*).

Cohort, sometimes called *birth cohort*, effects may be due to factors related to exposures associated with the date of birth, such as the introduction of a particular drug or practice during pregnancy. For example, a pregnancy practice associated with increased disease risk and adopted by the population of mothers during a particular time period could affect the risk during the life span of the entire generation born during that period. While it is common to refer to these effects as being associated with year of birth, they could also be the result of changes in exposure that occurred after birth. In many individuals, lifestyle factors tend to become fixed as their generation approaches adulthood. The quantification of these generational effects is referred to as *cohort effects*. To illustrate, consider a study of trends in suicide

rates in Japan which discovered that the generation of men who enlisted to fight in World War II had a lifelong increased risk of suicide when compared to other generations [6]. Presumably, these generational trends were due to experiences well after birth.

As an example of the alternative temporal perspectives, consider the birth rate, that is, the average number of births in a year per 1000 women for blacks by age during the period 1980–2002. These are analyzed for five-year age and period intervals. The final period only covers the years 2000–2002 because more recent data were not available, but we will assume that these rates are representative of the rates for 2000–2004. Figure 1 shows a graph of the age trends in the birth rates for each of the periods, and the vertical axis employs a log scale. The rate for a given age interval is plotted in the center of that interval (e.g., the point at age = 17.5 represents the interval from 15 to 20). A cohort may be identified for each age-period (age interval 15–20 and period interval 1980–1985), and the range in this case may be 10 years. For example, the earliest cohort for the first age-period would be someone nearly 20 at the beginning of 1980, who would have been born in 1960. The latest cohort would have just turned 15 at the end of 1984, and would have been born at the end of 1969. In general, the cohort interval is twice as long as the age and period intervals when the latter two are equal. In addition, the cohort intervals overlap, because the next cohort in our example would include individuals born from 1965 to 1974. Figure 2 shows a semilog plot of the age-specific birth rates for the different cohorts. A subtle difference between Figures 1 and 2 is that the segments in the cohort plot tend to be more nearly parallel, as can be seen in the lines connecting the rates for age 17.5 with 22.5. In the age-period-cohort modeling framework, the temporal factor that achieves the greatest degree of parallelism tends to be the predominant factor in the model, and this is identified in a more formal way by assessing statistical significance.

An inherent redundancy among these three temporal factors arises from the fact that knowing any two factors implies the value of the third. For example, if we know an individual's age (a) at a given date or period (p), then the cohort is the difference, ($c = p - a$). This linear dependence gives rise to an identifiability problem in a formal regression model that attempts to obtain quantitative estimates

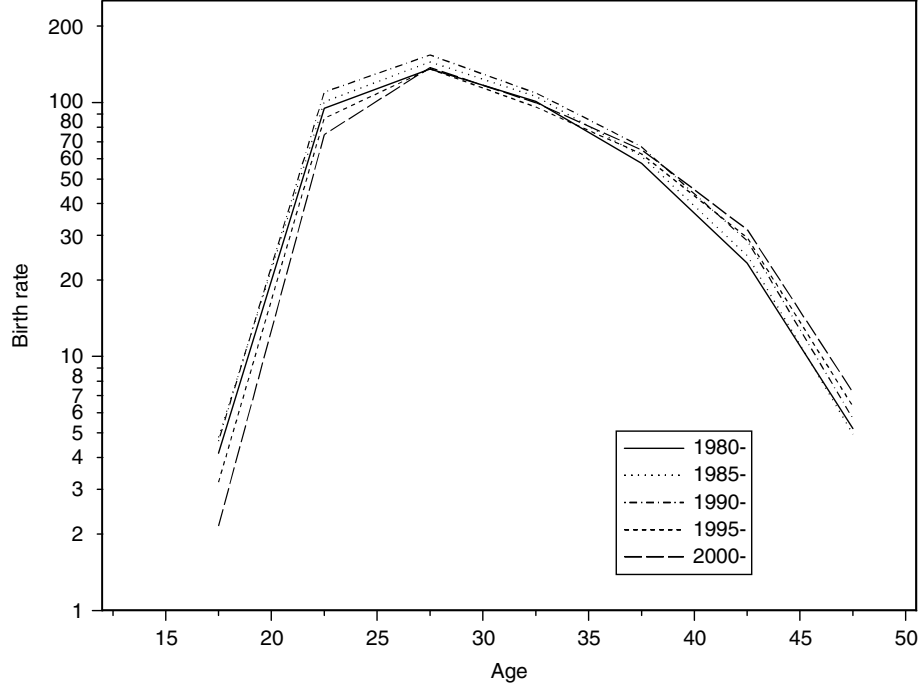


Figure 1 Period plot: a semilog plot of birth rates among US black women by age and period

of regression parameters associated with each temporal element. Suppose that the expected value of the outcome, Y (the log birth rate in our example) is linearly related to the temporal factors,

$$E[Y] = \beta_0 + a\beta_a + p\beta_p + c\beta_c. \quad (1)$$

Using the linear relationship between the temporal factors gives rise to

$$\begin{aligned} E[Y] &= \beta_0 + a\beta_a + p\beta_p + (p - a)\beta_c \\ &= \beta_0 + a(\beta_a - \beta_c) + p(\beta_p + \beta_c), \end{aligned} \quad (2)$$

which has only two identifiable parameters besides the intercept instead of the expected three. Another way of visualizing this phenomenon is that all combinations of age, period and cohort may be displayed in the Lexis diagram shown in Figure 3, which is obviously a representation of a two dimensional plane instead of the three dimensions expected for three separate factors.

In general, these analyses are not limited to linear effects applied to continuous measures of time, but instead they are applied to temporal intervals, such as mortality rates observed for five- or ten-year intervals

of age and period. When the widths of these intervals are equal, the model may be expressed as

$$E[Y_{ijk}] = \mu + \alpha_i + \pi_j + \gamma_k, \quad (3)$$

where μ is the intercept, α_i the effect of age for the i th ($i = 1, \dots, I$) interval, π_j the effect of period for the j th ($j = 1, \dots, J$) interval, and γ_k the effect of the k th cohort ($k = j - i + I = 1, \dots, K = I + J - 1$). The usual constraints in this model imply that $\sum \alpha_i = \sum \pi_j = \sum \gamma_k = 0$. The identifiability problem manifests itself through a single unidentifiable parameter [3], which can be more easily seen if we partition each temporal effect into components of overall linear trend, and curvature or departure from linear trend. For example, age can be given by $\alpha_i = i'\beta_\alpha + \tilde{\alpha}_i$, where $i' = i - 0.5(I + 1)$, β_α is the overall slope and $\tilde{\alpha}_i$ the curvature. The overall model can be expressed as

$$\begin{aligned} E[Y_{ijk}] &= \mu + (i'\beta_\alpha + \tilde{\alpha}_i) + (j'\beta_\pi + \tilde{\pi}_j) \\ &\quad + (k'\beta_\gamma + \tilde{\gamma}_k) \\ &= \mu + i'(\beta_\alpha - \beta_\gamma) + j'(\beta_\pi + \beta_\gamma) \\ &\quad + \tilde{\alpha}_i + \tilde{\pi}_j + \tilde{\gamma}_k, \end{aligned} \quad (4)$$

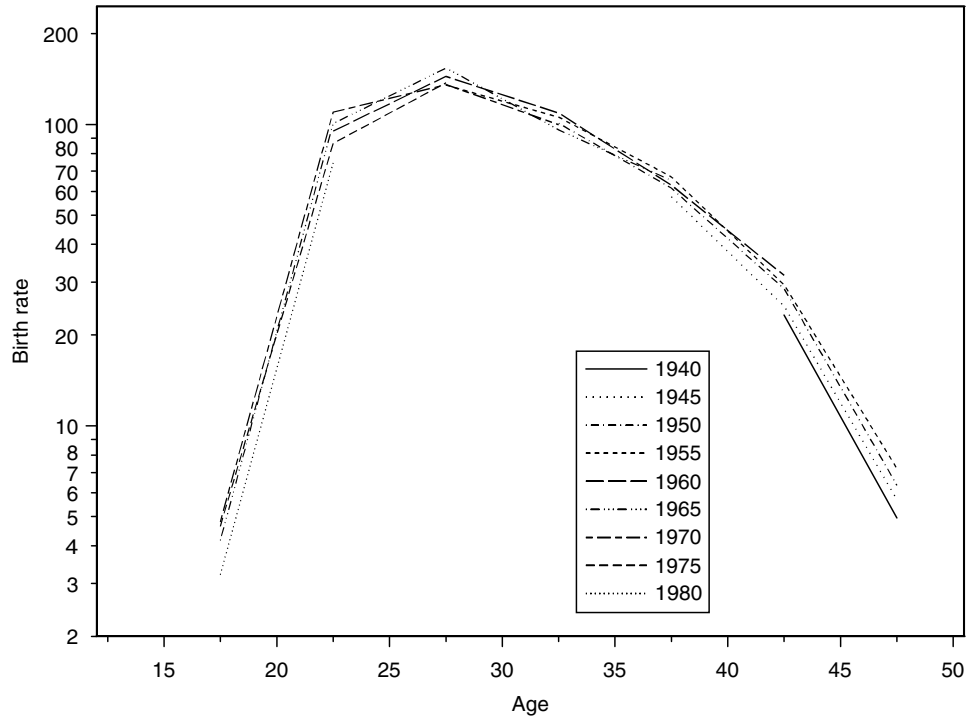


Figure 2 Cohort plot: a semilog plot of birth rates among US black women by age and cohort

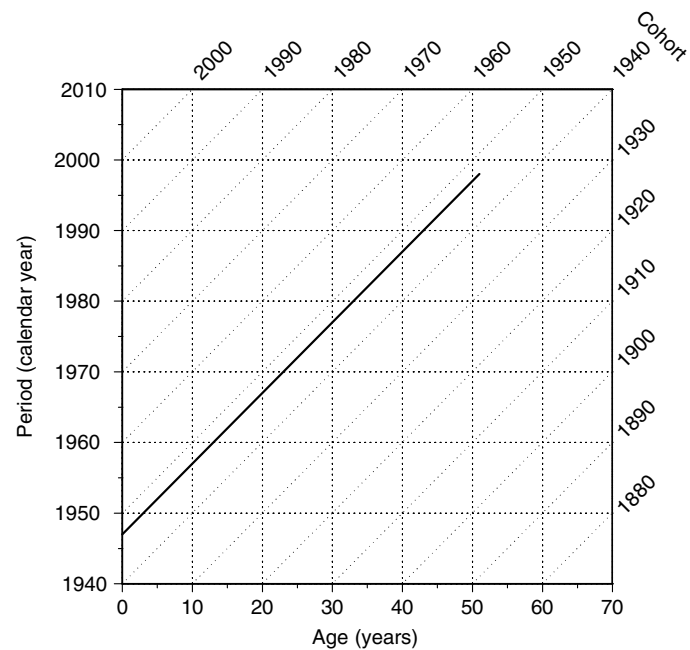


Figure 3 Lexis diagram showing the relationship between age, period, and cohort. The diagonal line traces age-period lifetime for an individual born in 1947

4 Age–Period–Cohort Analysis

because $k' = j' - i'$. Thus, each of the curvatures can be uniquely determined, but the overall slopes are hopelessly entangled so that only certain combinations can be uniquely estimated [4].

The implication of the identifiability problem is that the overall direction of the effect for any of the three temporal components cannot be determined from a regression analysis (*see Multiple Linear Regression*). Thus, we cannot even determine whether the trends are increasing or decreasing with cohort, for instance. Figure 4 displays several combinations of age, period and cohort parameters, each set of which provides an identical set of fitted rates. However, even though the specific trends cannot be uniquely estimated, certain combinations of the overall trend can be uniquely determined, such as $\beta_\pi + \beta_\gamma$ which is called the *net drift* [1, 2]. Alternative drift estimates covering shorter time spans can also be determined, and these have practical significance in that they describe the experience of following a particular age group in time, because both period and cohort will advance together. Curvatures, on the other hand, are completely determined including polynomial parameters for the square and higher powers, changes in slopes, and second differences. The significance test for any one of the temporal effects in

the presence of the other two will generally be a test of the corresponding curvature, and not the slope. Holford provides further detail on how software can be set up for fitting these models [5].

To illustrate the implications of the identifiability problem, and the type of valid interpretations that one can make by fitting an age-period-cohort model, we return to the data on birth rates among US black women. There are seven five-year age groups and five periods of identical width thus yielding $7 + 5 - 1 = 11$ cohorts. A general linear model (*see Generalized Linear Models (GLM)*) will be fitted to the log rates, introducing main effects for age, period and cohort. In situations in which the numerator and denominator for the rate are available, it is common to fit a **log-linear model** using Poisson regression (*see Generalized Linear Models (GLM)*), but the resulting interpretation issue will be identical for either model. An F test for the effect of cohort in the presence of age and period yields a value of 26.60 with $df_1 = 9$ and $df_2 = 15$. The numerator degrees of freedom, df_1 , are not 10 because the model with age and period effects implicitly includes their linear contribution, and thus the linear contribution for cohort. Therefore, this test can only evaluate the curvature for cohort. Similarly, the tests for age

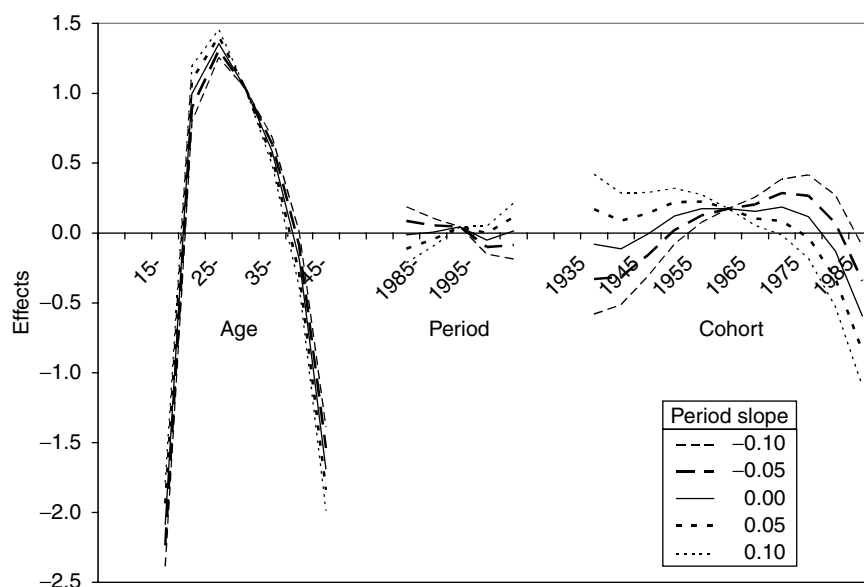


Figure 4 Age, period, and cohort effects for a log-linear model for birth rates in US black women, 1980–2001 by alternatively constrained period slopes

($F_{5,15} = 3397.47$, $p < 0.0001$) and period ($F_{3,15} = 4.50$, $p = 0.0192$) in the presence of the other two temporal factors are tests of curvature.

Figure 4 shows five sets of age, period and cohort parameters that may be obtained using least squares estimation. Each set of parameters provides an identical fit to the data, but there is obviously not a unique solution here but rather an infinite number of solutions. At the same time, once one of the slopes has been fixed (in this example the period slopes have been fixed), the other slopes can be identified. Notice that when the period slope is arbitrarily decreased, the underlying period trend is effectively rotated in a clockwise direction. Observing the corresponding cohort parameters, we can see that when the period trends are decreased, the trends for cohort are increased, that is, the corresponding estimates for the cohort parameters are rotated counterclockwise. Likewise, the age parameters experience a counterclockwise rotation, although in this example it is not easy to see because of the steep age trends. Sometimes there may be external information indicating a particular constraint for one of the temporal parameters, and once this can be applied then the other effects are also identified. However, such information must come from external sources because within the dataset itself it is impossible to disentangle the interrelationship so as to obtain a unique set of parameters.

In the absence of the detail required to make a constraint on one of the temporal parameters, it is safer to make inferences using estimable functions of the parameters, that is, functions that do not depend on an arbitrary constraint. Curvature, which would include both the overall departure from linear trend, as well as local changes of direction are estimable [1, 2, 4]. The latter would include polynomial terms of power greater than one (see **Polynomial Model**), change of slope, or second differences, which would compare the parameter at one point to the average of the parameters just before and just after that point. In addition, the sum of the period and cohort slope or drift is also estimable, thus providing a net indication of the trend.

In our example, we can see from the solid lines in Figure 4 that the first three periods show a gradual increasing trend, as do the first four cohorts. If we were to add these slopes, we would have an estimate of the early drift, which would be positive because both of the slope components are positive. Similarly,

the last three periods and the last three cohorts are negative, implying that the recent drift would also be negative. While the individual interpretation of the other lines shown in Figure 4 would be slightly different, the sum would be the same, thus indicating increasing early drift and decreasing late drift.

We can estimate the drift by taking the sum of the contrasts for linear trend in the first three periods and the first four cohorts, that is, $(-1, 0, 1, 0, 0)/2$ for period and $(-3, -1, 1, 3, 0, 0, 0, 0, 0)/10$ for cohort. This yields the result 0.1036 ($t_{15} = 6.33$, $p < 0.0001$), which indicates that the positive early drift is statistically significant. Similarly, the late drift, which uses the contrast $(0, 0, -1, 0, 1)/2$ for period and $(0, 0, 0, 0, 0, 0, -3, -1, 1, 3)/10$ for cohort, yields -0.1965 ($t_{15} = -12.02$, $p < 0.0001$), which is highly significant and negative.

In this discussion, we have concentrated on the analysis of data that have equal spaced intervals with age and period. The unequal case introduces further identifiability problems, which involve not only the overall linear trend, but certain short-term patterns, as well (see **Identification**). The latter can sometime appear to be cyclical trends; therefore, considerable care is needed in order to be certain that these are not just an artifact of the identifiability issues that arise for the unequal interval case.

References

- [1] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. I: age-period and age-cohort models, *Statistics in Medicine* 6, 449–467.
- [2] Clayton, D. & Schifflers, E. (1987). Models for temporal variation in cancer rates. II: age-period-cohort models, *Statistics in Medicine* 6, 469–481.
- [3] Fienberg, S.E. & Mason, W.M. (1978). Identification and estimation of age-period-cohort models in the analysis of discrete archival data, in *Sociological Methodology 1979*, Schuessler K.F., eds, Jossey-Bass, Inc., San Francisco, 1–67.
- [4] Holford, T.R. (1983). The estimation of age, period and cohort effects for vital rates, *Biometrics* 39, 311–324.
- [5] Holford, T.R. (2004). Temporal factors in public health surveillance: sorting out age, period and cohort effects, in *Monitoring the Health of Populations*, Brookmeyer R., Stroup D.F., eds, Oxford University Press, Oxford, 99–126.
- [6] Tango, T. & Kurashina, S. (1987). Age, period and cohort analysis of trends in mortality from major diseases in Japan, 1955 to 1979: peculiarity of the cohort born in the early Showa Era, *Statistics in Medicine* 6, 709–726.

THEODORE HOLFORD

Akaike's Criterion

CHARLES E. LANCE

Volume 1, pp. 38–39

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Akaike's Criterion

Akaike's information criterion (AIC) is now probably best known as an overall goodness-of-fit index (GFI) for confirmatory factor (*see* **Factor Analysis: Confirmatory**) and **structural equation models (SEMs)**. Originally developed for model selection in regression models [3], AIC has a remarkably solid theoretical foundation in information theory in the mathematical statistics literature [1–4]. The problem of model selection for which AIC was formulated is that of choosing a ‘best approximating’ model among a class of competing models, possibly with different numbers of estimated parameters, using an appropriate statistical criterion. Conceptually, AIC does this by balancing what Bozdogan [4] refers to as the risk of modeling, the extent to which the fitted model differs from the true model in the population, versus the risk of estimation, or discrepancy between population model parameters and sample-based estimates (*see* [8]).

AIC is often written as:

$$\chi^2 + 2q \quad (1)$$

where χ^2 is the maximum likelihood chi-squared statistic and q refers to the number of free parameters being estimated in the estimated model. According to Bozdogan [3], ‘the first term ... is a measure of inaccuracy, badness of fit, or bias when the **maximum likelihood estimators** of the models are used’ while the ‘... second term ... is a measure of complexity, of the penalty due to the increased unreliability, or compensation for the bias in the first term which depends upon the number of parameters used to fit the data’ (p. 356). Thus, when several models’ parameters are estimated using maximum likelihood, the models’ AICs can be compared to find a model with a minimum value of AIC. ‘This procedure is called the *minimum AIC procedure* and the model with the minimum AIC is called the *minimum AIC estimate* (MAICE) and is chosen to be the best model’ ([3], p. 356). Cudeck and Browne [6] and Browne and Cudeck [5] considered AIC and a closely related index proposed by Schwartz [11] as measures of cross-validation of SEMs; Cudeck and Brown proposed a ‘rescaled’ version of AIC, ostensibly ‘to eliminate the effect of sample size’ (p. 154).

Also, Bozdogan [3, 4] has extended Akaike's theory to develop a number of information complexity (ICOMP) measures of model fit based on various definitions of model complexity.

In one large-scale simulation, Marsh, Balla, and McDonald [9] found that AIC was very sensitive to sample size (an undesirable characteristic for a GFI) but suggested that AIC might be useful for comparing alternative, even nonnested models because of its absolute, as opposed to relative, nature. In a second large-scale simulation, Hu and Bentler [7] confirmed this finding and also found that AIC was the *least* sensitive to model misspecification error among the GFIs studied. In a thorough analysis of several SEM GFIs, McDonald and Marsh [10] explain how AIC tends to favor sparsely parameterized models in small samples and models that contain large numbers of free parameters (in the limit, saturated models) in large samples. Thus, despite its firm theoretical foundation and strong intuitive appeal, AIC has little empirical support to justify its application in practice. We agree with Hu and Bentler [7]: ‘We do not recommend [its] use’ (p. 446).

References

- [1] Akaike, H. (1981). Likelihood of a model and information criteria, *Journal of Econometrics* **16**, 3–14.
- [2] Akaike, H. (1987). Factor analysis and AIC, *Psychometrika* **52**, 317–332.
- [3] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika* **52**, 345–370.
- [4] Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity, *Journal of Mathematical Psychology* **44**, 62–91.
- [5] Browne, M.W. & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures, *Multivariate Behavioral Research* **24**, 445–455.
- [6] Cudeck, R. & Browne, M.W. (1983). Cross-validation of covariance structures, *Multivariate Behavioral Research* **18**, 147–167.
- [7] Hu, L. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterization model misspecification, *Psychological Methods* **4**, 424–453.
- [8] MacCallum, R.C. (2003). Working with imperfect models, *Multivariate Behavioral Research* **38**, 113–139.
- [9] Marsh, H.W., Balla, J.R. & McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis:

2 Akaike's Criterion

- the effect of sample size, *Psychological Bulletin* **103**, 391–410.
- [10] McDonald, R.P. & Marsh, H.W. (1990). Choosing a multivariate model: noncentrality and goodness of fit, *Psychological Bulletin* **107**, 247–255.
- [11] Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

CHARLES E. LANCE

Allelic Association

LON R. CARDON

Volume 1, pp. 40–43

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Allelic Association

Allelic association describes perhaps the simplest relationship between genes and phenotypes – differences in trait scores or disease frequencies with differences in alternative forms of a gene. For example, consider a genetic mutation that converts the DNA base pair Thymine to a Guanine, denoted $T \rightarrow G$. If having the T causes an increase in, say birth weight, then birth weight and the T allele are said to be associated. This concept of many small genetic effects dates back to the initial demonstration that discrete (Mendelian) changes in inherited material could lead to continuously varying phenotypes [8]. It is the cornerstone of the discipline of biometrical genetics [1, 7, 13].

For continuously measured phenotypes such as blood pressure or body weight, allelic association typically means that average levels of the trait differ according to the different alleles. For discrete traits such as diabetes or stroke, allelic association refers to different allele frequencies in groups of individuals who have the disease relative to those who do not. The general principle is the same in both cases: correlations between allele frequencies and outcomes. The statistical tools used to conduct the tests of significance are often different however [6].

Although formally the term ‘allelic association’ refers only to a specific allele, it generally encompasses **genotype** effects as well. For example, a diallelic locus with alleles T and G will produce genotypes TT, TG, and GG. Presence or absence of the T allele may be associated with higher trait values, but the genotypes themselves may offer a more precise association pattern, as is the case with different models of gene action (dominant, recessive, additive, multiplicative, etc.). Statistical tests of association based on individual alleles involve fewer parameters than genotype-based tests and thus are often preferred, but in practice, consideration of single-alleles versus genotypes is not just an esoteric statistical issue: the phenotype/allele/genotype relationship may have important consequences for gene identification and characterization. For example, mutations in the NOD2 locus for inflammatory bowel disease confer about a threefold increase in risk when considered as alleles, but as much as a 30- to 40-fold increase when considered as genotypes [10, 14].

As with any statistical correlation, the adage ‘correlation does not imply causation’ is also applicable in the domain of allelic association. Associated alleles may indeed cause changes in phenotypes, but they need not do so. Situations in which the associated alleles do cause trait changes are referred to as ‘direct’ associations, while situations in which different, but correlated, alleles cause the phenotype are referred to as ‘indirect’ associations [5]. The difference between direct and indirect association is depicted in Figure 1, in which the causal/susceptibility allele is depicted as a star and alleles at a correlated locus as circles. The frequency of the causal allele is higher in individuals with high trait values than with low trait values (70% versus 50%), as is the frequency of the noncausal correlated allele (50% versus 30%). Armed with just these frequencies, it would not be possible to determine which of the two loci is causal versus which is indirectly associated with the phenotype via a primary correlation with the disease allele.

In Figure 1, the correlation between the noncausal and causal loci is independent of the phenotype. They are correlated only because they are located close together on a chromosome and they have been transmitted together in the population over generations. We would not need to measure any traits to observe the correlation between the two genetic markers; it exists simply because of the history of the population studied. This correlation between alleles at different loci in a population is referred to as *linkage disequilibrium* (LD). Although the terms ‘linkage disequilibrium’ and ‘allelic association’ are sometimes used interchangeably in the literature, it is convenient and more precise to consider them as related but non-identical concepts. In general, the former refers to correlations between any two genetic loci, irrespective of disease, while the latter refers to a correlation between a genetic locus and a measured phenotype. This semantic distinction is important because LD and allelic association comprise different components of various disease-gene identification strategies. The widely used candidate-gene design involves a prior hypothesis about the role of some specific gene with a disease, and then evaluates genetic variants in that gene to test the correlation hypothesis [2]. These are thus the direct studies of allelic association. In contrast, the increasingly popular approaches involving hypothesis-free assessment of many loci scattered about large regions or the entire genome (positional

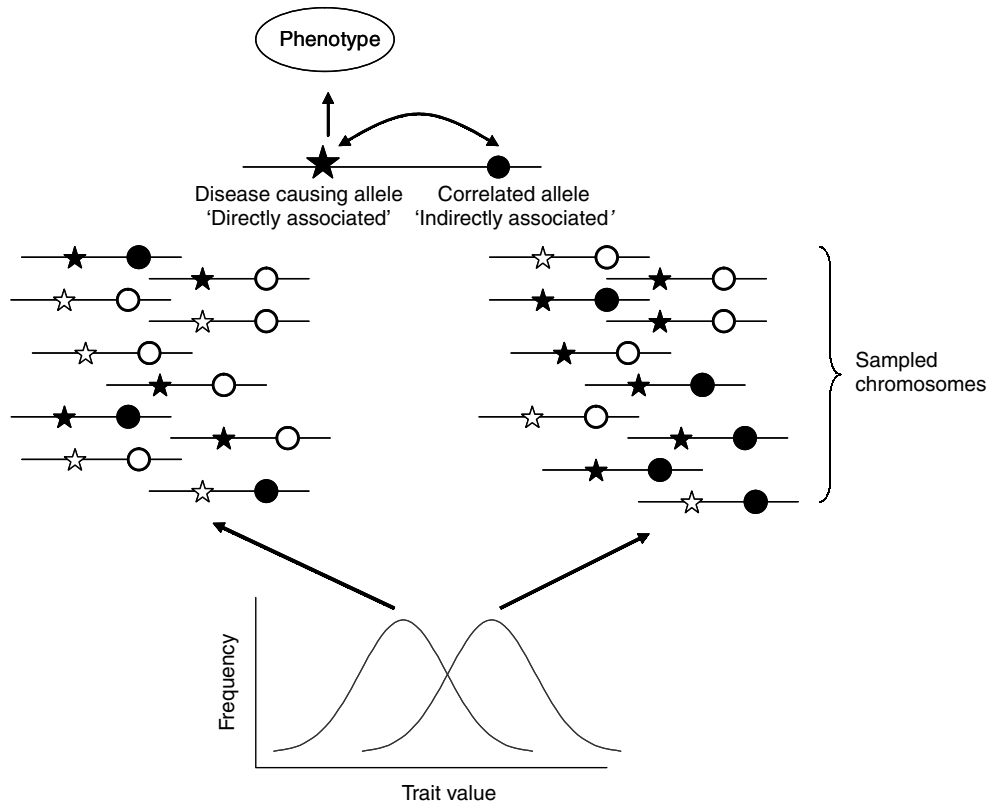


Figure 1 Example of indirect and direct association. The disease allele is shown as a filled star. Alleles at an anonymous marker are shown as circles. The two markers are correlated (in 'linkage disequilibrium') with one another. The frequency of the disease allele is greater in the individuals with high trait scores ($7/10 = 0.70$) versus those with low trait scores (0.50). Similarly, the correlated allele has higher frequency in high-trait individuals (0.50) versus low-trait individuals (0.30). Although the markers are correlated, they are not perfectly correlated, so it is difficult to distinguish the directly associated disease allele from the indirectly associated marker allele

cloning, candidate region, or whole-genome association designs) are indirect association strategies that rely on linkage disequilibrium between measured genetic markers and unmeasured causal loci [15].

From a sampling perspective, most association studies can be classified into two general groupings: case/control and family based. Historically, **case-control studies** have been used most widely, involving collections of sample of individuals who have a disease or trait of interest plus a sample of control individuals who do not have the trait (or who are randomly ascertained in some designs). Tests of allelic association involve comparisons of allele or genotype frequencies between the two groups. Matching between the case and control samples is a critical feature of this design, since differences between the

groups can be incorrectly ascribed to allelic association when in fact they reflect some unmeasured variable(s). Such spurious association outcomes are described as resulting from 'population stratification', or classical confounding in epidemiological terms.

Spurious association due to population stratification has worried geneticists considerably over the past two decades because human populations are known to have widely varying allele frequencies simply because of their different population histories [4]. Thus, one might expect many allele frequency differences between groups by chance alone. To address this concern, a number of family-based designs have been developed, popularized most widely in the Transmission Disequilibrium Test [17]. In this design, samples of affected offspring and their

two parents are collected (the disease status of the parents is usually irrelevant), and the frequencies of the alleles that are transmitted from parent to offspring form the 'case' group, while those that are present in the parents but not transmitted to the offspring form the 'control' group. The general idea is that drawing cases and controls from the same families renders the confounding allele frequency differences irrelevant. Similar strategies have been developed for continuous traits [9, 11, 12].

With the advent of family-based studies and with rapidly advancing studies of linkage disequilibrium across the human genome, the related concepts of linkage and association are becoming increasingly confused. The main distinction is that genetic linkage exists within families, while association extends to populations. More specifically, linkage refers to cosegregation of marker alleles with trait alleles within a family. Using the T $\rightarrow$ G example, a family would show evidence for linkage if members having high trait scores shared the T allele more often than expected by chance. However, another family would also show linkage if its members shared the G allele more often than expected by chance. In each case, an allele occurs in excess of expectations under random segregation, so each family is linked. In contrast, allelic association requires allelic overrepresentation across families. Thus, only if both families shared the same T (or G) allele would they offer joint evidence for association. In a simple sense, genetic linkage is allelic association within each family. The linkage-association distinction is important because linkage is useful for identifying large chromosome regions that harbor trait loci, but poor at identifying specific genetic variants, while association is more powerful in very tight regions but weak in identifying broad chromosomal locations. In addition, association analysis is more powerful than linkage for detecting alleles that are common in the population [16], whereas family-based linkage approaches offer the best available genetic approach to detect effects of rare alleles.

Association studies have not yielded many successes in detecting novel genes in the past, despite tens of thousands of attempts [18]. There are many postulated reasons for this lack of success, of which some of the most prominent are small sample sizes and poorly matched cases and controls [3]. In addition, there have been few large-effect complex trait

loci identified by any strategy, lending at least partial support to the traditional poly-/oligo-genic model, which posits that common trait variation results from many genes having individually small effects. To the extent that this model applies for a specific trait, so that each associated variant has an individually small effect on the trait, the sample size issue becomes even more problematic. Future studies are being designed to detect smaller effect sizes, which should help reveal the ultimate utility of the association approach for common traits.

References

- [1] Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*, Clarendon Press, Oxford.
- [2] Cardon, L.R. & Bell, J.I. (2001). Association study designs for complex diseases, *Nature Review. Genetics* **2**, 91–99.
- [3] Cardon, L.R. & Palmer, L.J. (2003). Population stratification and spurious allelic association, *Lancet* **361**, 598–604.
- [4] Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. (1994). *History and Geography of Human Genes*, Princeton University Press, Princeton.
- [5] Collins, F.S., Guyer, M.S. & Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation, *Science* **278**, 1580–1581.
- [6] Elston, R.C., Palmer, L.J., Olson, J.E., eds (2002). *Biostatistical Genetics and Genetic Epidemiology*, John Wiley & Sons, Chichester.
- [7] Falconer, D.S. & Mackay, T.F.C. (1996). *Quantitative Genetics*, Longman, Harlow.
- [8] Fisher, R.A. (1918). The correlations between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
- [9] Fulker, D.W., Cherny, S.S., Sham, P.C. & Hewitt, J.K. (1999). Combined linkage and association sib-pair analysis for quantitative traits, *American Journal of Human Genetics* **64**, 259–267.
- [10] Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.F., Sahbatou, M. & Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to crohn's disease, *Nature* **411**, 599–603.
- [11] Lange, C., DeMeo, D.L. & Laird, N.M. (2002). Power and design considerations for a general class of family-based association tests: quantitative traits, *American Journal of Human Genetics* **71**, 1330–1341.
- [12] Martin, E.R., Monks, S.A., Warren, L.L. & Kaplan, N.L. (2000). A test for linkage and association in general

4 Allelic Association

- pedigrees: the pedigree disequilibrium test, *American Journal of Human Genetics* **67**, 146–154.
- [13] Mather, K. & Jinks, J.L. (1982). *Biometrical Genetics*, Chapman & Hall, London.
- [14] Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., Achkar, J.P., Brant, S.R., Bayless, T.M., Kirschner, B.S., Hanauer, S.B., Nunez, G. & Cho, J.H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease, *Nature* **411**, 603–606.
- [15] Risch, N.J. (2000). Searching for genetic determinants in the new millennium, *Nature* **405**, 847–856.
- [16] Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science* **273**, 1516–1517.
- [17] Spielman, R., McGinnis, R. & Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics* **52**, 506–516.
- [18] Weiss, K.M. & Terwilliger, J.D. (2000). How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26**, 151–157.

LON R. CARDON

All-X Models

RONALD S. LANDIS

Volume 1, pp. 43–44

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

All-X Models

All-X models are measurement models tested using **structural equation modeling** techniques in which all **latent variables** are treated as exogenous. The term ‘All-X’ is used because the letter X is conventionally used to represent measures of exogenous latent factors. As an example, consider a model in which two endogenous variables (η_1 and η_2), each measured with three items, are predicted from two exogenous variables (ξ_1 and ξ_2), each also measured with three items. Figure 1 provides a graphical representation of this model.

In this figure, x1 through x6 and y1 through y6 are scale items, λ_{11} through λ_{62} represent the loadings of each item on its respective factor, δ_1 through δ_6 and ε_1 through ε_6 are the measurement errors associated with each item, Φ_{12} is the correlation/covariance between the underlying exogenous latent variables, ψ_{12} is the correlation/covariance between the residuals of endogenous latent variables, β_{12} is the correlation/covariance between the underlying endogenous latent variables, and γ_{11} and γ_{22} are the relationships between the exogenous and endogenous latent variables.

This full model includes both structural and measurement relationships. Eight unique matrices are

Table 1 Eight unique matrices required to assess all elements of the model shown in Figure 1

<i>Lambda y</i>	= factor loadings of measured variables on the latent dependent variable
<i>Lambda x</i>	= factor loadings of measured variables on the latent independent variables
<i>Beta</i>	= interrelationships among ‘endogenous’ latent variables
<i>Gamma</i>	= relationships between ‘exogenous’ and ‘endogenous’ variables
<i>Phi</i>	= relationships among the exogenous variables
<i>Psi</i>	= relationships between residuals of the endogenous variables
<i>Theta delta</i>	= uniquenesses associated with indicators of exogenous variables
<i>Theta epsilon</i>	= uniquenesses associated with indicators of endogenous variables

required to fully assess all elements of this model. These matrices are shown in Table 1.

The left-hand panel of Figure 1 illustrates the relationships tested in an all-X measurement model.

Specifically, the first three items (x1–x3) are indicators of the first dimension and the second three items (x4–x6) are indicators of the second dimension. Model specification requires estimation of three sets of parameters including (a) loadings of measured variables on the underlying exogenous construct(s),

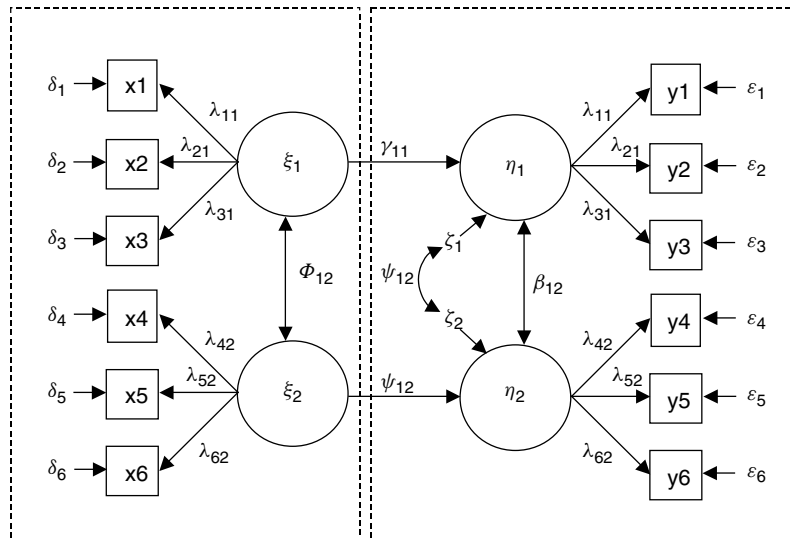


Figure 1 Model in which two endogenous variables (η_1 and η_2), each measured with three items, are predicted from two exogenous variables (ξ_1 and ξ_2), each also measured with three items

2 All-X Models

(b) measurement errors associated with the observed variables, and (c) relationships, if any, between the exogenous constructs.

Testing the all-X model requires the use of only three of the previously described eight matrices. The factor loadings would be contained in the lambda-x (λ_x) matrix, the measurement errors would be contained in the theta-delta (Θ_δ) matrix, and the

relationship between the factors would be captured in the phi (Φ) matrix.

(See also **Structural Equation Modeling: Overview**)

RONALD S. LANDIS

All-Y Models

RONALD S. LANDIS

Volume 1, pp. 44–44

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

All-Y Models

All-Y models are measurement models tested using confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) in which all variables are treated as endogenous. For example, consider the situation described and illustrated under the ‘**all-X models**’ entry. The all-Y measurement model is captured in the right-hand panel of this figure. The first three items (y1 through y3) are indicators of the first latent endogenous variable (η_1), and the second three items (y4 through y6) are indicators of the second latent endogenous variable (η_2). Model specification requires estimation of three sets of parameters including (1) loadings of measured variables on the underlying construct(s), (2) relationships among measurement errors associated with the observed variables, and (3) relationships between the residual errors for the latent endogenous constructs.

Testing the all-Y model requires the use of only three of the eight matrices described in the ‘all-X’ entry. The factor loadings would be contained in the lambda-y (λ_y) matrix, the measurement error variances would be contained in the theta-epsilon (Θ_ϵ) matrix, and the relationships among the residual errors for the latent endogenous constructs would be captured in the psi (ψ) matrix.

All-Y models are similar to all-X models, with the important difference being the treatment of **latent variables** as endogenous, as opposed to exogenous. This difference requires the estimation of relationships between the residuals of the latent endogenous variables rather than the estimations of relationships among exogenous latents in all-X models.

(*See also* **Structural Equation Modeling: Overview**; **Structural Equation Modeling: Software**)

RONALD S. LANDIS

Alternating Treatment Designs

GINA COFFEE HERRERA AND THOMAS R. KRATOCHWILL

Volume 1, pp. 44–46

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Alternating Treatment Designs

Overview

The alternating treatments design (ATD) is a type of single-participant design (*see* **Single-Case Designs**) characterized by rapid and random/semirandom shifts between two or more conditions [1]. Essentially, conditions are alternated as often as necessary to capture meaningful measurement of the behavior of interest. For example, if daily measurement is the most telling way to measure the behavior of interest, conditions are alternated daily. Similarly, depending on the behavior of interest, conditions could be alternated bi-daily, weekly, by session, or by any other schedule that is appropriate for the behavior. In addition to the frequency with which conditions are changed, the order by which conditions are changed is a significant component of the ATD. Usually, conditions in an ATD are shifted semirandomly. The alternations are semirandom, rather than random, because there are restrictions on the number of times conditions can be sequentially implemented. Overall, the ATD is most commonly used to compare two or more treatments through the examination of treatment divergence and overlap [3]. Other uses include comparing a treatment to no treatment or inspecting treatment components.

To illustrate the use of the ATD in a classroom setting, consider a recent study by Skinner, Hurst, Teeple, and Meadows [5]. In this study, the researchers examined the effects of different mathematics assignments (control and experimental) on the on-task behavior and the rate of problem completion in students with emotional disturbance. The experimental assignment was similar to the control assignment with the addition of brief mathematics problems interspersed after every third problem. In a self-contained classroom, the researchers observed four students across 14 days (one 15-minute session per day) as the students completed the mathematics assignments. The assignment (control or experimental) was randomly selected on days 1, 5, 9, and 13. Then, the assignments were alternated daily so that if the students completed the experimental assignment on day 1, they would complete the control assignment

on day 2, and so on. Results suggested that the interspersal procedure produced an increase in problem completion and in on-task behavior.

Considerations

This research example illustrates certain considerations of using an ATD [3]. In particular, when using an ATD, questions regarding the number of conditions, baseline data, alternations, and analyses emerge. First, with regard to the number of conditions, it is important to understand that as the number of conditions increases, the complexity of the ATD increases in terms of drawing comparisons among conditions of the design. Consequently, it is generally recommended that the number of conditions not exceed three, and that each condition has at least two data points (although four or more are preferable). In the example, the researchers used two conditions, each with seven data points.

In terms of baseline data, a unique feature of the ATD, in comparison to other single-participant designs, is that baseline data are not required when one wants to examine the relative effectiveness of treatments that are already known to be effective (baseline data are usually required when a treatment's effectiveness has not been demonstrated). For example, because in the mathematics assignments Skinner et al. [5] used treatments that were already known to be effective, baseline data were not required. However, regardless of the known effectiveness, including baseline data before the ATD and as a condition within the ATD can provide useful information about individual change and the effectiveness of the treatments, while ruling out extraneous variables as the cause of change.

In addition to the number of conditions and baseline data, alternations are also aspects of the ATD that must be considered. First, as the number of alternations increase, the number of opportunities for divergence and overlap also increase. Because comparisons are made after examining divergence and overlap of conditions, increasing the number of alternations can yield more accurate results in the analyses. In general, although the number of alternations cannot be less than two, the maximum number of alternations varies depending on the study because the unit of measurement (e.g., day, week, session) and the duration of the treatment effect

2 Alternating Treatment Designs

influence the number of alternations. The variables in the example allowed for seven alternations.

Related to the number of alternations, but with more serious ramifications, is the way that conditions are alternated (e.g., order and schedule). The manner in which the conditions are alternated is important because it can threaten the validity of the ATD. In particular, the validity of the ATD can be threatened by sequential confounding, **carryover effects**, and alternation effects – a type of carryover effect [1]. Fortunately, although these concerns have the potential of threatening the validity of the ATD, they can usually be addressed by random or semi-random alternations and by monitoring for carryover effects. In addition to implementing a randomization strategy, the researchers in the example reported that naturally occurring absences also contributed to controlling for carryover effects.

A final consideration of using an ATD is the data analysis. Analysis of data in an ATD is important so that one can understand the effects of the treatment(s). As with other single-participant designs, data points in an ATD can be analyzed by visual inspection – that is, by assessing the level, trend, and variability within each condition [3]. This is the methodology used in the example [5] and is the most common way of analyzing data in single-participant designs. In addition to visual inspection, however, data from ATDs can also be analyzed with inferential statistics such as **randomization tests** [4]. Using this type of inferential analysis is unique to the ATD and can be accomplished by using available software packages [4] or doing hand calculations. It should be noted, however, that randomization tests are appropriate when alternations are truly random. Additional nonparametric tests that can be used to analyze data from an ATD include Wilcoxon's matched-pairs, signed-ranks tests (see **Paired Observations, Distribution Free Methods**), **sign tests**, and Friedman's analysis of variance (see **Friedman's Test**) [2].

Overall, depending on one's needs, the ATD is a useful design and presents certain advantages over other single-participant designs [3]. One advantage is that the ATD does not require the withdrawal of a treatment. This aspect of the ATD can be useful in avoiding or minimizing the ethical and practical issues that withdrawal of treatment can present. Another advantage of the ATD is that comparisons between treatments can be made more quickly than in other single-participant designs – sometimes in as little time as one session. A final advantage of the ATD, as discussed previously, is that the ATD does not require baseline data.

References

- [1] Barlow, D.H. & Hayes, S.C. (1979). Alternating treatments design: one strategy for comparing the effects of two treatments in a single subject, *Journal of Applied Behavior Analysis* **12**, 199–210.
- [2] Edgington, E.S. (1992). Nonparametric tests for single-case experiments, in *Single-case Research Design and Analysis: New Directions for Psychology and Education*, T.R. Kratochwill & J.R. Levin, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 133–157.
- [3] Hayes, S.C., Barlow, D.H. & Nelson-Gray, R.O. (1999). *The Scientist Practitioner: Research and Accountability in the Age of Managed Care*, 2nd Edition, Allyn & Bacon, Boston.
- [4] Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs, *Behaviour Research and Therapy* **32**, 783–786.
- [5] Skinner, C.H., Hurst, K.L., Teeple, D.F. & Meadows, S.O. (2002). Increasing on-task behavior during mathematics independent seat-work in students with emotional disturbance by interspersing additional brief problems, *Psychology in the Schools* **39**, 647–659.

GINA COFFEE HERRERA AND THOMAS
R. KRATOCHWILL

Analysis of Covariance

BRADLEY E. HUITEMA

Volume 1, pp. 46–49

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Covariance

The control of nuisance variation is important in the design, analysis, and interpretation of experiments, quasi-experiments, and observational studies. The analysis of covariance (ANCOVA) is a method for controlling nuisance variation statistically. It can be used in place of, or in addition to, other approaches such as blocking (*see* **Block Random Assignment**) and **matching**. A variable that is of little experimental interest (such as verbal aptitude), but is believed to affect the response on the dependent variable (e.g., sales performance), is a **nuisance variable**. A measure of such a variable is called a covariate in the context of ANCOVA; it should be measured before treatments are applied.

There are two potential advantages of applying ANCOVA rather than a conventional **analysis of variance (ANOVA)** to data from a traditional randomized-groups design. First, the dependent variable means are adjusted to remove the variation that is predictable from chance differences on the covariate. Second, the ANCOVA error mean square is usually smaller than the corresponding ANOVA error mean square; this leads to narrower **confidence intervals**, increased **power**, and larger **effect size estimates**. The extent to which the means are adjusted and the error term is decreased depends upon several issues, the most important of which are the research design and the degree of relationship between the covariate and the dependent variable.

Design Issues

Because the randomized-groups design (*see* **Analysis of Variance: Classification**) yields groups that are probabilistically equivalent on all variables before treatments are applied, differences between covariate means are generally small (unless very small samples are used) and little adjustment of means is likely. But, if a useful covariate is employed, the size of the error mean square associated with the ANCOVA F test will be much smaller than the corresponding term in ANOVA. Hence, the major advantage of ANCOVA in a randomized-groups design is the reduction in the size of the error term.

When ANCOVA is applied to the biased assignment **quasi-experimental design** (often called the

regression-discontinuity design [4, 8]) where treatment groups are formed exclusively on the basis of the covariate score, the adjustment of the means will be large. This is a situation in which neither the unadjusted means nor the corresponding ANOVA F test supplies meaningful information about treatment effects. In contrast, the adjusted means and the corresponding ANCOVA F test supply relevant information about treatment effects.

Observational studies are also sometimes analyzed using ANCOVA. Although the covariate(s) often provides a major reduction in bias associated with group differences on confounding variables, it does not usually remove all such bias. Bias is likely to remain for at least two reasons. The first is measurement error in the covariate(s) and the second is omission of important (but often unknown) covariates in the model. Hence, ANCOVA results based on observational studies should be considered to have lower **Internal Validity** than results based on randomized experiments and quasi-experiments.

Nature of the Covariate

A covariate (X) is considered to be useful if it has a reasonably high correlation with the dependent variable (Y). It is typical for X and Y to be measures of different constructs (such as age and job performance), but in the application of ANCOVA to the randomized-groups pretest–posttest design, the pretest is used as the covariate and the posttest is used as the dependent variable. Because this pre–post design uses the same metric for both pre- and posttesting, several analytic approaches have been recommended. This design is ideal for illustrating the advantages of ANCOVA relative to other popular analysis methods.

Example

Consider the data from a randomized-groups pretest–posttest design presented in Table 1.

The purpose of the study is to evaluate three methods of training; the pretest is a measure of achievement obtained before training and the posttest is the measure of achievement obtained after training. Four methods of analyzing these data are described here. The first one, a one-factor ANOVA on posttest scores, is the least satisfactory because it ignores potentially

2 Analysis of Covariance

Table 1 Four analyses applied to a pretest–posttest randomized-groups experiment (taken from [4])

Treatment	Pretest (X)	Posttest (Y)	Difference ($Y - X$)
1	2	4	2
1	4	9	5
1	3	5	2
2	3	7	4
2	4	8	4
2	4	8	4
3	3	9	6
3	3	8	5
3	2	6	4

Group	$\bar{X}$	$\bar{Y}$	$\bar{Y} - \bar{X}$	Adjusted $\bar{Y}$
1	3.00	6.00	3.00	6.24
2	3.67	7.67	4.00	6.44
3	2.67	7.67	5.00	8.64

ANOVA on posttest scores

Source	SS	df	MS	F	p
Between	5.56	2	2.78	0.86	.47
Within	19.33	6	3.22		
Total	24.89	8			

ANOVA on difference scores

Source	SS	df	MS	F	p
Between	6.00	2	3.00	2.25	.19
Within	8.00	6	1.33		
Total	14.00	8			

Split-plot ANOVA

Source	SS	df	MS	F	p
Between subjects	22.78	8			
Groups	4.11	2	2.06	0.66	.55
Error a	18.67	6	3.11		
Within subjects	79.00	9			
Times	72.00	1	72.00	108.00	.00
Times $\times$ gps.	3.00	2	1.50	2.25	.19
Error b	4.00	6	0.67		
Total	101.78	17			

ANCOVA on posttest scores (using pretest as covariate)

Source	SS	df	MS	F	p
Adjusted treatment	8.96	2	4.48	7.00	.04
Residual within gps.	3.20	5	0.64		
Residual total		12.16	7		

useful information contained in the pretest. It is included here to demonstrate the relative advantages of the other methods, all of which use the pretest information in some way.

The second approach is a one-factor ANOVA on the differences between the pretest and the posttest scores, often referred to as an analysis of change scores. The third approach is to treat the data as a two-factor split-plot factorial design (*see Analysis of Variance: Classification*) in which the three groups constitute the levels of the between-subjects factor and the two testing times (i.e., pre and post) constitute the levels of the repeated measurement factor. The last approach is a one-factor ANCOVA in which the pretest is used as the covariate and the posttest is used as the dependent variable.

The results of all four analytic approaches are summarized in Table 1. The group means based on variables X , Y , and $Y - X$ differences are shown, along with the adjusted Y means. Below the means are the summary tables for the four inferential analyses used to test for treatment effects. Before inspecting the P values associated with these analyses, notice the means on X . Even though random assignment was employed in forming the three small groups, it can be seen that there are annoying differences among these means. These group differences on the covariate may seem to cloud the interpretation of the differences among the means on Y . It is natural to wonder if the observed differences on the outcome are simply reflections of chance differences that were present among these groups at pretesting. This issue seems especially salient when the rank order of the Y means is the same as the rank order of the X means. Consequently, we are likely to lament the fact that random assignment has produced groups with different covariate means and to ponder the following question: ‘If the pretest (covariate) means had been exactly equivalent for all three groups, what are the predicted values of the posttest means?’ ANCOVA provides an answer to this question in the form of adjusted means. The direction of the adjustment follows the logic that a group starting with an advantage (i.e., a high $\bar{X}$) should have a downward adjustment to $\bar{Y}$, whereas a group starting with a disadvantage (i.e., a low $\bar{X}$) should have an upward adjustment to $\bar{Y}$.

Now consider the inferential results presented below the means in the table; it can be seen that the conclusions of the different analyses vary greatly.

The P values for the ANOVA on the posttest scores and the ANOVA on the difference scores are .47 and .19 respectively. The split-plot analysis provides three tests: a main-effects test for each factor and a test on the interaction. Only the interaction test is directly relevant to the question of whether there are differential effects of the treatments. The other tests can be ignored. The interaction test turns out to be just another, more cumbersome, way of evaluating whether we have sufficient evidence to claim that the difference-score means are the same for the three treatment groups. Hence, the null hypothesis of no interaction in the split-plot ANOVA is equivalent to the null hypothesis of equality of means in the one-factor ANOVA on difference scores.

Although these approaches are generally far more satisfactory than is a one-factor ANOVA on the posttest, the most satisfactory method is usually a one-factor ANCOVA on the posttest using the pretest as the covariate. Notice that the P value for ANCOVA ($p = .04$) is much smaller than those found using the other methods; it is the only one that leads to the conclusion that there is sufficient information to claim a treatment effect.

The main reason ANOVA on difference scores is usually less satisfactory than ANCOVA is that the latter typically has a smaller error mean square. This occurs because ANOVA on difference scores implicitly assumes that the value of the population within group regression slope is 1.0 (whether it actually is or not), whereas in the case of ANCOVA, the within group slope is estimated from the data. This difference is important because the error variation in both analyses refers to deviations from the within group slope. If the actual slope is far from 1.0, the ANCOVA error mean square will be much smaller than the error mean square associated with the ANOVA on $Y - X$ differences. The example data illustrate this point. The estimated within group slope is 2.2 and the associated ANCOVA error mean square is less than one-half the size of the ANOVA error mean square.

In summary, this example shows that information on the pretest can be used either as a covariate or to form pretest–posttest differences, but it is more effective to use it as a covariate. Although there are conditions in which this will not be true, ANCOVA is usually the preferred analysis of the randomized-groups pretest–posttest design. By the way, no additional

advantage is obtained by combining both approaches in a single analysis. If we carry out ANCOVA using the pretest as the covariate and the difference scores rather than the posttest scores as the dependent variable, the error mean square and the P value from this analysis will be identical to those shown in Table 1.

Assumptions

Several assumptions in addition to those normally associated with ANOVA (viz., homogeneity of population variances, normality of population error distributions, and independence of errors) are associated with the ANCOVA model. Among the most important are the assumptions that the relationship between the covariate and the dependent variable is linear, that the covariate is measured without error, that the within group regression slopes are homogeneous, and that the treatments do not affect the covariate.

Alternatives and Extensions

A simple alternative to a one-factor ANCOVA is to use a two-factor (treatments by blocks) ANOVA in which block levels are formed using scores on X (see **Randomized Block Designs**). Although this approach has the advantage of not requiring a linear relationship between X and Y , it also has several disadvantages including the reduction of error degrees of freedom and the censoring of information on X . Comparisons of the two approaches usually reveal higher power for ANCOVA, especially if the treatment groups can be formed using restricted randomization rather than simple random assignment.

Alternatives to conventional ANCOVA are now available to accommodate violations of any of the assumptions listed above [4]. Some of these alternatives require minor modifications of conventional ANCOVA computational procedures. Others such as those designed for dichotomously scaled dependent variables, robust analysis [3], complex matching [7], random treatment effects [6], and intragroup dependency of errors [6] require specialized software. Straightforward extensions of covariance analysis are available for experimental designs having more than one factor (multiple-factor ANCOVA), more than one dependent variable (multivariate ANCOVA), and more than one covariate (multiple ANCOVA). Most of these extensions are described in standard references on experimental design [1, 2, 5].

References

- [1] Keppel, G. & Wickens, T.D. (2004). *Design and Analysis: A Researcher's Handbook*, 4th Edition, Pearson Prentice-Hall, Upper Saddle River.
- [2] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [3] Hettmansperger, T.P. & McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*, John Wiley & Sons, New York.
- [4] Huitema, B.E. (1980). *Analysis of Covariance and Alternatives*, John Wiley & Sons, New York.
- [5] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Erlbaum, Mahwah.
- [6] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models*, 2nd Edition, Sage, Thousand Oaks.
- [7] Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics* **4**, 87–94.
- [8] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, New York.

(See also **Generalized Linear Models (GLM)**; **Linear Multilevel Models**; **Logistic Regression**; **Multiple Linear Regression**)

BRADLEY E. HITEMA

Analysis of Covariance: Nonparametric

VANCE W. BERGER

Volume 1, pp. 50–52

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Covariance: Nonparametric

Many studies are designed to clarify the relationship between two variables, a predictor (often an exposure in epidemiological study or a treatment in a medical trial) and an outcome. Yet, in almost all of these studies there is a third type of variable measured, a covariate. A covariate may be a confounder if it is both predictive of the outcome and associated with the predictor. For example, a simple comparison of survival times (an outcome) between cancer patients and subjects receiving a flu vaccine (so the predictor is binary, cancer vs. at risk for influenza) would likely demonstrate better survival for the subjects than for the patients. An uncritical interpretation of this finding might then be that the flu vaccine is superior to cancer treatment at extending survival, and so even cancer patients should have the flu vaccine (in lieu of their usual cancer treatment).

The problem with this interpretation, aside from its obvious conflict with intuition, is that it ignores the confounder, disease severity. The patients tend to have more severe disease than the subjects, so disease severity is associated with the predictor. In addition, those with less severe disease will tend to live longer, regardless of how they are treated, than those with a more severe disease, and so disease severity is also associated with the outcome. It might be of greater interest, then, to adjust or control for disease severity when making the survival comparison. Intuitively, one would wish to make the survival comparisons across treatment groups within levels of disease severity, so that apples are compared with apples and oranges are compared with oranges. More formally, a ‘covariate cannot be responsible for confounding within a stratum that is internally homogeneous with respect to the covariate’ [10]. Koch et al. [11] make explicit five benefits in adjusting an analysis for one or several covariates, including:

1. bias reduction through adjustment for baseline imbalances in covariate distributions (especially in observational studies);
2. better power through variance reduction;
3. creating comparable comparison groups;
4. clarifying the predictive ability of one or several covariates;

5. clarifying the uniformity of treatment effects over subpopulations defined by particular values or ranges of covariates.

In a subsequent article [12], some of the same authors (and others) stated that the first benefit does not apply to randomized trials (*see Randomization*) because ‘randomization provides statistically equivalent groups at baseline (that is, any departures from equivalence are random)’. Yet, systematic baseline imbalances (selection bias) can, in fact, occur even in properly randomized trials [5, 6], and so even the first benefit applies to randomized trials.

The fewer assumptions required, the better. Unfortunately, the validity of the popular **analysis of covariance** (ANCOVA) model is predicated on a variety of assumptions including normality of residuals, equal variances, linearity, and independence. When these assumptions are not met, the ANCOVA may not be robust [13]. By not requiring such assumptions, a nonparametric analysis offers better robustness properties. In Table 11 (pp. 590, 591) of [11] are listed numerous covariance techniques for categorical data, along with the assumptions required by each. Generally, even methods that are considered nonparametric, such as those discussed in [1] and [19], rely on the **central limit theorem**, chi-squared distributional assumptions for quadratic forms, and/or link functions connecting covariates to outcomes. The ideal situation is when no assumptions are required and inference can proceed on the basis of randomization (which is known, and hence is not an assumption).

In a randomized trial, exact design-based analyses are permutation tests (*see Permutation Based Inference; Linear Models: Permutation Methods*) [2, 7], which tend to be conducted without adjusting for covariates. In fact, there are several ways to build covariate adjustment into a permutation test. We note that while a continuous variable may be seen as an extension of a categorical variable (each outcome constitutes its own category), there is a qualitative difference in the way adjustment is conducted for continuous covariates. Specifically, adjustment for an ordinal covariate tends to be conducted by comparing treatment groups only within values of the covariate; there is no attempt to make comparisons across covariate values. In contrast, models tend to be fit when adjusting for continuous covariates, thereby allowing such disparate comparisons to be made.

2 Analysis of Covariance: Nonparametric

We will consider only categorical covariates, but distinguish nominal (including binary) covariates from ordinal covariates.

Koch et al. [11] present a data set with 59 patients, two treatment groups (active and placebo), five response status levels (excellent, good, moderate, fair, poor), and age as a continuous covariate. In their Table 5 (p. 577), the response variable is dichotomized into good and excellent versus moderate, fair, or poor. We then dichotomize the age into 54 or less (younger) versus 55 or over (older). Dichotomizing an ordinal response variable can result in a loss of **power** [3, 14, 16], and dichotomizing an ordinal covariate can result in a reversal of the direction of the effect [9], but we do so for the sake of simplicity. The data structure is then a $2 \times 2 \times 2$ table:

Now if the randomization was unrestricted other than the restriction that 32 patients were to receive placebo and 27 were to receive the active treatment, there would be $59!/([32!27!])$ ways to select 32 patients out of 59 to constitute the placebo group. An unadjusted permutation test would compare the test statistic of the observed table to the reference distribution consisting of the test statistics computed under the null hypothesis [2, 7] of all permuted tables (possible realizations of the randomization). To adjust for age, one can use an adjusted test statistic that combines age-specific measures of the treatment effect. This could be done with an essentially unrestricted (other than the requirement that the row margins be fixed) permutation test.

It is also possible to adjust for age by using an unadjusted test statistic (a simple difference across treatment groups in response rates) with a restricted permutation sample space (only those tables that retain the age*treatment distribution). One could also use both the adjusted test statistic and the adjusted permutation sample space. We find that for that data set in Table 1, the unadjusted difference in proportions (active minus placebo) is 0.4051, whereas the

weighted average of age-specific differences in proportions is 0.3254. That is, the differences in proportions are computed within each age group (young and old), and these differences are weighted by the relative frequencies of each age group, to obtain 0.3254. Monte Carlo (*see Monte Carlo Simulation*) P values based on resampling from the permutation distribution with 25 000 permutations yields a P value of 0.0017 unadjusted. This is the proportion of permutations with unadjusted test statistics at least as large as 0.4051. The proportion with adjusted test statistics at least as large as 0.3254 is 0.0036, so this is the first adjusted P value.

Next, consider only the restricted set of permutations that retain the age imbalance across treatment groups, that is, permutations in which 17 younger and 15 older cases are assigned to the placebo and 7 younger and 20 older cases are assigned to the active treatment. The number of permissible permutations can be expressed as the product of the number of ways of assigning the younger cases, $24!/([17!7!])$, and the number of ways of assigning the older cases, $35!/([15!20!])$. The proportion of Monte Carlo permutations with unadjusted test statistics at least as large as 0.4051 is 0.0086, so this is the second adjusted P value. Considering only the restricted set of permutations that retain the age imbalance across treatment groups, the proportion of permutations with adjusted test statistics at least as large as 0.3254 is 0.0082, so this is the third adjusted P value, or the doubly adjusted P value [4]. Of course, the set of numerical values of the various P values for a given set of data does not serve as a basis for selecting one adjustment technique or another. Rather, this decision should be based on the relative importance of testing and estimation, because obtaining a valid P value by comparing a distorted estimate to other equally distorted estimates does not help with valid estimation. The double adjustment technique might be ideal for ensuring both valid testing and valid estimation [4].

Table 1 A simplified $2 \times 2 \times 2$ contingency table. (Reproduced from Koch, G.G., Amara, I.A., Davis, G.W. & Gillings, D.B. (1982). A review of some statistical methods for covariance analysis of categorical data, *Biometrics* **38**, 563–595 [11])

	Younger		Older		Total
	Unfavorable	Favorable	Unfavorable	Favorable	
Placebo	15	2	11	4	32
Active	5	2	6	14	27
Total	20	4	17	18	59

Another exact permutation adjustment technique applies to ordered categorical covariates measured on the same scale as the ordered categorical response variable [8]. The idea here is to consider the information-preserving composite end point [3], which consists of the combination of the baseline value (the covariate) and the subsequent outcome measure. Instead of assigning arbitrary numerical scores and then considering a difference from baseline (as is frequently done in practice), this approach is based on a partial ordering on the set of possible values for the pair (baseline, final outcome), and then a U test.

Regardless of the scale on which the covariate is measured, it needs to be a true covariate, meaning that it is not influenced by the treatments, because adjustment for variables measured subsequent to randomization is known to lead to unreliable results [17, 18]. Covariates measured after randomization have been called *pseudocovariates* [15], and the subgroups defined by them have been called *improper subgroups* [20].

References

- [1] Akritas, M.G., Arnold, S.F. & Brunner, E. (1997). Non-parametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of the American Statistical Association* **92**, 258–265.
- [2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [3] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [4] Berger, V.W. (2005). Nonparametric adjustment techniques for binary covariates, *Biometrical Journal* In press.
- [5] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias”, *Journal of Modern Applied Statistical Methods* **2**, 80–86.
- [6] Berger, V.W. & Exner, D.V. (1999). Detecting selection bias in randomized clinical trials, *Controlled Clinical Trials* **20**, 319–327.
- [7] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**, 74–82.
- [8] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**(1), 48–55.
- [9] Brenner, H. (1998). A potential pitfall in control of covariates in epidemiologic studies, *Epidemiology* **9**(1), 68–71.
- [10] Greenland, S., Robins, J.M. & Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**, 29–46.
- [11] Koch, G.G., Amara, I.A., Davis, G.W. & Gillings, D.B. (1982). A review of some statistical methods for covariance analysis of categorical data, *Biometrics* **38**, 563–595.
- [12] Koch, G.G., Tangen, C.M., Jung, J.W. & Amara, I.A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them, *Statistics in Medicine* **17**, 1863–1892.
- [13] Lachenbruch, P.A. & Clements, P.J. (1991). ANOVA, Kruskal-Wallis, normal scores, and unequal variance, *Communications in Statistics – Theory and Methods* **20**(1), 107–126.
- [14] Moses, L.E., Emerson, J.D. & Hosseini, H. (1984). Analyzing data from ordered categories, *New England Journal of Medicine* **311**, 442–448.
- [15] Prorok, P.C., Hankey, B.F. & Bundy, B.N. (1981). Concepts and problems in the evaluation of screening programs, *Journal of Chronic Diseases* **34**, 159–171.
- [16] Rahrfs, V.W. & Zimmermann, H. (1993). Scores: Ordinal data with few categories – how should they be analyzed? *Drug Information Journal* **27**, 1227–1240.
- [17] Robins, J.M. & Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high-vs. low-dose AZT treatment arms in an AIDS randomized trial, *Journal of the American Statistical Association* **89**, 737–749.
- [18] Rosenbaum, P.R. (1984). The consequences of adjusting for a concomitant variable that has been affected by the treatment, *Journal of the Royal Statistical Society, Part A* **147**, 656–666.
- [19] Tangen, C.M. & Koch, G.G. (2000). Non-parametric covariance methods for incidence density analyses of time-to-event data from a randomized clinical trial and their complementary roles to proportional hazards regression, *Statistics in Medicine* **19**, 1039–1058.
- [20] Yusuf, S., Wittes, J., Probstfield, J. & Tyroler, H.A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials, *Journal of the American Medical Association* **266**(1), 93–98.

(See also **Stratification**)

VANCE W. BERGER

Analysis of Variance

RONALD C. SERLIN

Volume 1, pp. 52–56

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Variance

Statistical tests have been employed since the early eighteenth century, and the basic model that underlies much of modern statistics has been explored since its formulation by Legendre and **Gauss** in the early nineteenth century. Yet, it was not until the early twentieth century, when **R. A. Fisher** introduced the analysis of variance [2–4, 7], that a systematic method based on exact sampling distributions was put into place. The method allowed for inferences to be drawn from sample statistics to population characteristics of interest. The analysis of variance united methods that were previously only approximate, were developed separately, and were seemingly unrelated, into one exact procedure whose results could be reported in a single compact table.

In the inferential procedure explicated by Fisher, the actual data in an experiment are considered to be a random sample from a hypothetical infinite population that was assumed to be appropriately modeled by distributions specified by relatively few parameters [2]. For instance, in terms of a particular observable measure, the population could consist of data that follow a normal distribution. The parameters of this distribution are the mean and variance (it was Fisher who introduced the word ‘variance’ to denote the square of the standard deviation). Or, alternatively, four normally distributed populations could have been sampled, possibly having different means and variances. The question then naturally arises as to whether corresponding characteristics of the populations are equal.

In order to investigate this matter, we formulate what Fisher called a null hypothesis, usually denoted by H_0 . This hypothesis specifies a function of the parameters related to the question of interest. In the present case, we might specify the hypothesis in terms of the population means μ_k

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K, \quad (1)$$

where K is the number of populations whose means we are comparing. According to Fisher [6], the term ‘null hypothesis’ was chosen by analogy with a device used to measure electrical resistance. In this device, an indicator dial showing no deflection indicated that the correct value was determined. In the same way, if the null hypothesis is true, then any

discrepancies of the sample means from one another are due to the errors of random sampling.

Fisher at one time revealingly referred to the analysis of variance as the analysis of variation [7]. The analysis consists of partitioning the overall variability present in the data into parts reflecting the variability accounted for by explanatory factors and the variability due to chance. These different sources of variability are measured by particular sums of squared deviations. For instance, the total variability to be studied is equal to the sum of the squares of the differences between the observations and the overall sample mean. If the amount of variability accounted for by the explanatory factor was sufficiently larger than the magnitude we would expect simply because of the vagaries of random sampling, we would conclude that there was a nonzero effect in the population due to the factor under examination [7].

The basic partition of the total variability utilizes the algebraic identity

$$\begin{aligned} \sum_{i=1}^N (Y_{ik} - \bar{Y})^2 &= \sum_{k=1}^K N_k (\bar{Y}_k - \bar{Y})^2 \\ &\quad + \sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2, \end{aligned} \quad (2)$$

or

$$SST = SSB + SSW. \quad (3)$$

SST is the sum of squares total and refers to the overall sum of squared deviations of the observations about the combined group mean $\bar{Y}$. SSB is the sum of squares between groups and refers to the variability of the individual group means $\bar{Y}_k$ about the combined group mean. SSW is the sum of squares within groups and refers to the variability of the observations about their separate group means. The sum of squares between groups is that part of the total variability that reflects possible differences among the groups.

Let us also assume in the present case that the K population variances are equal. Then in sampling from K normal populations with equal variances and equal means under the null hypothesis, we are in effect sampling K samples from a single population. And if the null hypothesis is true, the variability among the sample means should be reflective of the known variance of means in repeated sampling from

2 Analysis of Variance

a single population. This variance is

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{N}, \quad (4)$$

where $\sigma_{\bar{Y}}^2$ is the variance of all possible sample means, σ^2 denotes the population variance, and N is the total sample size. With the sample estimate of a population variance given by

$$S^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{(N - 1)},$$

$\sigma_{\bar{Y}}^2$ can be estimated in the sample by

$$S_{\bar{Y}}^2 = \frac{\sum_{k=1}^K (\bar{Y}_k - \bar{Y})^2}{(K - 1)}.$$

Finally, then,

$$\begin{aligned} NS_{\bar{Y}}^2 &= N \sum_{k=1}^K \frac{(\bar{Y}_k - \bar{Y})^2}{(K - 1)} = \frac{1}{(K - 1)} \\ &\times \sum_{k=1}^K N(\bar{Y}_k - \bar{Y})^2 = \frac{SSB}{(K - 1)} \end{aligned}$$

is an estimate of the population variance if the null hypothesis is true. The ratio of SSB to $(K - 1)$ is called the *mean square* between, MSB , where $(K - 1)$ is the number of degrees of freedom associated with this variance estimate.

If the null hypothesis is true, then MSB provides an unbiased estimate of the population variance. If the null hypothesis is false, however, MSB overestimates the population variance, because the sample means will vary about their own population means and will tend to lie further from the combined group mean than they would under H_0 . We can get an indication, then, of the possible falseness of the null hypothesis by comparing the MSB to an independent estimate of the population variance.

Fortunately, we have available a second estimate of the population variance, based on the K separate sample variances. Fisher [3] showed that the best way to combine these separate estimates into a single pooled estimator of the common population variance

is to divide the sum of squares within by $(N - K)$, resulting in the mean square within,

$$MSW = \frac{\sum_{k=1}^K \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2}{(N - K)} = \frac{\sum_{k=1}^K (N_k - 1)S_k^2}{(N - K)},$$

based on $(N - K)$ degrees of freedom. Fisher showed that the mean square within is an unbiased estimate of the population variance, regardless of the truth of the null hypothesis. Finally, Fisher showed [1] that these two estimates are statistically independent of one another.

We have, then, two independent estimates of the common population variance, one of which, the mean square between, is unbiased only if the null hypothesis is true, otherwise tending to be too large. It is a remarkable feature of the analysis of variance that a test of the hypothesis concerning the equality of population means can be performed by testing the equality of two variances, σ^2 and $N\sigma_{\bar{Y}}^2$.

Fisher [4] first formulated an approximate test that both MSB and MSW estimate the same value by using the logarithm of the ratio MSB/MSW , which he labeled z . He did so for two reasons. First, the distribution of the logarithm of the sample variance approaches the normal distribution with increasing sample sizes, so he could use well-known procedures based on the normal distribution to perform the test involving the two variances. More importantly, the variability of a sample variance involves the population variance, which is unknown, whereas the variability of the logarithm of the sample variance only involves the sample size. Therefore, the unknown population variance would not enter a procedure based on the approximate distribution of z .

The population variance would need to be eliminated from the problem at hand in order for the test to be exact and not approximate. This is precisely what Fisher accomplished. Helmer had shown that for a normally distributed variable Y , with mean μ and variance σ^2 , $\sum_{i=1}^p (Y_i - \mu)^2 / \sigma^2$ follows a chi-square distribution with p degrees of freedom, denoted χ_p^2 . Pizzetti [9] showed that $\sum_{i=1}^p (Y_i - \bar{Y})^2 / \sigma^2$ follows a χ_{p-1}^2 distribution. Consider then the sum of squares between. Each of the sample means $\bar{Y}_k$ is normally distributed with variance $\sigma_{\bar{Y}}^2 = \sigma^2 / N_k$, the overall average of the sample means is $\bar{Y}$, and we have seen that under the

null hypothesis, the K samples are effectively drawn from the same population. Pizzetti's result tells us that

$$\sum_{k=1}^K \frac{(\bar{Y}_k - \bar{Y})^2}{\sigma^2/N_k} = \sum_{k=1}^K \frac{N_k(\bar{Y}_k - \bar{Y})^2}{\sigma^2} = \frac{SSB}{\sigma^2}$$

follows a χ_{K-1}^2 distribution, and so MSB/σ^2 has a distribution that is $1/(K-1)$ times a χ_{K-1}^2 distributed variable. Similarly, we find that MSW/σ^2 has a distribution that is $1/(N-K)$ times a χ_{N-K}^2 distributed variable. Finally, then, we see that the ratio

$$\frac{[MSB/\sigma^2]}{[MSW/\sigma^2]} = \frac{MSB}{MSW}$$

is distributed as the ratio of two independent chi-square distributed variables, each divided by its degrees of freedom. It is the distribution of this ratio that Fisher derived. **Snedecor** named the ratio F in Fisher's honor, reputedly [11] 'for which officiousness Fisher never seems to have forgiven him'.

The distribution of the F -ratio is actually a family of distributions. The particular distribution appropriate to the problem at hand is determined by two parameters, the number of degrees of freedom associated with the numerator and denominator estimates of the variance. As desired, the ratio of the two mean squares reflects the relative amounts of variability attributed to the explanatory factor of interest and to chance. The F distribution allows us to specify a cutoff, called a *critical value* (see **Critical Region**). F -ratios larger than the critical value lead us to conclude that [5] 'either there is something in the [mean differences], or a coincidence has occurred ...' A small percentage of the time, we can obtain a large F -ratio even when the null hypothesis is true, on the basis of which we would conclude incorrectly that H_0 is false. Fisher often set the rate at which we would commit this error, known as a Type I error, at 0.05 or 0.01; the corresponding critical values would be the 95th or 99th cumulative percentiles of the F distribution.

Neyman and **Pearson** [10] introduced the concept of an alternative hypothesis, denoted H_1 , which reflected the conclusion to be drawn regarding the population parameters if the null hypothesis were rejected. They also pointed to a second kind of error that could occur, the failure to reject a false null hypothesis, called a *Type II error*, with an

associated Type II error rate. The rate of correct rejection of H_0 is known as the power of the test. Fisher rarely acknowledged the contributions of the Neyman and Pearson method of hypothesis testing. He did, nevertheless, derive the non-null distribution of the F -ratio, which allowed the power of tests to be calculated.

As an example of an analysis of variance, consider the data reported in a study [8] of the effects of four treatments on posttraumatic stress disorder (PSD), summarized in Table 1. The dependent variable is a posttest measure of the severity of PSD. On the basis of pooling the sample variances, the MSW is found to equal 55.71. From the sample means and sample sizes, the combined group mean is calculated to be 15.62, and using these values and $K = 4$, the MSB is determined to equal 169.32. Finally, the F -ratio is found to be 3.04. These results are usually summarized in an analysis of variance table, shown in Table 2. From a table of the F distribution with numerator degrees of freedom equal to $K - 1 = 3$, denominator degrees of freedom equal to $N - K = 41$, and Type I error rate set to 0.05, we find the critical value is equal to 2.83. Because the observed F -ratio exceeds the critical value, we conclude that the test is significant and that the null hypothesis is false.

This example points to a difficulty associated with the analysis of a variance test, namely, that although

Table 1 Summary data from study of traumatic stress disorder

	Treatment group ^a			
	SIT	PE	SC	WL
N_k	14	10	11	10
$\bar{Y}_k$	11.07	15.40	18.09	19.50
S_k^2	15.76	122.99	50.84	51.55

Note: ^aSIT = stress inoculation therapy, PE = prolonged exposure, SC = supportive counseling, WL = wait-list control.

Table 2 Analysis of variance table for data in Table 1

Source	df	SS	MS	F
Between	3	507.97	169.32	3.04
Within	41	2284.13	55.71	
Total	44	2792.10		

we have concluded that the population means differ, we still do not know in what ways they differ. It seems likely that this more focused information would be particularly useful. One possible solution would involve testing the means for equality in a pairwise fashion, but this approach would engender its own problems. Most importantly, if each pairwise test were conducted with a Type I error rate of 0.05, then the rate at which we would falsely conclude that the means are not all equal could greatly exceed 0.05. Fisher introduced a method, known as a **multiple comparison procedure** for performing the desired pairwise comparisons, but it failed to hold the Type I error rate at the desired level in all circumstances. Many other multiple comparison procedures have since been developed that either bypass the F test or take advantage of its properties and successfully control the overall Type I error rate.

References

- [1] Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error, *Monthly Notes of the Royal Astronomical Society* **80**, 758–770.
- [2] Fisher, R.A. (1922a). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A* **222**, 309–368.
- [3] Fisher, R.A. (1922b). The goodness of fit of regression formulae and the distribution of regression coefficients, *Journal of the Royal Statistical Society* **85**, 597–612.
- [4] Fisher, R.A. (1924). On a distribution yielding the error functions of several well known statistics, *Proceedings of the International Congress of Mathematics* **2**, 805–813.
- [5] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture, Great Britain* **33**, 503–513.
- [6] Fisher, R.A. (1990). *Statistical Inference and Analysis: Selected Correspondence of R.A. Fisher*, J.H. Bennett, ed., Oxford University Press, London.
- [7] Fisher, R.A. & Mackenzie, W.A. (1923). Studies in crop variation II: the manurial response of different potato varieties, *Journal of Agricultural Science* **13**, 311–320.
- [8] Foa, E.B., Rothbaum, B.O., Riggs, D.S. & Murdock, T.B. (1991). Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling, *Journal of Consulting and Clinical Psychology* **59**, 715–723.
- [9] Hald, Anders. (2000). Studies in the history of probability and statistics XLVII. Pizzetti's contributions to the statistical analysis of normally distributed observations, 1891, *Biometrika* **87**, 213–217.
- [10] Neyman, J. & Pearson, E. (1933). The testing of statistical hypotheses in relation to probabilities a priori, *Proceedings of the Cambridge Philosophical Society* **29**, 492–510.
- [11] Savage, L.J. (1976). On rereading Fisher, *The Annals of Statistics* **4**, 441–500.

(See also **Generalized Linear Models (GLM); History of Multivariate Analysis of Variance; Repeated Measures Analysis of Variance**)

RONALD C. SERLIN

Analysis of Variance: Cell Means Approach

ROGER E. KIRK AND B. NEBIYOU BEKELE

Volume 1, pp. 56–66

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Variance: Cell Means Approach

Analysis of variance, ANOVA, is typically introduced in textbooks by means of the classical ANOVA model. This model for a completely randomized design is

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)} \quad (i = 1, \dots, n; j = 1, \dots, p), \quad (1)$$

where Y_{ij} is the observation for participant i in treatment level j , μ is the grand mean, α_j is the treatment effect for population j , and $\varepsilon_{i(j)}$ is the error effect that is i.i.d. $N(0, \sigma_\varepsilon^2)$. The focus of the model is on treatment effects, α_j 's. However, according to Urquhart, Weeks, and Henderson [9], **Ronald A. Fisher's** early development of ANOVA was conceptualized by his colleagues in terms of cell means, μ_j . It was not until later that a cell mean was given a linear structure in terms of the grand mean plus a treatment effect, that is, $\mu_j = \mu + \alpha_j$.

The cell means model is an alternative ANOVA model. This model for a **completely randomized design** is

$$Y_{ij} = \mu_j + \varepsilon_{i(j)} \quad (i = 1, \dots, n; j = 1, \dots, p), \quad (2)$$

where μ_j is the population mean for treatment level j and $\varepsilon_{i(j)}$ is the error effect that is i.i.d. $N(0, \sigma_\varepsilon^2)$. The cell means model replaces two parameters of the classical model, the grand mean and treatment effect, with one parameter, the cell mean. It seems that Fisher did not use either model (1) or (2). However, the cell means model is consistent with his early development of ANOVA as an analysis of differences among observed means [5, p. 12]. The advantages of the cell means model approach to the ANOVA are well documented [1–8, 10]. The advantages are most evident for multitreatment experiments with unequal cell n 's or empty cells (see **Missing Data**). Two versions of the cell means model are described: the unrestricted model and the restricted model. To simplify the presentation, a **fixed-effects model** will be assumed throughout the discussion.

Unrestricted Cell Means Model

The unrestricted model is illustrated using an experiment from Kirk [2, p. 166]. The experiment is concerned with the effects of sleep deprivation on hand steadiness. Because of space limitations, only a portion of the data is used. $N = 12$ participants were randomly assigned to $p = 3$ conditions of sleep deprivation, 12, 24, and 36 hours, with the restriction that $n = 4$ participants were assigned to each condition. The dependent variable was a measure of hand steadiness; large scores indicate lack of steadiness. The data are presented in Table 1. The null hypothesis for these data is $H_0: \mu_1 = \mu_2 = \mu_3$. Three equivalent null hypotheses that can be used with the cell means model are

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 & H_0: \mu_1 - \mu_3 &= 0 \\ \mu_2 - \mu_3 &= 0 & \mu_2 - \mu_3 &= 0 \\ H_0: \mu_1 - \mu_2 &= 0 \\ \mu_1 - \mu_3 &= 0. \end{aligned} \quad (3)$$

In matrix notation, the first null hypothesis is written as

$$\begin{array}{ccc} \text{Coefficient} & \text{Mean} & \text{Null} \\ \text{matrix} & \text{vector} & \text{vector} \\ \mathbf{C}' & \boldsymbol{\mu} & \mathbf{0} \\ \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} & \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} & = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \end{array} \quad (4)$$

where $\mathbf{C}'$ is a $(p-1) \times p$ coefficient matrix of full row rank that defines the null hypothesis, $\boldsymbol{\mu}$ is a $p \times 1$ vector of population means, and $\mathbf{0}$ is a $(p-1) \times 1$ vector of zeros.

For a completely randomized ANOVA, the total sum of square, $SSTOTAL$, can be partitioned into the sum of squares between groups, $SSBG$, and the sum of squares within groups, $SSWG$. Formulas

Table 1 Hand unsteadiness data for three conditions of sleep deprivation

a_1	a_2	a_3
12 hours	24 hours	36 hours
2	3	5
1	5	7
6	4	6
3	4	10
$\bar{Y}_{ij} = 3$	4	7

2 Analysis of Variance: Cell Means Approach

for computing these sums of squares using vectors, matrices, and a scalar are

$$\begin{aligned} SSTOTAL &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1} \\ SSBG &= (\mathbf{C}'\hat{\boldsymbol{\mu}} - \mathbf{0})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\hat{\boldsymbol{\mu}} - \mathbf{0}) \\ SSWG &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}'\mathbf{X}'\mathbf{y}, \end{aligned} \quad (5)$$

where $\mathbf{y}$ is an $N \times 1$ vector of observations and $\mathbf{X}$ is an $N \times p$ structural matrix that indicates the treatment level in which an observation appears. The $\mathbf{X}$ matrix contains ones and zeros such that each row has only one one and each column has as many ones as there are observations from the corresponding population. For the data in Table 1, $\mathbf{y}$ and $\mathbf{X}$ are

$$\begin{array}{c} \mathbf{y} \\ \mathbf{X} \end{array} \begin{array}{c} \left\{ \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right\} \begin{bmatrix} 2 \\ 1 \\ 6 \\ 3 \\ 3 \\ 5 \\ 4 \\ 4 \\ 5 \\ 7 \\ 6 \\ 10 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

$\mathbf{J}$ is an $N \times N$ matrix of ones and is obtained from the product of an $N \times 1$ vector of ones, $\mathbf{1}$, (column sum vector) and a $1 \times N$ vector of ones, $\mathbf{1}'$ (row sum vector). $\hat{\boldsymbol{\mu}}$ is a $p \times 1$ vector of sample means and is given by

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\mu}} \quad (6)$$

$$\begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 12 \\ 16 \\ 28 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 7 \end{bmatrix}.$$

$(\mathbf{X}'\mathbf{X})^{-1}$ is a $p \times p$ diagonal matrix whose elements are the inverse of the sample n 's for each treatment level.

An advantage of the cell means model is that a representation, $\mathbf{C}'\hat{\boldsymbol{\mu}} - \mathbf{0}$, of the null hypothesis, $\mathbf{C}'\boldsymbol{\mu} = \mathbf{0}$, always appears in the formula for $SSBG$. Hence there is never any ambiguity about the hypothesis that this sum of squares is used to test. Because $\mathbf{0}$ in $\mathbf{C}'\hat{\boldsymbol{\mu}} - \mathbf{0}$ is a vector of zeros, the

formula for $SSBG$ simplifies to

$$SSBG = (\mathbf{C}'\hat{\boldsymbol{\mu}})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\hat{\boldsymbol{\mu}}). \quad (7)$$

The between groups, within groups, and total sums of squares for the data in Table 1 are, respectively,

$$\begin{aligned} SSBG &= (\mathbf{C}'\hat{\boldsymbol{\mu}})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\hat{\boldsymbol{\mu}}) \\ &= 34.6667 \\ SSWG &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}'\mathbf{X}'\mathbf{y} \\ &= 326.0000 - 296.0000 = 30.0000 \\ SSTOTAL &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1} \\ &= 326.0000 - 261.3333 = 64.6667. \end{aligned} \quad (8)$$

The between and within groups mean square are given by

$$\begin{aligned} MSBG &= \frac{SSBG}{(p-1)} = \frac{34.6667}{2} = 17.3333 \\ MSWG &= \frac{SSWG}{[p(n-1)]} = \frac{30.0000}{[3(4-1)]} = 3.3333. \end{aligned} \quad (9)$$

The F statistic and P value are

$$F = \frac{MSBG}{MSWG} = \frac{17.3333}{3.3333} = 5.20, \quad P = .04. \quad (10)$$

The computation of the three sums of squares is easily performed with any computer package that performs matrix operations.

Restricted Cell Means Model

A second form of the cell means model enables a researcher to test a null hypothesis subject to one or more restrictions. This cell means model is called a restricted model. The restrictions represent assumptions about the means of the populations that are sampled. Consider a randomized block ANOVA design where it is assumed that the treatment and blocks do not interact. The restricted cell means model for this design is

$$Y_{ij} = \mu_{ij} + \varepsilon_{i(j)} \quad (i = 1, \dots, n; \quad j = 1, \dots, p), \quad (11)$$

where μ_{ij} is the population mean for block i and treatment level j , $\varepsilon_{i(j)}$ is the error effect that is i.i.d. $N(0, \sigma_\varepsilon^2)$, and μ_{ij} is subject to the restrictions that

$$\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0 \text{ for all } i, i', j, \text{ and } j'. \quad (12)$$

The restrictions on μ_{ij} state that all block-treatment interaction effects equal zero. These restrictions, which are a part of the model, are imposed when the cell n_{ij} 's are equal to one as in a randomized block design and it is not possible to estimate error effects separately from interaction effects.

Consider a randomized block design with $p = 3$ treatment levels and $n = 4$ blocks. Only four blocks

$$\begin{array}{ccc} \text{Hypothesis} & \text{Mean} & \text{Null} \\ \text{matrix} & \text{vector} & \text{vector} \\ \mathbf{H}'_{BL} & \begin{bmatrix} \mu_{BL} \\ \mu_{1\cdot} \\ \mu_{2\cdot} \\ \mu_{3\cdot} \\ \mu_{4\cdot} \end{bmatrix} & \begin{bmatrix} \mathbf{0}_{BL} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array} \quad (14)$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{BL} \\ \mu_{1\cdot} \\ \mu_{2\cdot} \\ \mu_{3\cdot} \\ \mu_{4\cdot} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The randomized block design has $np = h$ sample means. The $1 \times h$ vector of means is $\boldsymbol{\mu}' = [\hat{\mu}_{11} \hat{\mu}_{21} \hat{\mu}_{31} \hat{\mu}_{41} \hat{\mu}_{12} \hat{\mu}_{22}, \dots, \hat{\mu}_{43}]$. The coefficient matrices for computing sums of squares for treatment A , blocks, and the block-treatment interaction are denoted by $\mathbf{C}'_A$, $\mathbf{C}'_{BL}$, and $\mathbf{R}'$, respectively. These coefficient matrices are easily obtained from Kronecker products, $\otimes$, as follows:

$$\begin{aligned} & \mathbf{H}'_A \otimes \frac{1}{n} \mathbf{1}'_{BL} = \frac{1}{n} \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix} \\ & \frac{1}{p} \mathbf{1}'_A \otimes \mathbf{H}'_{BL} = \frac{1}{p} \begin{bmatrix} 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix} \quad (15) \\ & \mathbf{H}'_A \otimes \mathbf{H}'_{BL} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{bmatrix}, \end{aligned}$$

are used because of space limitations; ordinarily a researcher would use many more blocks. The null hypotheses for treatment A and blocks are, respectively,

$$\begin{array}{ll} H_0: \mu_{1\cdot} - \mu_{2\cdot} = 0 & H_0: \mu_{1\cdot} - \mu_{2\cdot} = 0 \\ \mu_{2\cdot} - \mu_{3\cdot} = 0 & \mu_{2\cdot} - \mu_{3\cdot} = 0 \\ \mu_{3\cdot} - \mu_{4\cdot} = 0 & \mu_{3\cdot} - \mu_{4\cdot} = 0 \end{array} \quad (13)$$

In matrix notation, the hypotheses can be written as

$$\begin{array}{ccc} \text{Hypothesis} & \text{Mean} & \text{Null} \\ \text{matrix} & \text{vector} & \text{vector} \\ \mathbf{H}'_A & \begin{bmatrix} \mu_A \\ \mu_{1\cdot} \\ \mu_{2\cdot} \\ \mu_{3\cdot} \end{bmatrix} & \begin{bmatrix} \mathbf{0}_A \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{array}$$

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{1\cdot} \\ \mu_{2\cdot} \\ \mu_{3\cdot} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $\mathbf{1}'_{BL}$ is a $1 \times n$ vector of ones and $\mathbf{1}'_A$ is a $1 \times p$ vector of ones. The null hypothesis for treatment A can be expressed as $1/n \mathbf{C}'_A \boldsymbol{\mu} = \mathbf{0}_A$, where $\boldsymbol{\mu}$ is an $h \times 1$ vector of population means and $\mathbf{0}_A$ is a $(p-1) \times 1$ vector of zeros. The restrictions on the model, $\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$ for all i, i', j , and j' , can be expressed as $\mathbf{R}' \boldsymbol{\mu} = \boldsymbol{\theta}_A$, where $\boldsymbol{\theta}_A$ is an $s = h - p - n + 1$ vector of zeros. Without any loss of generality, the fractions in the $1/n \mathbf{C}'_A$ and $1/p \mathbf{C}'_{BL}$ matrices can be eliminated by replacing the fractions with 1 and -1 . To test $\mathbf{C}'_A \boldsymbol{\mu} = \mathbf{0}_A$ subject to the restrictions that $\mathbf{R}' \boldsymbol{\mu} = \boldsymbol{\theta}_A$, we can form an augmented treatment matrix and an augmented vector of zeros as follows:

$$\mathbf{Q}'_A = \begin{bmatrix} \mathbf{R}' \\ \mathbf{C}'_A \end{bmatrix} \quad \boldsymbol{\eta}_A = \begin{bmatrix} \boldsymbol{\theta}_A \\ \mathbf{0}_A \end{bmatrix}, \quad (16)$$

4 Analysis of Variance: Cell Means Approach

where $\mathbf{Q}'_A$ consists of the s rows of the $\mathbf{R}'$ matrix and the $p - 1$ rows of the $\mathbf{C}'_A$ matrix that are not identical to the rows of $\mathbf{R}'$, inconsistent with them, or linearly dependent on them and $\boldsymbol{\eta}_A$ is an $s + p - 1$ vector of zeros. The joint null hypothesis $\mathbf{Q}'_A \boldsymbol{\mu} = \boldsymbol{\eta}_A$ combines the restrictions that all interactions are equal to zero with the hypothesis that differences among the treatment means are equal to zero. The sum of squares that is used to test this joint null hypothesis is

$$SSA = (\mathbf{Q}'_A \hat{\boldsymbol{\mu}})' (\mathbf{Q}'_A \mathbf{Q}_A)^{-1} (\mathbf{Q}'_A \hat{\boldsymbol{\mu}}) - SSRES, \quad (17)$$

where $SSRES = (\mathbf{R}' \hat{\boldsymbol{\mu}})' (\mathbf{R}' \mathbf{R})^{-1} (\mathbf{R}' \hat{\boldsymbol{\mu}})$. To test hypotheses about contrasts among treatment means, restricted cell means rather than unrestricted cell means should be used. The vector of restricted cell means, $\hat{\boldsymbol{\mu}}_R$, is given by

$$\hat{\boldsymbol{\mu}}_R = \hat{\boldsymbol{\mu}} - \mathbf{R}(\mathbf{R}' \mathbf{R})^{-1} \mathbf{R}' \hat{\boldsymbol{\mu}}. \quad (18)$$

The formula for computing the sum of squares for blocks follows the pattern described earlier for treatment A . We want to test the null hypothesis for blocks, $\mathbf{C}'_{BL} \boldsymbol{\mu} = \mathbf{0}_{BL}$, subject to the restrictions that $\mathbf{R}' \boldsymbol{\mu} = \boldsymbol{\theta}_{BL}$. We can form an augmented matrix for blocks and an augmented vector of zeros as follows:

$$\mathbf{Q}'_{BL} = \begin{bmatrix} \mathbf{R}' \\ \mathbf{C}'_{BL} \end{bmatrix} \quad \boldsymbol{\eta}_{BL} = \begin{bmatrix} \boldsymbol{\theta}_{BL} \\ \mathbf{0}_{BL} \end{bmatrix}, \quad (19)$$

where $\mathbf{Q}'_{BL}$ consists of the s rows of the $\mathbf{R}'$ matrix and the $n - 1$ rows of the $\mathbf{C}'_{BL}$ matrix that are not identical to the rows of $\mathbf{R}'$, inconsistent with them, or linearly dependent on them, and $\boldsymbol{\eta}_{BL}$ is an $s + n - 1$ vector of zeros. The joint null hypothesis $\mathbf{Q}'_{BL} \boldsymbol{\mu} = \boldsymbol{\eta}_{BL}$ combines the restrictions that all interactions are equal to zero with the hypothesis that differences among the block means are equal to zero. The sum of squares that is used to test this joint null hypothesis is

$$SSBL = (\mathbf{Q}'_{BL} \hat{\boldsymbol{\mu}})' (\mathbf{Q}'_{BL} \mathbf{Q}_{BL})^{-1} (\mathbf{Q}'_{BL} \hat{\boldsymbol{\mu}}) - SSRES. \quad (20)$$

The total sum of squares is given by

$$SSTOTAL = \hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}' \hat{\mathbf{J}} \hat{\boldsymbol{\mu}} h^{-1}. \quad (21)$$

The treatment and block mean squares are given by, respectively, $MSA = SSA/(p - 1)$ and $MSBL = SSBL/[(n - 1)(p - 1)]$. The F statistics are $F = MSA/MSRES$ and $F = MSBL/MSRES$.

The previous paragraphs have provided an overview of the restricted cell means model. For many restricted models, the formulas for computing SSA and $SSBL$ can be simplified. The simplified formulas are described by Kirk [2, pp. 290–297]. An important advantage of the cell means model is that it can be used when observations are missing; the procedures for a randomized block design are described by Kirk [2, pp. 297–301]. Another advantage of the cell means model that we will illustrate in the following section is that it can be used when there are empty cells in a multitreatment design.

Unrestricted Cell Means Model for a Completely Randomized Factorial Design

The expectation of the classical sum of squares model equation for a two-treatment, completely randomized factorial design is

$$E(Y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

$$(i = 1, \dots, n; j = 1, \dots, p; k = 1, \dots, q), \quad (22)$$

where Y_{ijk} is the observation for participant i in treatment combination $a_j b_k$, μ is the grand mean, α_j is the treatment effect for population j , β_k is the treatment effect for population k , and $(\alpha\beta)_{jk}$ is the interaction of treatment levels j and k . If treatments A and B each have three levels, the classical model contains 16 parameters: $\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, (\alpha\beta)_{11}, (\alpha\beta)_{12}, \dots, (\alpha\beta)_{33}$. However, only nine cell means are available to estimate these parameters. Thus, the model is overparameterized – it contains more parameters than there are means from which to estimate them. Statisticians have developed a number of ways to get around this problem [4, 5, 8]. Unfortunately, the solutions do not work well when there are missing observations or empty cells.

The cell means model equation for a two-treatment, completely randomized factorial design is

$$Y_{ijk} = \mu_{jk} + \varepsilon_{i(jk)} \quad (i = 1, \dots, n; j = 1, \dots, p; k = 1, \dots, q), \quad (23)$$

where μ_{jk} is the population mean for treatment combination $a_j b_k$ and $\varepsilon_{i(jk)}$ is the error effect that is i.i.d. $N(0, \sigma_e^2)$. This model has none of the problems associated with overparameterization. A population mean can be estimated for each cell that contains

one or more observations. Thus, the model is fully parameterized. And unlike the classical ANOVA model, the cell means model does not impose a structure on the analysis of data. Consequently, the model can be used to test hypotheses about any linear combination of cell means. It is up to the researcher to decide which tests are meaningful or useful based on the original research hypotheses, the way the experiment was conducted, and the data that are available. As we will show, however, if one or more cells are empty, linear combinations of cell means must be carefully chosen because some tests may be uninterpretable.

An experiment described by Kirk [2, pp. 367–370] is used to illustrate the computational procedures for the cell means model. The experiment examined the effects of $p = 3$ levels of treatments A and $q = 3$ levels of treatment B on $N = 45$ police recruits' attitudes toward minorities. The attitude data are shown in Table 2.

Using matrices and vectors, the null hypotheses for treatments A and B can be expressed as

$$\begin{aligned} \mathbf{H}'_A \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_A \\ \mu_{1.} \\ \mu_{2.} \\ \mu_{3.} \end{bmatrix} &= \begin{bmatrix} 0_A \\ 0 \end{bmatrix} \\ \mathbf{H}'_B \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_B \\ \mu_{.1} \\ \mu_{.2} \\ \mu_{.3} \end{bmatrix} &= \begin{bmatrix} 0_B \\ 0 \end{bmatrix}. \end{aligned} \quad (24)$$

The formulas for computing the sums of squares are

$$SSTOTAL = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1}$$

$$SSA = (\mathbf{C}'_A \hat{\boldsymbol{\mu}})'[\mathbf{C}'_A(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_A]^{-1}(\mathbf{C}'_A \hat{\boldsymbol{\mu}})$$

$$SSB = (\mathbf{C}'_B \hat{\boldsymbol{\mu}})'[\mathbf{C}'_B(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_B]^{-1}(\mathbf{C}'_B \hat{\boldsymbol{\mu}})$$

$$SSA \times B = (\mathbf{C}'_{AB} \hat{\boldsymbol{\mu}})'[\mathbf{C}'_{AB}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{AB}]^{-1}(\mathbf{C}'_{AB} \hat{\boldsymbol{\mu}})$$

$$SSWCELL = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}'\mathbf{X}'\mathbf{y}, \quad (25)$$

where $\mathbf{y}$ is an $N \times 1$ vector of observations and $\mathbf{X}$ is an $N \times pq$ structural matrix that indicates the treatment level in which an observation appears. For the data in Table 2, $\mathbf{y}$ and $\mathbf{X}$ are

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix} 24 \\ 33 \\ 37 \\ 29 \\ 42 \\ 44 \\ 36 \\ 25 \\ 27 \\ 43 \\ \vdots \\ 42 \\ 52 \\ 53 \\ 49 \\ 64 \end{bmatrix} \\ \mathbf{X} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

$\mathbf{J}$ is an $N \times N$ matrix of ones. $\hat{\boldsymbol{\mu}}$ is a $pq \times 1$ vector of sample means and is given by $\hat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [33, 35, 38, 30, 31, \dots, 52]'$. The coefficient matrices for computing sums of squares are obtained using Kronecker products as follows:

$$\begin{aligned} \mathbf{H}'_A \otimes \frac{1}{q}\mathbf{1}'_B &= \frac{1}{q} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \frac{1}{q} \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix} \\ \frac{1}{p}\mathbf{1}'_A \otimes \mathbf{H}'_B &= \frac{1}{p} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} = \frac{1}{p} \begin{bmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{bmatrix} \\ \mathbf{H}'_A \otimes \mathbf{H}'_B &= \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix}. \end{aligned} \quad (26)$$

6 Analysis of Variance: Cell Means Approach

Table 2 Attitude data for 45 police recruits

a_1 b_1	a_1 b_2	a_1 b_3	a_2 b_1	a_2 b_2	a_2 b_3	a_3 b_1	a_3 b_2	a_3 b_3
24	44	38	30	35	26	21	41	42
33	36	29	21	40	27	18	39	52
37	25	28	39	27	36	10	50	53
29	27	47	26	31	46	31	36	49
42	43	48	34	22	45	20	34	64
$\sum_{i=1}^n \bar{Y}_{.jk} =$								
33	35	38	30	31	36	20	40	52

Without any loss of generality, the fractions in $1/q\mathbf{C}'_A$ and $1/p\mathbf{C}'_B$ can be eliminated by replacing the fractions with 1 and -1 . The sums of squares and mean squares for the data in Table 2 are

$$\begin{aligned}
 SSA &= 190.000 & MSA &= \frac{95.000}{(3-1)} = 95.000 \\
 SSB &= 1543.333 & MSB &= \frac{1543.333}{(3-1)} = 771.667 \\
 SSA \times B &= 1236.667 & MSA \times B &= \frac{1236.667}{(3-1)(3-1)} \\
 & & &= 309.167 \\
 SSWCELL &= 2250.000 & MSWCELL &= \frac{2250.000}{(45-9)} \\
 & & &= 62.500 \\
 SSTOTAL &= 5220.000 & & (27)
 \end{aligned}$$

The F statistics and P values for treatments A and B , and the $A \times B$ interaction are, respectively, $F = MSA/MSWCELL = 1.52$, $p = .23$; $F = MSB/MSWCELL = 12.35$, $p < .0001$; $F = MSA \times B/MSWCELL = 4.95$, $p = .003$.

Because the $A \times B$ interaction is significant, some researchers would perform tests of simple main-effects (see **Interaction Effects**). The computations are easy to perform with the cell means model. The null hypothesis for the simple main-effects of treatment A at b_1 , for example, is

$$\begin{aligned}
 H_0: \mu_{11} - \mu_{21} &= 0 \\
 \mu_{21} - \mu_{31} &= 0.
 \end{aligned} \quad (28)$$

The coefficient matrix for computing the sums of squares of treatment A at b_1 is obtained using the

Kronecker product as follows:

$$\begin{aligned}
 &\mathbf{H}'_A \\
 &\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{c}'_{A \text{ at } b_1} \\ 1 & 0 & 0 \end{bmatrix} \\
 &\mathbf{C}'_{A \text{ at } b_1} \\
 &= \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}. \quad (29)
 \end{aligned}$$

The coefficient vector $\mathbf{c}'_{A \text{ at } b_1} = [1 \ 0 \ 0]$ selects the first level of treatment B . The second level of treatment B can be selected by using the coefficient vector $\mathbf{C}'_{A \text{ at } b_2} = [0 \ 1 \ 0]$, and so on. The simple main-effects sums of squares of treatment A at b_1 is

$$\begin{aligned}
 SSA \text{ at } b_1 &= (\mathbf{C}'_{A \text{ at } b_1} \hat{\boldsymbol{\mu}})' [\mathbf{C}'_{A \text{ at } b_1} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_{A \text{ at } b_1}]^{-1} \\
 &\times (\mathbf{C}'_{A \text{ at } b_1} \hat{\boldsymbol{\mu}}) = 463.3333. \quad (30)
 \end{aligned}$$

The F statistic and Bonferroni adjusted P value [2, p. 381] are $F = [SSA \text{ at } b_1 / (p-1)] / MSWCELL = 231.167 / 62.500 = 3.71$, $p > .05$.

When the $A \times B$ interaction is significant, rather than performing tests of simple main-effects, we prefer to test treatment-contrast interactions and contrast-contrast interactions. The significant $A \times B$ interaction indicates that at least one contrast for the treatment B means interacts with treatment A and vice versa. Suppose that after examining the data, we want to determine if the treatment B contrast denoted by $\psi_{1(B)}$ interacts with treatment A , where $\psi_{1(B)} = \mathbf{c}'_{1(B)} \boldsymbol{\mu}_B$ and $\mathbf{c}'_{1(B)} = [1 \ -1 \ 0]$. The null hypothesis is $H_0: \alpha \times \psi_{1(B)} = \delta$ for all j , where δ is a constant for a given hypothesis. The coefficient matrix for computing the sum of squares for this treatment-contrast interaction is as follows:

$$\begin{aligned}
 &\mathbf{H}'_A \\
 &\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{c}'_{1(B)} \\ 1 & -1 & 0 \end{bmatrix} \\
 &\mathbf{C}'_{A \times \psi_{1(B)}} \\
 &= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \end{bmatrix}. \quad (31)
 \end{aligned}$$

The treatment-contrast interaction sum of squares is

$$\begin{aligned}
 SSA \psi_B &= (\mathbf{C}'_{A \times \psi_{1(B)}} \hat{\boldsymbol{\mu}})' [\mathbf{C}'_{A \times \psi_{1(B)}} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_{A \times \psi_{1(B)}}]^{-1} \\
 &\times (\mathbf{C}'_{A \times \psi_{1(B)}} \hat{\boldsymbol{\mu}}) = 571.6667. \quad (32)
 \end{aligned}$$

The F statistic and simultaneous test procedure adjusted P value [2, p. 381] are $F = [SSA \psi_B / (p -$

1)]/MSWCELL = 285 833/62.500 = 4.57, $p > .05$. Because the test is not significant, the null hypothesis that $\alpha \times \psi_{1(B)} = \delta$ for all j remains tenable.

Cell Means Model with Missing Observations and Empty Cells

The cell means model is especially useful for analyzing data when the cell n_{jk} 's are unequal and one or more cells are empty. A researcher can test hypotheses about any linear combination of population means that can be estimated from the data. The challenge facing the researcher is to formulate interesting and interpretable null hypotheses using the means that are available.

Suppose that for reasons unrelated to the treatments, observation Y_{511} in Table 2 is missing. When the cell n_{jk} 's are unequal, the researcher has a choice between computing unweighted means, μ , or weighted means, $\bar{\mu}$ (see **Analysis of Variance: Multiple Regression Approaches**). Unweighted means are simple averages of cell means. Hypotheses for these means were described in the previous section. For treatments A and B , the means are given by, respectively,

$$\mu_{j\cdot} = \sum_{k=1}^q \frac{\mu_{jk}}{q} \text{ and } \mu_{\cdot k} = \sum_{j=1}^p \frac{\mu_{jk}}{p}. \quad (33)$$

general, the use of weighted means is not recommended unless the sample n_{jk} 's are proportional to the population n_{jk} 's.

For the case in which observation Y_{511} in Table 2 is missing, null hypotheses for treatment A using unweighted and weighted means are, respectively,

$$\begin{aligned} H_0: \frac{\mu_{11} + \mu_{12} + \mu_{13}}{3} - \frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} &= 0 \\ \frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} - \frac{\mu_{31} + \mu_{32} + \mu_{33}}{3} &= 0 \end{aligned} \quad (35)$$

and

$$\begin{aligned} H_0: \frac{4\mu_{11} + 5\mu_{12} + 5\mu_{13}}{14} \\ - \frac{5\mu_{21} + 5\mu_{22} + 5\mu_{23}}{15} &= 0 \\ \frac{5\mu_{21} + 5\mu_{22} + 5\mu_{23}}{15} - \frac{5\mu_{31} + 5\mu_{32} + 5\mu_{33}}{15} &= 0. \end{aligned} \quad (36)$$

The coefficients of the unweighted means are $\pm 1/q$ and 0; the coefficients of the weighted means are $\pm n_{jk}/n_j$ and 0. The unweighted and weighted coefficient matrices and sums of squares are, respectively,

$$\begin{aligned} \mathbf{C}'_{1(A)} &= \begin{bmatrix} 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 \end{bmatrix} \\ SSA &= (\mathbf{C}'_{1(A)}\hat{\boldsymbol{\mu}})'[\mathbf{C}'_{1(A)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{1(A)}]^{-1}(\mathbf{C}'_{1(A)}\hat{\boldsymbol{\mu}}) = 188.09 \\ \mathbf{C}'_{2(A)} &= \begin{bmatrix} 4/14 & 5/14 & 5/14 & -5/15 & -5/15 & -5/15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5/15 & 5/15 & 5/15 & -5/15 & -5/15 & -5/15 \end{bmatrix} \\ SSA &= (\mathbf{C}'_{2(A)}\hat{\boldsymbol{\mu}})'[\mathbf{C}'_{2(A)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{2(A)}]^{-1}(\mathbf{C}'_{2(A)}\hat{\boldsymbol{\mu}}) = 187.51 \end{aligned} \quad (37)$$

Weighted means are weighted averages of cell means in which the weights are the sample n_{jk} 's,

$$\bar{\mu}_{j\cdot} = \sum_{k=1}^q \frac{n_{jk}\mu_{jk}}{n_{j\cdot}} \text{ and } \bar{\mu}_{\cdot k} = \sum_{j=1}^p \frac{n_{jk}\mu_{jk}}{n_{\cdot k}}. \quad (34)$$

The value of weighted means is affected by the sample n_{jk} 's. Hence, the means are data dependent. In

with 2 degrees of freedom, the number of rows in $\mathbf{C}'_{1(A)}$ and $\mathbf{C}'_{2(A)}$.

Null hypotheses for treatment B using unweighted and weighted means are, respectively,

$$H_0: \frac{\mu_{11} + \mu_{21} + \mu_{31}}{3} - \frac{\mu_{12} + \mu_{22} + \mu_{32}}{3} = 0$$

8 Analysis of Variance: Cell Means Approach

$$\frac{\mu_{12} + \mu_{22} + \mu_{32}}{3} - \frac{\mu_{13} + \mu_{23} + \mu_{33}}{3} = 0 \quad (38)$$

and

$$\begin{aligned} H_0: & \frac{4\mu_{11} + 5\mu_{21} + 5\mu_{31}}{14} \\ & - \frac{5\mu_{12} + 5\mu_{22} + 5\mu_{32}}{15} = 0 \\ & \frac{5\mu_{12} + 5\mu_{22} + 5\mu_{32}}{15} - \frac{5\mu_{13} + 5\mu_{23} + 5\mu_{33}}{15} = 0. \end{aligned} \quad (39)$$

The coefficients of the unweighted means are $\pm 1/p$ and 0; the coefficients of the weighted means are $\pm n_{jk}/n_{\cdot k}$ and 0. The unweighted and weighted coefficient matrices and sums of squares are, respectively,

$$\begin{aligned} \mathbf{C}'_{1(B)} &= \begin{bmatrix} 1/3 & -1/3 & 0 & 1/3 & -1/3 & 0 & 1/3 & -1/3 & 0 \\ 0 & 1/3 & -1/3 & 0 & 1/3 & -1/3 & 0 & 1/3 & -1/3 \end{bmatrix} \\ SSB &= (\mathbf{C}'_{1(B)}\hat{\boldsymbol{\mu}})'[\mathbf{C}'_{1(B)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{1(B)}]^{-1}(\mathbf{C}'_{1(B)}\hat{\boldsymbol{\mu}}) = 1649.29 \\ \mathbf{C}'_{2(B)} &= \begin{bmatrix} 4/14 & -5/15 & 0 & 5/14 & -5/15 & 0 & 5/14 & -5/15 & 0 \\ 0 & 5/15 & -5/15 & 0 & 5/15 & -5/15 & 0 & 5/15 & -5/15 \end{bmatrix} \\ SSB &= (\mathbf{C}'_{2(B)}\hat{\boldsymbol{\mu}})'[\mathbf{C}'_{2(B)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{2(B)}]^{-1}(\mathbf{C}'_{2(B)}\hat{\boldsymbol{\mu}}) = 1713.30, \end{aligned} \quad (40)$$

with 2 degrees of freedom, the number of rows in $\mathbf{C}'_{1(B)}$ and $\mathbf{C}'_{2(B)}$.

When one or more cells are empty, the analysis of the data is more challenging. Consider the police attitude data in Table 3 where an observation Y_{511}

Table 3 Police recruit attitude data, observation Y_{511} is missing and cells a_1b_3 and a_2b_2 are empty

a_1 b_1	a_1 b_2	a_1 b_3	a_2 b_1	a_2 b_2	a_2 b_3	a_3 b_1	a_3 b_2	a_3 b_3
24	44		30		26	21	41	42
33	36		21		27	18	39	52
37	25		39		36	10	50	53
29	27		26		46	31	36	49
	43		34		45	20	34	64
$\sum_{i=1}^n \bar{Y}_{jk} = 30.75$	35		30		36	20	40	52

in cell a_1b_1 is missing and cells a_1b_3 and a_2b_2 are empty. The experiment was designed to test along with other hypotheses, the following null hypothesis for treatment A

$$\begin{aligned} H_0: & \frac{\mu_{11} + \mu_{12} + \mu_{13}}{3} - \frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} = 0 \\ & \frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} - \frac{\mu_{31} + \mu_{32} + \mu_{33}}{3} = 0. \end{aligned} \quad (41)$$

Unfortunately, this hypothesis is untestable because μ_{13} and μ_{22} cannot be estimated. The hypothesis

$$\begin{aligned} H_0: & \frac{\mu_{11} + \mu_{12}}{2} - \frac{\mu_{21} + \mu_{23}}{2} = 0 \\ & \frac{\mu_{21} + \mu_{23}}{2} - \frac{\mu_{31} + \mu_{32} + \mu_{33}}{3} = 0 \end{aligned} \quad (42)$$

is testable because data are available to estimate each of the population means. However, the hypothesis is uninterpretable because different levels of treatment B appear in each row: (b_1 and b_2) versus (b_1 and b_3) in the first row and (b_1 and b_3) versus (b_1 , b_2 , and b_3) in the second row. The following hypothesis is both testable and interpretable.

$$\begin{aligned} H_0: & \frac{\mu_{11} + \mu_{12}}{2} - \frac{\mu_{31} + \mu_{32}}{2} = 0 \\ & \frac{\mu_{21} + \mu_{23}}{2} - \frac{\mu_{31} + \mu_{33}}{2} = 0. \end{aligned} \quad (43)$$

For a hypothesis to be interpretable, the estimators of population means for each contrast in the hypothesis should share the same levels of the other treatment(s). For example, to estimate $\mu_{1\cdot} = 1/2(\mu_{11} + \mu_{12})$ and $\mu_{3\cdot} = 1/2(\mu_{31} + \mu_{32})$, it is necessary to average over b_1 and b_2 and ignore b_3 . The

null hypothesis for treatment A can be expressed in matrix notation as $\mathbf{C}'_A \boldsymbol{\mu} = \mathbf{0}$, where

$$\mathbf{C}'_A = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & -1/2 & -1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & -1/2 & 0 & -1/2 \end{bmatrix}$$

$$\boldsymbol{\mu}' = [\mu_{11} \ \mu_{12} \ \mu_{21} \ \mu_{23} \ \mu_{31} \ \mu_{32} \ \mu_{33}]. \quad (44)$$

The fractions in $\mathbf{C}'_A$ can be eliminated by replacing the fractions with 1 and -1 . For the data in Table 3 where Y_{511} is missing and two of the cells are empty, the sum of squares for testing (43) is

$$SSA = (\mathbf{C}'_A \hat{\boldsymbol{\mu}})' [\mathbf{C}'_A (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_A]^{-1} (\mathbf{C}'_A \hat{\boldsymbol{\mu}}) = 110.70 \quad (45)$$

with 2 degrees of freedom, the number of rows in $\mathbf{C}'_A$.

Testable and interpretable hypotheses for treatment B and the $A \times B$ interaction are, respectively,

$$H_0: \frac{1}{2}(\mu_{11} + \mu_{31}) - \frac{1}{2}(\mu_{12} + \mu_{32}) = 0 \quad (\text{Treatment } B)$$

$$\frac{1}{2}(\mu_{21} + \mu_{31}) - \frac{1}{2}(\mu_{23} + \mu_{33}) = 0 \quad (46)$$

and

$$H_0: \mu_{11} - \mu_{31} - \mu_{12} + \mu_{32} = 0 \quad (A \times B \text{ interaction})$$

$$\mu_{21} - \mu_{31} - \mu_{23} + \mu_{33} = 0. \quad (47)$$

If there were no empty cells, the null hypothesis for the $A \times B$ interaction would have $h - p - q + 1 = 9 - 3 - 3 + 1 = 4$ interaction terms. However, because of the empty cells, only two of the interaction terms can be tested. If the null hypothesis for the $A \times B$ interaction is rejected, we can conclude that at least one function of the form $\mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k'}$ does not equal zero. However, failure to reject the null hypothesis does not imply that all functions of the form $\mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k'}$ equal zero because we are unable to test two of the interaction terms.

When cells are empty, it is apparent that to test hypotheses, the researcher must be able to state the hypotheses in terms of linearly independent contrasts in $\mathbf{C}'\boldsymbol{\mu}$. Thus, the researcher is forced to consider what hypotheses are both interesting and interpretable. This is not the case when the classical

ANOVA model is used. When this model is used, the hypotheses that are tested are typically left to a computer package, and the researcher is seldom aware of exactly what hypotheses are being tested.

Unrestricted Cell Means Model for ANCOVA

The cell means model can be used to perform an **analysis of covariance** (ANCOVA). This application of the model is described using a completely randomized ANCOVA design with N observations, p treatment levels, and one covariate. The adjusted between-groups sum of squares, A_{adj} , and the adjusted within-groups sum of squares, E_{adj} , for a completely randomized ANCOVA design are given by

$$A_{adj} = (A_{yy} + E_{yy}) - \frac{(A_{zy} + E_{zy})^2}{A_{zz} + E_{zz}} - E_{adj} \text{ and}$$

$$E_{adj} = E_{yy} - \frac{(E_{zy})^2}{E_{zz}}. \quad (48)$$

The sums of squares in the formula, A_{yy} , E_{yy} , A_{zy} , and so on, can be expressed in matrix notation using the cell means model by defining

$$\begin{aligned} A_{yy} &= (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_y)' (\mathbf{C}'_A (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_A)^{-1} (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_y) \\ A_{zy} &= (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_z)' (\mathbf{C}'_A (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_A)^{-1} (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_y) \\ A_{zz} &= (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_z)' (\mathbf{C}'_A (\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}_A)^{-1} (\mathbf{C}'_A \hat{\boldsymbol{\mu}}_z) \\ \hat{\boldsymbol{\mu}}_y &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ \hat{\boldsymbol{\mu}}_z &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{z} \\ E_{yy} &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}_y' \mathbf{X}'\mathbf{y} \\ E_{zy} &= \mathbf{z}'\mathbf{y} - \hat{\boldsymbol{\mu}}_z' \mathbf{X}'\mathbf{y} \\ E_{zz} &= \mathbf{z}'\mathbf{z} - \hat{\boldsymbol{\mu}}_z' \mathbf{X}'\mathbf{z}, \end{aligned} \quad (49)$$

where $\hat{\boldsymbol{\mu}}_y$ is a $p \times 1$ vector of dependent variable cell means, $\mathbf{X}$ is a $N \times p$ structural matrix, $\hat{\boldsymbol{\mu}}_z$ is a $p \times 1$ vector of covariate means, $\mathbf{y}$ is an $N \times 1$ vector of dependent variable observations, and $\mathbf{z}$ is an $N \times 1$ vector of covariates. The adjusted between- and within-groups mean squares are given by, respectively,

$$MSA_{adj} = \frac{A_{adj}}{(p-1)} \text{ and } MSE_{adj} = \frac{E_{adj}}{(N-p-1)}. \quad (50)$$

The F statistic is $F = MSA_{adj}/MSE_{adj}$ with $p - 1$ and $N - p - 1$ degrees of freedom.

The cell means model can be extended to other ANCOVA designs and to designs with multiple covariates (see **Analysis of Covariance**). Lack of space prevents a description of the computational procedures.

Some Advantages of the Cell Means Model

The simplicity of the cell means model is readily apparent. There are only two models for all ANOVA designs: an unrestricted model and a restricted model. Furthermore, only three kinds of computational formulas are required to compute all sums of squares: treatment and interaction sums of squares have the general form $(\mathbf{C}'\hat{\boldsymbol{\mu}})'[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}(\mathbf{C}'\hat{\boldsymbol{\mu}})$, the within-groups and within-cells sums of squares have the form $\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\mu}}'\mathbf{X}'\mathbf{y}$, and the total sum of squares has the form $\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1}$.

The cell means model has an important advantage relative to the classical overparameterized model: the ease with which experiments with missing observations and empty cells can be analyzed. In the classical model, questions arise regarding which functions are estimable and which hypotheses are testable. However, these are nonissues with the cell means model. There is never any confusion about what functions of the means are estimable and what their best linear unbiased estimators are. And it is easy to discern which hypotheses are testable. Furthermore, the cell means model is always of full rank with the $\mathbf{X}'\mathbf{X}$ matrix being a diagonal matrix whose elements are the cell sample sizes. The number of parameters in the model exactly equals the number of cells that contain one or more observations.

There is never any confusion about what null hypotheses are tested by treatment and interaction

mean squares. The researcher specifies the hypothesis of interest when the contrasts in $\mathbf{C}'\boldsymbol{\mu}$ are specified. Hence, a sample representation of the null hypothesis always appears in formulas for treatment and interaction mean squares. Finally, the cell means model gives the researcher great flexibility in analyzing data because hypotheses about any linear combination of available cell means can be tested (see **Multiple Comparison Procedures**).

References

- [1] Hocking, R.R. & Speed, F.M. (1975). A full rank analysis of some linear model problems, *Journal of the American Statistical Association* **70**, 706–712.
- [2] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [3] Milliken, G.A. & Johnson, D.E. (1984). *Analysis of Messy Data*, Vol. 1: *Designed Experiment*, Wadsworth, Belmont.
- [4] Searle, S.R. (1971). *Linear Models*, Wiley, New York.
- [5] Searle, S.R. (1987). *Linear Models for Unbalanced Data*, Wiley, New York.
- [6] Speed, F.M. (June 1969). A new approach to the analysis of linear models. NASA Technical Memorandum, NASA TM X-58030.
- [7] Speed, F.M., Hocking, R.R. & Hackney, O.P. (1978). Methods of analysis of linear models with unbalanced data, *Journal of the American Statistical Association* **73**, 105–112.
- [8] Timm, N.H. & Carlson, J.E. (1975). Analysis of variance through full rank models, *Multivariate Behavioral Research Monographs*. No. 75-1.
- [9] Urquhart, N.S., Weeks, D.L. & Henderson, C.R. (1973). Estimation associated with linear models: a revisitation, *Communications in Statistics* **1**, 303–330.
- [10] Woodward, J.A., Bonett, D.G. & Brecht, M. (1990). *Introduction to Linear Models and Experimental Design*, Harcourt Brace Jovanovich, San Diego.

(See also **Regression Model Coding for the Analysis of Variance**)

ROGER E. KIRK AND B. NEBIYOU BEKELE

Analysis of Variance: Classification

ROGER E. KIRK

Volume 1, pp. 66–83

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Variance: Classification

A variety of **analysis of variance (ANOVA)** designs are available to researchers. Unfortunately, there is no standard nomenclature for the designs. For example, the simplest ANOVA design is variously referred to as a completely randomized design, one-way classification design, single-factor experiment, randomized group design, simple randomized design, single variable experiment, and one-way ANOVA. The use of multiple designations for designs is confusing. In this article, I describe a nomenclature and acronyms for ANOVA designs that are based on three simple ANOVA designs. These designs—completely randomized design, **randomized block design**, and **Latin square design**—are the building blocks with which more complex designs are constructed. Four characteristics of the designs are described: layout of the designs, partition of the total sum of squares and degrees of freedom, hypotheses that can be tested, and advantages and disadvantages of the designs. To simplify the presentation, I assume that all treatment levels of interest are included in the designs, that is, a fixed-effects model is appropriate. Space limitations prevent me from describing the computational procedures or the assumptions associated with the designs. For this information, the reader is referred to the many excellent experimental design books [1, 2, 4, 6, 7, 9–12].

Completely Randomized Design

The simplest ANOVA design in terms of assignment of participants (experimental units) to treatment levels and the statistical analysis of data is called a **completely randomized design**. The design is appropriate for experiments that have one treatment with $p \geq 2$ treatment levels. I use the terms ‘treatment,’ ‘factor,’ and ‘independent variable’ interchangeably. A treatment is identified by the capital letter A . A specific level of treatment A is denoted by the lower case letter a and a number or letter subscript, for example, $a_1, a_2, \dots, a_p$. A particular but unspecified treatment level is denoted by a_j where j ranges over the values $1, \dots, p$. In a completely randomized design, N participants are randomly assigned to the p levels

of the treatment. It is desirable to assign an equal or approximately equal number of participants, n_j , to each treatment level [10]. The abbreviated designation for a completely randomized design is CR- p , where CR stands for ‘completely randomized’ and p denotes the number of treatment levels.

Consider an experiment to evaluate three therapies for helping cigarette smokers break the habit. The designations for the three treatment levels are: a_1 = cognitive-behavioral therapy, a_2 = hypnosis, and a_3 = drug therapy. The dependent variable could be the number of participants who are no longer smoking after six months of treatment, the change in their cigarette consumption, or any one of a number of other measures.

Assume that $N = 45$ smokers who want to stop are available to participate in the experiment. The 45 smokers are randomly assigned to the three treatment levels with the restriction that $n = 15$ smokers are assigned to each level. The phrase ‘randomly assigned’ is important. Random assignment (see **Randomization**) helps to distribute the idiosyncratic characteristics of the participants over the three treatment levels so that the characteristics do not selectively bias the outcome of the experiment. If, for example, a disproportionately large number of very heavy smokers was assigned to one of the treatment levels, the comparison of this therapy with the other therapies could be compromised. The **internal validity** of a completely randomized design depends on random assignment.

The layout for the CR-3 design is shown in Figure 1. The partition of the total sum of squares, $SSTOTAL$, and total degrees of freedom, $np - 1$, are as follows:

$$\begin{aligned} SSTOTAL &= SSBG + SSWG \\ np - 1 &= p - 1 + p(n - 1), \end{aligned} \quad (1)$$

where $SSBG$ denotes the between-groups sum of squares and $SSWG$ denotes the within-groups sum of squares. The null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3$, where the μ_j ’s denote population means. The F statistic for testing the hypothesis is

$$F = \frac{SSBG/(p - 1)}{SSWG/[p(n - 1)]} = \frac{MSBG}{MSWG}. \quad (2)$$

The numerator of this statistic estimates error effects (error variation) plus any effects attributable to the treatment. The denominator provides an independent

		Treat. level	
Group ₁	Participant ₁	a_1	$\bar{Y}_{\cdot 1}$
	Participant ₂	a_1	
	$\vdots$	$\vdots$	
	Participant ₁₅	a_1	
Group ₂	Participant ₁₆	a_2	$\bar{Y}_{\cdot 2}$
	Participant ₁₇	a_2	
	$\vdots$	$\vdots$	
	Participant ₃₀	a_2	
Group ₃	Participant ₃₁	a_3	$\bar{Y}_{\cdot 3}$
	Participant ₃₂	a_3	
	$\vdots$	$\vdots$	
	Participant ₄₅	a_3	

Figure 1 Layout for a completely randomized design (CR-3 design) where 45 participants were randomly assigned to three levels of treatment A with the restriction that 15 participants were assigned to each level. The dependent-variable means for participants in treatment levels a_1 , a_2 , and a_3 are denoted by $\bar{Y}_{\cdot 1}$, $\bar{Y}_{\cdot 2}$, and $\bar{Y}_{\cdot 3}$, respectively

estimate of error effects. Hence, the F statistic can be thought of as the ratio of error and treatment effects:

$$F = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})}. \quad (3)$$

A large F statistic suggests that the dependent-variable population means or treatment effects are not all equal. An F statistic close to one suggests that there are no or negligible differences among the population means.

The advantages of a CR- p design are (a) simplicity in the randomization and statistical analysis and (b) flexibility with respect to having an equal or unequal number of participants in the treatment levels. A disadvantage is that differences among participants are controlled by random assignment. For this control to be effective, the participants should be relatively homogeneous or a relatively large number of participants should be used. The design described next, a **randomized block design**, enables a researcher to isolate and remove one source of variation among participants that ordinarily would be included in the estimate of the error effects. As a result, the randomized block design is usually more powerful than the completely randomized design.

Randomized Block Design

The *randomized block design*, denoted by RB- p , is appropriate for experiments that have one treatment with $p \geq 2$ treatment levels and n blocks. A block can contain a single participant who is observed under all p treatment levels or p participants who are similar with respect to a variable that is positively correlated with the dependent variable. If each block contains one participant, the order in which the treatment levels are administered is randomized independently for each block, assuming that the nature of the treatment permits this. If a block contains p matched participants, the participants in each block are randomly assigned to the treatment levels.

Consider the cigarette example again. It is reasonable to expect that the longer one has smoked, the more difficult it is to break the habit. Length of smoking is a **nuisance variable**, an undesired source of variation that can affect the dependent variable and contribute to the estimate of error effects. Suppose that the 45 smokers in the experiment are ranked in terms of the length of time they have smoked. The three smokers who have smoked for the shortest length of time are assigned to block 1, the next three smokers are assigned to block 2, and so on for the 15 blocks. Smokers in each block are similar with respect to length of time they have smoked. Differences among the 15 blocks reflect the effects of length of smoking. The layout for this design is shown in Figure 2. The total sum of squares and total degrees of freedom for the design are partitioned as follows:

$$SSTOTAL = SSA + SSBLOCKS$$

$$+ SSRESIDUAL$$

$$np - 1 = (p - 1) + (n - 1) + (n - 1)(p - 1), \quad (4)$$

where SSA denotes the treatment sum of squares and $SSBLOCKS$ denotes the block sum of squares. The $SSRESIDUAL$ is the interaction between the treatment and blocks; it is used to estimate error effects. Two null hypotheses can be tested.

$$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$$

(treatment A population means are equal)

$$H_0: \mu_{1\cdot} = \mu_{2\cdot} = \dots = \mu_{15\cdot}$$

(block population means are equal) (5)

	Treat. level	Treat. level	Treat. level	
Block ₁	a_1	a_2	a_3	$\bar{Y}_{1\cdot}$
Block ₂	a_1	a_2	a_3	$\bar{Y}_{2\cdot}$
Block ₃	a_1	a_2	a_3	$\bar{Y}_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Block ₁₅	a_1	a_2	a_3	$\bar{Y}_{15\cdot}$
	$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	$\bar{Y}_{\cdot 3}$	

Figure 2 Layout for a randomized block design (RB-3 design) where each block contains three matched participants who have smoked for about the same length of time. The participants in each block were randomly assigned to the three treatment levels. The mean cigarette consumption for participants in treatment levels a_1 , a_2 , and a_3 is denoted by $\bar{Y}_{\cdot 1}$, $\bar{Y}_{\cdot 2}$, and $\bar{Y}_{\cdot 3}$, respectively; the mean cigarette consumption for participants in Block₁, Block₂, ..., Block₁₅ is denoted by $\bar{Y}_{1\cdot}$, $\bar{Y}_{2\cdot}$, ..., $\bar{Y}_{15\cdot}$, respectively

The F statistics are

$$F = \frac{SSA/(p-1)}{SSRES/[(n-1)(p-1)]} = \frac{MSA}{MSRES}$$

and

$$F = \frac{SSBL/(n-1)}{SSRES/[(n-1)(p-1)]} = \frac{MSBL}{MSRES}. \quad (6)$$

The test of the block null hypothesis is of little interest because the blocks represent the nuisance variable of length of smoking.

The advantages of this design are (a) simplicity in the statistical analysis, and (b) the ability to isolate a nuisance variable so as to obtain greater power to reject a false null hypothesis. Disadvantages of the design include (a) the difficulty of forming homogeneous block or observing participants p times when p is large, and (b) the restrictive assumptions (sphericity and additive model) of the design. For a description of these assumptions, see Kirk [10, pp. 271–282].

Latin Square Design

The last building block design to be described is the Latin square design, denoted by LS- p . The design gets its name from an ancient puzzle that was concerned with the number of ways that Latin letters could be arranged in a square matrix so that each letter appeared once in each row and once

	c_1	c_2	c_3
b_1	a_1	a_2	a_3
b_2	a_2	a_3	a_1
b_3	a_3	a_1	a_2

Figure 3 Three-by-three Latin square where a_j denotes one of the $j = 1, \dots, p$ levels of treatment A , b_k denotes one of the $k = 1, \dots, p$ levels of nuisance variable B , and c_l denotes one of the $l = 1, \dots, p$ levels of nuisance variable C . Each level of treatment A appears once in each row and once in each column

in each column. A 3×3 Latin square is shown in Figure 3. The randomized block design enables a researcher to isolate one nuisance variable (variation among blocks) while evaluating the treatment effects. A Latin square design extends this procedure to two nuisance variables: variation associated with the rows of the square and variation associated with the columns of the square. As a result, the Latin square design is generally more powerful than the randomized block design.

In the cigarette smoking experiment, the nuisance variable of length of time that participants have smoked could be assigned to the rows of the square: b_1 = less than a year, b_2 = 1–3 years, and b_3 = more than three years. A second nuisance variable, number of cigarettes smoked per day, could be assigned to the columns of the square: c_1 = less than one pack, c_2 = 1–2 packs, and c_3 = more than two packs. The layout for this design is shown in Figure 4 and is based on the $a_j b_k c_l$ combinations in Figure 3. Five smokers who fit each of the $a_j b_k c_l$ combinations are randomly sampled from a large population of smokers who want to break the habit. The total sum of squares and total degrees of freedom are partitioned as follows:

$$SSTOTAL = SSA + SSB + SSC + SSRESIDUAL + SSWCELL$$

$$np^2 - 1 = (p-1) + (p-1) + (p-1) + (p-1)(p-2) + p^2(n-1), \quad (7)$$

where SSA denotes the treatment sum of squares, SSB denotes the row sum of squares, and SSC denotes the column sum of squares. $SSWCELL$ denotes the within cell sum of squares and estimates error effects. Three

		Variable comb.	
Group ₁	Participant ₁	$a_1b_1c_1$	$\bar{Y}_{\cdot 111}$
	$\vdots$	$\vdots$	
	Participant ₅	$a_1b_1c_1$	
Group ₂	Participant ₆	$a_1b_2c_3$	$\bar{Y}_{\cdot 123}$
	$\vdots$	$\vdots$	
	Participant ₁₀	$a_1b_2c_3$	
Group ₃	Participant ₁₁	$a_1b_3c_2$	$\bar{Y}_{\cdot 132}$
	$\vdots$	$\vdots$	
	Participant ₁₅	$a_1b_3c_2$	
Group ₄	Participant ₁₆	$a_2b_1c_2$	$\bar{Y}_{\cdot 212}$
	$\vdots$	$\vdots$	
	Participant ₂₀	$a_2b_1c_2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group ₉	Participant ₄₁	$a_3b_3c_1$	$\bar{Y}_{\cdot 331}$
	$\vdots$	$\vdots$	
	Participant ₄₅	$a_3b_3c_1$	

Figure 4 Layout for a Latin square design (LS-3 design) that is based on the Latin square in Figure 3. Treatment A represents three kinds of therapy, nuisance variable B represents length of time that a person has smoked, and nuisance variable C represents number of cigarettes smoked per day. Participants in Group₁, for example, received behavioral therapy (a_1), had smoked for less than one year (b_1), and smoked less than one pack of cigarettes per day (c_1). The mean cigarette consumption for the participants in the nine groups is denoted by $\bar{Y}_{\cdot 111}, \bar{Y}_{\cdot 123}, \dots, \bar{Y}_{\cdot 331}$

null hypotheses can be tested.

$$\begin{aligned}
 H_0: \mu_{1..} &= \mu_{2..} = \mu_{3..} \\
 &\text{(treatment } A \text{ population means are equal)} \\
 H_0: \mu_{\cdot 1.} &= \mu_{\cdot 2.} = \mu_{\cdot 3.} \\
 &\text{(row population means are equal)} \\
 H_0: \mu_{\cdot \cdot 1} &= \mu_{\cdot \cdot 2} = \mu_{\cdot \cdot 3} \\
 &\text{(columns population means are equal)} \quad (8)
 \end{aligned}$$

The F statistics are

$$\begin{aligned}
 F &= \frac{SSA/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSA}{MSWCELL} \\
 F &= \frac{SSB/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSB}{MSWCELL} \\
 F &= \frac{SSC/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSC}{MSWCELL}. \quad (9)
 \end{aligned}$$

The advantage of the Latin square design is the ability to isolate two nuisance variables to obtain greater power to reject a false null hypothesis. The disadvantages are (a) the number of treatment levels, rows, and columns must be equal, a balance that may be difficult to achieve; (b) if there are any interactions among the treatment levels, rows, and columns, the test of the treatment is positively biased; and (c) the randomization is relatively complex.

Three building block designs have been described that provide the organizational framework for the classification scheme and nomenclature in this article. The following ANOVA designs are extensions of or variations of one of the building block designs or a combination of two or more building block designs.

Generalized Randomized Block Design

A *generalized randomized block design*, denoted by GRB- p , is a variation of a randomized block design. Instead of having n blocks of homogeneous participants, the GRB- p design has w groups of homogeneous participants. The $z = 1, \dots, w$ groups, like the blocks in a randomized block design, represent a nuisance variable. The GRB- p design is appropriate for experiments that have one treatment with $p \geq 2$ treatment levels and w groups each containing np homogeneous participants. The total number of participants in the design is $N = npw$. The np participants in each group are randomly assigned to the p treatment levels with the restriction that n participants are assigned to each level. The layout for the design is shown in Figure 5.

In the smoking experiment, suppose that 30 smokers are available to participate. The 30 smokers are ranked with respect to the length of time that they have smoked. The $np = (2)(3) = 6$ smokers who have smoked for the shortest length of time are assigned to group 1, the next six smokers are assigned to group 2, and so on. The six smokers in each group are then randomly assigned to the three treatment levels with the restriction that $n = 2$ smokers are assigned to each level.

The total sum of squares and total degrees of freedom are partitioned as follows:

		Treat. level		Treat. level		Treat. level
Group ₁	Participant ₁	a_1	Participant ₃	a_2	Participant ₅	a_3
	Participant ₂	a_1	Participant ₄	a_2	Participant ₆	a_3
Group ₂	Participant ₇	a_1	Participant ₉	a_2	Participant ₁₁	a_3
	Participant ₈	a_1	Participant ₁₀	a_2	Participant ₁₂	a_3
Group ₃	Participant ₁₃	a_1	Participant ₁₅	a_2	Participant ₁₇	a_3
	Participant ₁₄	a_1	Participant ₁₆	a_2	Participant ₁₈	a_3
Group ₄	Participant ₁₉	a_1	Participant ₂₁	a_2	Participant ₂₃	a_3
	Participant ₂₀	a_1	Participant ₂₂	a_2	Participant ₂₄	a_3
Group ₅	Participant ₂₅	a_1	Participant ₂₇	a_2	Participant ₂₉	a_3
	Participant ₂₆	a_1	Participant ₂₈	a_2	Participant ₃₀	a_3

Figure 5 Generalized randomized block design (GRB-3 design) with $p = 3$ treatment levels and $w = 5$ groups of $np = (2)(3) = 6$ homogeneous participants. The six participants in each group were randomly assigned to the three treatment levels with the restriction that two participants were assigned to each level

$$\begin{aligned}
 SSTOTAL &= SSA + SSG + SSA \times G \\
 &\quad + SSWCELL \\
 npw - 1 &= (p - 1) + (w - 1) + (p - 1)(w - 1) \\
 &\quad + pw(n - 1), \quad (10)
 \end{aligned}$$

where SSG denotes the groups sum of squares and $SSA \times G$ denotes the interaction of treatment A and groups. The within cells sum of squares, $SSWCELL$, is used to estimate error effects. Three null hypotheses can be tested.

$$\begin{aligned}
 H_0: \mu_1. &= \mu_2. = \mu_3. \\
 &\text{(treatment } A \text{ population means are equal)} \\
 H_0: \mu_{.1} &= \mu_{.2} = \dots = \mu_{.5} \\
 &\text{(group population means are equal)} \\
 H_0: \mu_{jz} - \mu_{j'z} - \mu_{jz'} + \mu_{j'z'} &= 0 \text{ for all } j, j', z, \\
 &\text{and } z' \text{ (treatment } A \text{ and groups do not interact)} \quad (11)
 \end{aligned}$$

The F statistics are

$$\begin{aligned}
 F &= \frac{SSA/(p - 1)}{SSWCELL/[pw(n - 1)]} = \frac{MSA}{MSWCELL} \\
 F &= \frac{SSG/(w - 1)}{SSWCELL/[pw(n - 1)]} = \frac{MSG}{MSWCELL}
 \end{aligned}$$

$$F = \frac{SSA \times G/(p - 1)(w - 1)}{SSWCELL/[pw(n - 1)]} = \frac{MSA \times G}{MSWCELL}. \quad (12)$$

The generalized randomized block design enables a researcher to isolate a nuisance variable—an advantage that it shares with the randomized block design. Furthermore, for the fixed-effects model, the design uses the pooled variation in the pw cells to estimate error effects rather than an interaction as in the randomized block design. Hence, the restrictive sphericity assumption of the randomized block design is replaced with the assumption of homogeneity of within cell population variances.

Graeco-Latin Square and Hyper-Graeco-Latin Square Designs

A *Graeco-Latin square design*, denoted by GLS- p , is constructed by superimposing one Latin square on a second Latin square that is orthogonal to the first. Two Latin squares are orthogonal if when they are superimposed, every treatment level of one square occurs once with every treatment level of the other square. An example of a 3×3 Graeco-Latin square is shown in Figure 6. The layout for a GLS-3 design with $n = 5$ participants in each cell is shown in Figure 7. The design is based on

	c_1	c_2	c_3
b_1	a_1d_1	a_2d_2	a_3d_3
b_2	a_2d_3	a_3d_1	a_1d_2
b_3	a_3d_2	a_1d_3	a_2d_1

Figure 6 Three-by-three Graeco-Latin square, where a_j denotes one of the $j = 1, \dots, p$ levels of treatment A , b_k denotes one of the $k = 1, \dots, p$ levels of nuisance variable B , c_l denotes one of the $l = 1, \dots, p$ levels of nuisance variable C , and d_m denotes one of the $m = 1, \dots, p$ levels of nuisance variable D . Each level of treatment A and nuisance variable D appear once in each row and once in each column and each level of A occurs once with each level of D

		Treat. comb.	
Group ₁	Participant ₁	$a_1b_1c_1d_1$	$\bar{Y}_{\cdot 1111}$
	$\vdots$	$\vdots$	
	Participant ₅	$a_1b_1c_1d_1$	
Group ₂	Participant ₆	$a_1b_2c_3d_2$	$\bar{Y}_{\cdot 1232}$
	$\vdots$	$\vdots$	
	Participant ₁₀	$a_1b_2c_3d_2$	
Group ₃	Participant ₁₁	$a_1b_3c_2d_3$	$\bar{Y}_{\cdot 1323}$
	$\vdots$	$\vdots$	
	Participant ₁₅	$a_1b_3c_2d_3$	
Group ₄	Participant ₁₆	$a_2b_1c_2d_2$	$\bar{Y}_{\cdot 2122}$
	$\vdots$	$\vdots$	
	Participant ₂₀	$a_2b_1c_2d_2$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group ₉	Participant ₄₁	$a_3b_3c_1d_2$	$\bar{Y}_{\cdot 3312}$
	$\vdots$	$\vdots$	
	Participant ₄₅	$a_3b_3c_1d_2$	

Figure 7 Layout for a 3×3 Graeco-Latin square design (GLS-3 design) that is based on the Graeco-Latin square in Figure 6

the $a_jb_kc_ld_m$ combinations in Figure 6. A Graeco-Latin square design enables a researcher to isolate three nuisance variables. However, the design is rarely used in the behavioral and social sciences for several reasons. First, the design is restricted to research problems in which three nuisance variables and a treatment each have the same number of levels. It is difficult to achieve this balance for four variables. Second, if interactions among the variables occur, one or more tests will be positively biased.

The prevalence of interactions in behavioral research limits the usefulness of the design.

A *hyper-Graeco-Latin square design*, denoted by HGLS- p , is constructed like a Graeco-Latin square design, but combines more than two orthogonal Latin squares. It shares the advantages and disadvantages of the Latin square and Graeco-Latin square designs.

Cross-over Design

A **cross-over design**, denoted by CO- p , is so called because participants are administered first one treatment level and then ‘crossed over’ to receive a second, and perhaps a third or even a fourth treatment level. The crossover design is appropriate for experiments having one treatment with $p \geq 2$ levels and two nuisance variables. One of the nuisance variables is blocks of participants (experimental units); the other nuisance variable is periods of time and must have p levels. Each treatment level must occur an equal number of times in each time period. Hence, the design requires $n = hp$ blocks, where h is a positive integer.

The simplest cross-over design has two treatment levels, a_1 and a_2 . Half of the participants receive a_1 followed by a_2 ; the other half receive a_2 followed by a_1 . The design can be used when it is reasonable to assume that participants revert to their original state before the second treatment level is administered. For example, the effects of a drug should be eliminated before a second drug is administered. **Carry-over effects**—treatment effects that continue after a treatment has been discontinued—are a potential threat to the **internal validity** of the design. Sometimes carry-over effects can be eliminated or at least minimized by inserting a rest or washout period between administrations of the treatment levels. Alternatively, a complex cross-over design can be used that provides a statistical adjustment for the carry-over effects of the immediately preceding treatment level. These designs are discussed by Cochran and Cox [3], Federer [5], and Jones and Kenward [8].

Cross-over designs have features of randomized block and Latin square designs. For example, each participant receives all p treatment levels and serves as his or her own control, as in a randomized block design with repeated measures (see **Repeated Measures Analysis of Variance**).

And, as in a Latin square design, each treatment level occurs an equal number of times in each time period, and the effects of two nuisance variables—blocks and time periods—can be isolated. Cross-over designs are often used in clinical trials, agricultural research, and marketing research. The layout for a CO-2 design with eight blocks is shown in Figure 8.

The total sum of squares and total degrees of freedom for the design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSTP + SSBLOCKS \\ &\quad + SSRESIDUAL \\ np - 1 &= (p - 1) + (p - 1) + (n - 1) \\ &\quad + (n - 2)(p - 1), \end{aligned} \quad (13)$$

where $SSTP$ denotes the time-periods sum of squares. Three null hypotheses can be tested.

$$\begin{aligned} H_0: \mu_{.1.} &= \mu_{.2.} \\ &\text{(treatment A population means are equal)} \\ H_0: \mu_{..1} &= \mu_{..2} \\ &\text{(time-periods population means are equal)} \\ H_0: \mu_{1..} &= \mu_{2..} = \cdots = \mu_{8..} \\ &\text{(block population means are equal)} \end{aligned} \quad (14)$$

	Time period b_1	Time period b_2
	Treat. level	Treat. level
Block ₁	a_1	a_2
Block ₂	a_2	a_1
Block ₃	a_2	a_1
Block ₄	a_1	a_2
Block ₅	a_2	a_1
Block ₆	a_1	a_2
Block ₇	a_2	a_1
Block ₈	a_1	a_2

Figure 8 Layout for a cross-over design with two treatment levels and two time periods (CO-2 design). Each block contains $p = 2$ treatment levels, and each treatment level occurs an equal number of times in each time period. The participants were randomly assigned to the blocks

The F statistics are

$$\begin{aligned} F &= \frac{SSA/(p - 1)}{SSRESIDUAL/[(n - 2)(p - 1)]} \\ &= \frac{MSA}{MSRESIDUAL} \\ F &= \frac{SSTP/(p - 1)}{SSRESIDUAL/[(n - 2)(p - 1)]} \\ &= \frac{MSTP}{MSRESIDUAL} \\ F &= \frac{SSBLOCKS/(n - 1)}{SSRESIDUAL/[(n - 2)(p - 1)]} \\ &= \frac{MSBLOCKS}{MSRESIDUAL}. \end{aligned} \quad (15)$$

The design shares the advantages and limitations of the randomized block and Latin square designs. In particular, it must be reasonable to assume that there are no interactions among the blocks, time periods, and treatment. If this assumption is not satisfied, a test of one or more of the corresponding effects is biased. Finally, statistical adjustments are required if carry-over effects of the immediately preceding treatment level are present.

Incomplete Block Designs

The name *incomplete block design* refers to a large class of designs in which p treatment levels of a single treatment are assigned to blocks of size k , where $k < p$. In many respects, the designs are similar to randomized block designs. However, incomplete block designs are typically used when a researcher wants to evaluate a large number of treatment levels. In such situations, the use of a randomized block design, an alternative design choice, may not be feasible because of the difficulty in observing participants over and over again or the difficulty in forming sufficiently homogeneous blocks.

The layout for a **balanced incomplete block design** with seven treatment levels, denoted by BIB-7, is shown in Figure 9. This design enables a researcher to evaluate seven treatment levels using blocks of size three. Each pair of treatment levels occurs in some block an equal number of times. Balanced incomplete block designs with many treatment levels, say $p > 10$, may require a prohibitively large number of blocks. For example, a BIB-10 design with

	Treat. level	Treat. level	Treat. level
Block ₁	a_1	a_2	a_4
Block ₂	a_2	a_3	a_5
Block ₃	a_3	a_4	a_6
Block ₄	a_4	a_5	a_7
Block ₅	a_5	a_6	a_1
Block ₆	a_6	a_7	a_2
Block ₇	a_7	a_1	a_3

Figure 9 Layout for a balanced incomplete block design with seven treatment levels in blocks of size three (BIB-7 design)

blocks of size three requires 30 blocks. A smaller number of blocks can be used if the design is partially balance, that is, each pair of treatment levels does not occur within some block an equal number of times. Such designs are called *partially balanced incomplete block designs* and are denoted by PBIB- p .

Another large group of incomplete block designs are *lattice designs*. Unlike the incomplete block designs just described, the layout and analysis of lattice designs are facilitated by establishing a correspondence between the p treatment levels of the design and the treatment combinations of a factorial design. Factorial designs are described later. Lattice designs can be balanced, denoted by LBIB- p , partially balanced, denoted by LPBIB- p , and unbalanced, denoted by LUBIB- p .

Yet another group of incomplete block designs is based on an incomplete Latin square. One example is the *Youden square design*, denoted by YBIB- p (see **Balanced Incomplete Block Designs**). The design combines features of balanced incomplete block designs and Latin square designs. A Youden square design is constructed by omitting one or more columns of a Latin square. Hence the design is not really a square. Consider the layout for a YBIB-4 design shown in Figure 10. A 4×4 Latin square is shown below the YBIB-4 design. An examination of the YBIB-4 design reveals that it contains the treatment levels in columns 1–3 of the Latin square.

Incomplete block designs are rarely used in the behavioral and social sciences for several reasons. First, the designs are most useful when the number of treatment levels, p , is very large. Researchers in the behavioral and social sciences seldom design experiments with $p > 7$. Second, researchers are often interested in simultaneously evaluating two or more

	Column level c_1	Column level c_2	Column level c_3
Block ₁	a_1	a_2	a_3
Block ₂	a_2	a_3	a_4
Block ₃	a_3	a_4	a_1
Block ₄	a_4	a_1	a_2

	c_1	c_2	c_3	c_4
b_1	a_1	a_2	a_3	a_4
b_2	a_2	a_3	a_4	a_1
b_3	a_3	a_4	a_1	a_2
b_4	a_4	a_1	a_2	a_3

Figure 10 The top figure shows the layout for a Youden square design with four treatment levels in blocks of size three (YBIB-4 design). The design contains the treatment levels in columns c_1 , c_2 , and c_3 of the Latin square shown in the lower figure

treatments and associated interactions. All of the designs described so far are appropriate for experiments with one treatment. The designs described next can be used when a researcher is interested in simultaneously evaluating several treatments and associated interactions.

Completely Randomized Factorial Design

Factorial designs differ from those described previously in that two or more treatments can be evaluated simultaneously and, in addition, the **interaction** between the treatments can be evaluated. Each level of one treatment must occur once with each level of other treatments and vice versa, that is, the treatments must be crossed. Although there are many kinds of factorial design, they are all constructed by combining two or more building block designs.

The simplest factorial design from the standpoint of randomization procedures and data analysis is the *completely randomized factorial design*. The design is denoted by CRF- pq , where CR indicates the building block design and F indicates that it is a factorial design. The design is constructed by combining the levels of a CR- p design for treatment A with those of a second CR- q design for treatment B so that each level of the CR- p design appears once with each level of the CR- q design and vice versa. The levels of the two treatments, A and B , are said to be completely crossed.

Hence, the design has $p \times q$ treatment combinations, $a_1b_1, a_1b_2, \dots, a_pb_q$. The layout for a CRF-23 design with $p = 2$ levels of treatment A and $q = 3$ levels of treatment B is shown in Figure 11. In this example, 30 participants are randomly assigned to the $2 \times 3 = 6$ treatment combinations with the restriction that $n = 5$ participants are assigned to each combination.

The total sum of squares and total degrees of freedom for the design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSB + SSA \times B \\ &\quad + SSWCELL \\ npq - 1 &= (p - 1) + (q - 1) + (p - 1)(q - 1) \\ &\quad + pq(n - 1), \end{aligned} \quad (16)$$

where $SSA \times B$ denotes the interaction of treatments A and B . Three null hypotheses can be tested.

$$H_0: \mu_{1\cdot} = \mu_{2\cdot}.$$

(treatment A population means are equal)

$$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$$

(treatment B population means are equal)

$$H_0: \mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k'} = 0 \text{ for all } j, j', k,$$

and k' (treatments A and B do not interact)

(17)

The F statistics are

$$\begin{aligned} F &= \frac{SSA/(p - 1)}{SSWCELL/[pq(n - 1)]} = \frac{MSA}{MSWCELL} \\ F &= \frac{SSB/(q - 1)}{SSWCELL/[pq(n - 1)]} = \frac{MSB}{MSWCELL} \\ F &= \frac{SSA \times B/(p - 1)(q - 1)}{SSWCELL/[pq(n - 1)]} = \frac{MSA \times B}{MSWCELL}. \end{aligned} \quad (18)$$

The advantages of the design are as follows: (a) All participants are used in simultaneously evaluating the effects of two or more treatments. The effects of each treatment are evaluated with the same precision as if the entire experiment had been devoted to that treatment alone. Thus, the design permits efficient use of resources. (b) The interactions among the treatments can be evaluated. The disadvantages of the design are as follows. (a) If numerous treatments are included in the experiment, the number of participants required can be prohibitive. (b) A factorial design lacks simplicity in the interpretation of results if interaction effects are present. Unfortunately, interactions among variables in the behavioral sciences and education are common. (c) The use of a factorial design commits a researcher to a relatively large experiment. Small exploratory experiments may indicate much more promising lines of investigation than those originally envisioned. Relatively small experiments permit greater freedom in the pursuit of serendipity.

Randomized Block Factorial Design

A two-treatment *randomized block factorial design* is denoted by RBF- pq . The design is constructed by combining the levels of an RB- p design with those of an RB- q design so that each level of the RB- p design appears once with each level of the RB- q design and vice versa. The design uses the blocking technique described in connection with a randomized block design to isolate variation attributable to a nuisance variable while simultaneously evaluating two or more treatments and associated interactions.

		Treat. comb.	
Group ₁	Participant ₁	a_1b_1	$\bar{Y}_{\cdot 11}$
	$\vdots$	$\vdots$	
	Participant ₅	a_1b_1	
Group ₂	Participant ₆	a_1b_2	$\bar{Y}_{\cdot 12}$
	$\vdots$	$\vdots$	
	Participant ₁₀	a_1b_2	
Group ₃	Participant ₁₁	a_1b_3	$\bar{Y}_{\cdot 13}$
	$\vdots$	$\vdots$	
	Participant ₁₅	a_1b_3	
Group ₄	Participant ₁₆	a_2b_1	$\bar{Y}_{\cdot 21}$
	$\vdots$	$\vdots$	
	Participant ₂₀	a_2b_1	
Group ₅	Participant ₂₁	a_2b_2	$\bar{Y}_{\cdot 22}$
	$\vdots$	$\vdots$	
	Participant ₂₅	a_2b_2	
Group ₆	Participant ₂₆	a_2b_3	$\bar{Y}_{\cdot 23}$
	$\vdots$	$\vdots$	
	Participant ₃₀	a_2b_3	

Figure 11 Layout for a two-treatment, completely randomized factorial design (CRF-23 design) where 30 participants were randomly assigned to six combinations of treatments A and B with the restriction that five participants were assigned to each combination

	Treat. comb.	Treat. comb.	Treat. comb.	Treat. comb.	Treat. comb.	Treat. comb.	
Block ₁	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2	$\bar{Y}_{1..}$
Block ₂	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2	$\bar{Y}_{2..}$
Block ₃	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2	$\bar{Y}_{3..}$
Block ₄	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2	$\bar{Y}_{4..}$
Block ₅	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2	$\bar{Y}_{5..}$
	$\bar{Y}_{.11}$	$\bar{Y}_{.12}$	$\bar{Y}_{.21}$	$\bar{Y}_{.22}$	$\bar{Y}_{.31}$	$\bar{Y}_{.32}$	

Figure 12 Layout for a two-treatment, randomized block factorial design (RBF-32 design)

An RBF-32 design has blocks of size $3 \times 2 = 6$. If a block consists of matched participants, n blocks of six matched participants must be formed. The participants in each block are randomly assigned to the $a_1b_1, a_1b_2, \dots, a_pb_q$ treatment combinations. Alternatively, if repeated measures are obtained, each participant must be observed six times. For this case, the order in which the treatment combinations are administered is randomized independently for each block, assuming that the nature of the treatments permits this. The layout for the design is shown in Figure 12.

The total sum of squares and total degrees of freedom for an RBF-32 design are partitioned as follows:

$$\begin{aligned}
 SSTOTAL &= SSA + SSB + SSA \times B \\
 &\quad + SSRESIDUAL \\
 npq - 1 &= (p - 1) + (q - 1) + (p - 1)(q - 1) \\
 &\quad + (n - 1)(pq - 1). \quad (19)
 \end{aligned}$$

Three null hypotheses can be tested.

$$\begin{aligned}
 H_0: \mu_{1.} &= \mu_{2.} = \mu_{3.} \\
 &\quad \text{(treatment } A \text{ population means are equal)} \\
 H_0: \mu_{.1} &= \mu_{.2} \\
 &\quad \text{(treatment } B \text{ population means are equal)} \\
 H_0: \mu_{jk} - \mu_{j'k} - \mu_{jk'} + \mu_{j'k'} &= 0 \text{ for all } j, j', k, \\
 &\quad \text{and } k' \text{ (treatments } A \text{ and } B \text{ do not interact)} \quad (20)
 \end{aligned}$$

The F statistics are

$$\begin{aligned}
 F &= \frac{SSA/(p - 1)}{SSRESIDUAL/[(n - 1)(pq - 1)]} \\
 &= \frac{MSA}{MSRESIDUAL}
 \end{aligned}$$

$$\begin{aligned}
 F &= \frac{SSB/(q - 1)}{SSRESIDUAL/[(n - 1)(pq - 1)]} \\
 &= \frac{MSB}{MSRESIDUAL} \\
 F &= \frac{SSA \times B/(p - 1)(q - 1)}{SSRESIDUAL/[(n - 1)(pq - 1)]} \\
 &= \frac{MSA \times B}{MSRESIDUAL}. \quad (21)
 \end{aligned}$$

The design shares the advantages and disadvantages of the randomized block design. It has an additional disadvantage: if treatment A or B has numerous levels, say four or five, the block size becomes prohibitively large. Designs that reduce the block size are described next.

Split-plot Factorial Design

The *split-plot factorial* design is appropriate for experiments with two or more treatments where the number of treatment combinations exceeds the desired block size. The term *split-plot* comes from agricultural experimentation where the levels of, say, treatment A are applied to relatively large plots of land—the whole plots. The whole plots are then split or subdivided and the levels of treatment B are applied to the subplots within each whole plot.

A two-treatment, split-plot factorial design is constructed by combining a CR- p design with an RB- q design. In the split-plot factorial design, the assignment of participants to treatment combinations is carried out in two stages. To illustrate, again consider the smoking example. Suppose that we are interested in comparing the three therapies and also in comparing two lengths of the therapy, $b_1 =$ three months and $b_2 =$ six months. If 60 smokers are available, they can be ranked in terms of the length of time that they

have smoked. The two participants who have smoked for the shortest time are assigned to one block, the next two smokers to another block, and so on. This procedure produces 30 blocks in which the two smokers in a block are similar in terms of the length of time they have smoked. In the first stage of randomization, the 30 blocks are randomly assigned to the three levels of treatment A with the restriction that 10 blocks are assigned to each level of treatment A . In the second stage of randomization, the two smokers in each block are randomly assigned to the two levels of treatment B with the restriction that b_1 and b_2 appear equally often in each level of treatment A . An exception to this randomization procedure must be made when treatment B is a temporal variable, such as successive learning trials or periods of time. Trial two, for example, cannot occur before trial one.

The layout for this split-plot factorial design with three levels of treatment A and two levels of treatment B is shown in Figure 13. The total sum of squares and total degrees of freedom are partitioned as follows:

$$\begin{aligned}
 SSTOTAL &= SSA + SSBL(A) + SSB + SSA \times B \\
 &\quad + SSRESIDUAL \\
 npq - 1 &= (p - 1) + p(n - 1) + (q - 1) \\
 &\quad + (p - 1)(q - 1) + p(n - 1)(q - 1),
 \end{aligned} \tag{22}$$

where $SSBL(A)$ denotes the sum of squares of blocks that are nested in the p levels of treatment

		Treat. comb. b_1	Treat. comb. b_2	
a_1	Group ₁	Block ₁	a_1b_1	$\bar{Y}_{\cdot 1}$
		$\vdots$	$\vdots$	
		Block _{n}	a_1b_1	
a_2	Group ₂	Block _{$n+1$}	a_2b_1	$\bar{Y}_{\cdot 2}$
		$\vdots$	$\vdots$	
		Block _{$2n$}	a_2b_1	
a_3	Group ₃	Block _{$2n+1$}	a_3b_1	$\bar{Y}_{\cdot 3}$
		$\vdots$	$\vdots$	
		Block _{$3n$}	a_3b_1	
		$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	

Figure 13 Layout for a two-treatment, split-plot factorial design (SPF-3·2 design). Treatment A is confounded with groups

A . Three null hypotheses can be tested.

$$H_0: \mu_{\cdot 1 \cdot} = \mu_{\cdot 2 \cdot} = \mu_{\cdot 3 \cdot}$$

(treatment A population means are equal)

$$H_0: \mu_{\cdot \cdot 1} = \mu_{\cdot \cdot 2}$$

(treatment B population means are equal)

$$H_0: \mu_{\cdot jk} - \mu_{\cdot j'k} - \mu_{\cdot jk'} + \mu_{\cdot j'k'} = 0 \text{ for all } j, j', k, \text{ and } k' \text{ (treatments } A \text{ and } B \text{ do not interact) (23)}$$

The F statistics are

$$\begin{aligned}
 F &= \frac{SSA/(p - 1)}{SSBL(A)/[p(n - 1)]} \\
 &= \frac{MSA}{MSBL(A)} \\
 F &= \frac{SSB/(q - 1)}{SSRESIDUAL/[p(n - 1)(q - 1)]} \\
 &= \frac{MSB}{MSRESIDUAL} \\
 F &= \frac{SSA \times B/(p - 1)(q - 1)}{SSRESIDUAL/[p(n - 1)(q - 1)]} \\
 &= \frac{MSA \times B}{MSRESIDUAL}.
 \end{aligned} \tag{24}$$

Treatment A is called a *between-blocks effect*. The error term for testing between-blocks effects is $MSBL(A)$. Treatment B and the $A \times B$ interaction are *within-blocks effects*. The error term for testing the within-blocks effects is $MSRESIDUAL$. The designation for a two-treatment, split-plot factorial design is SPF- $p \cdot q$. The p preceding the dot denotes the number of levels of the between-blocks treatment; the q after the dot denotes the number of levels of the within-blocks treatment. Hence, the design in Figure 13 is an SPF-3·2 design. A careful examination of the randomization and layout of the between-blocks effects reveals that they resemble those for a CR-3 design. The randomization and layout of the within-blocks effects at each level of treatment A resemble those for an RB-2 design.

The block size of the SPF-3·2 design in Figure 13 is three. The RBF-32 design in Figure 12 contains the same $3 \times 2 = 6$ treatment combinations, but the block size is six. The advantage of the split-plot factorial – the smaller block size – is achieved by confounding groups of blocks with treatment A . Consider the sample means $\bar{Y}_{\cdot 1}$, $\bar{Y}_{\cdot 2}$, and $\bar{Y}_{\cdot 3}$ in

Figure 13. The differences among the means reflect the differences among the three groups of smokers as well as the differences among the three levels of treatment A. To put it another way, we cannot tell how much of the differences among the three sample means is attributable to the differences among Group₁, Group₂, and Group₃, and how much is attributable to the differences among treatments levels a_1 , a_2 , and a_3 . For this reason, the groups and treatment A are said to be *completely confounded* (see **Confounding Variable**).

The use of confounding to reduce the block size in an SPF- $p \cdot q$ design involves a trade-off that needs to be made explicit. The RBF-32 design uses the same error term, *MSRESIDUAL*, to test hypotheses for treatments A and B and the $A \times B$ interaction. The two-treatment, split-plot factorial design, however, uses two error terms: *MSBL(A)* is used to test treatment A; a different and usually much smaller error term, *MSRESIDUAL*, is used to test treatment B and the $A \times B$ interaction. As a result, the **power** of the tests of treatment B and the $A \times B$ interaction is greater than that for treatment A. Hence, a split-plot factorial design is a good design choice if a researcher is more interested in treatment B and the $A \times B$ interaction than in treatment A. When both treatments and the $A \times B$ interaction are of equal interest, a randomized block factorial design is a better choice if the larger block size is acceptable. If a large block size is not acceptable and the researcher is primarily interested in treatments A and B, an alternative design choice is the confounded factorial design. This design, which is described next, achieves a reduction in block size by confounding groups of blocks with the $A \times B$ interaction. As a result, tests of treatments A and B are more powerful than the test of the $A \times B$ interaction.

Confounded Factorial Designs

Confounded factorial designs are constructed from either randomized block designs or Latin square designs. A simple confounded factorial design is denoted by RBCF- p^k . The RB in the designation indicates the building block design, C indicates that an interaction is completely confounded with groups, k indicates the number of treatments, and p indicates the number of levels of each treatment. The layout for an RBCF-2² design is shown in Figure 14. The

		Treat. comb.	Treat. comb.	
$(AB)_{jk}$ Group ₁	Block ₁	a_1b_1	a_2b_2	$\bar{Y}_{\cdot 111} + \bar{Y}_{\cdot 221}$
	$\vdots$	$\vdots$	$\vdots$	
	Block _{n}	a_1b_1	a_2b_2	
$(AB)_{jk}$ Group ₂	Block _{$n+1$}	a_1b_2	a_2b_1	$\bar{Y}_{\cdot 122} + \bar{Y}_{\cdot 212}$
	$\vdots$	$\vdots$	$\vdots$	
	Block _{$2n$}	a_1b_2	a_2b_1	

Figure 14 Layout for a two-treatment, randomized block confounded factorial design (RBCF-2² design). The $A \times B$ interaction is confounded with groups

total sum of squares and total degrees of freedom are partitioned as follows:

$$\begin{aligned}
 SSTOTAL &= SSA \times B + SSBL(G) + SSA + SSB \\
 &\quad + SSRESIDUAL \\
 nvw - 1 &= (w - 1) + w(n - 1) + (p - 1) + (q - 1) \\
 &\quad + w(n - 1)(v - 1), \quad (25)
 \end{aligned}$$

where *SSBL(G)* denotes the sum of squares of blocks that are nested in the w groups and v denotes the number of combinations of treatments A and B in each block. Three null hypotheses can be tested.

$$H_0: \mu_{\cdot jk} - \mu_{\cdot j'k} - \mu_{\cdot jk'} + \mu_{\cdot j'k'} = 0 \text{ for all } j,$$

$$j', k, \text{ and } k' \text{ (treatments A and B do not interact)}$$

$$H_0: \mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot}$$

$$\text{(treatment A population means are equal)}$$

$$H_0: \mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot}$$

$$\text{(treatment B population means are equal)} \quad (26)$$

The F statistics are

$$\begin{aligned}
 F &= \frac{SSA \times B / (w - 1)}{SSBL(G) / [w(n - 1)]} \\
 &= \frac{MSA \times B}{MSBL(G)} \\
 F &= \frac{SSA / (p - 1)}{SSRESIDUAL / [w(n - 1)(v - 1)]} \\
 &= \frac{MSA}{MSRESIDUAL} \\
 F &= \frac{SSB / (q - 1)}{SSRESIDUAL / [w(n - 1)(v - 1)]}
 \end{aligned}$$

$$= \frac{MSB}{MSRESIDUAL}. \quad (27)$$

In this example, groups are confounded with the $A \times B$ interaction. Hence, a test of the interaction is also a test of differences among the groups and vice versa.

A randomized block factorial design with two levels of treatments A and B has blocks of size four. The RBCF- 2^2 design confounds the $A \times B$ interaction with groups and thereby reduces the block size to two. The power of the tests of the two treatments is usually much greater than the power of the test of the $A \times B$ interaction. Hence, the RBCF- 2^2 design is a good design choice if a small block size is required and the researcher is primarily interested in the tests of the treatments.

For experiments with three treatments, A , B , and C , each having p levels, a researcher can confound one interaction, say $A \times B \times C$, with groups and not confound the treatments or two-treatment interactions. The design is called a *partially confounded factorial design* and is denoted by RBPF- p^3 , where P indicates partial confounding. A two-treatment, confounded factorial design that is based on a Latin square is denoted by LSCF- p^2 .

Two kinds of confounding have been described: group-treatment confounding in an SPF- $p \cdot q$ design and group-interaction confounding in RBCF- p^k and LSCF- p^k designs. A third kind of confounding, treatment-interaction confounding, is described next.

Fractional Factorial Design

Confounded factorial designs reduce the number of treatment combinations that appear in a block. Fractional factorial designs use treatment-interaction confounding to reduce the number of treatment combinations that appear in an experiment. For example, the number of treatment combinations that must be included in a multitreatment experiment can be reduced to some fraction—1/2, 1/3, 1/4, 1/8, 1/9, and so on—of the total number of treatment combinations in an unconfounded factorial design.

Fractional factorial designs are constructed using completely randomized, randomized block, and Latin square building block designs. The resulting designs are denoted by CRFF- $p^k - i$, RBFF- $p^k - i$, and LSFF- p^k , respectively, where CR, RB, and LS denote the building block design. The letters FF in the

acronyms indicate a fractional factorial design, k indicates the number of treatments, p indicates the number of levels of each treatment, and i indicates the fraction of the treatment combinations in the design. For example, if $k = 2$ and $i = 1$, the design contains 1/2 of the combinations of a complete factorial design; if $i = 2$, the design contains 1/4 of the combinations.

To conserve space, I show the layout in Figure 15 for a very small CRFF- 2^{3-1} design with three treatments. Ordinarily, a fractional factorial design would have many more treatments. The total sum of squares and degrees of freedom are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA[B \times C] + SSB[A \times C] \\ &\quad + SSC[A \times B] + SSWCELL \\ npq - 1 &= (p - 1) + (q - 1) + (r - 1) \\ &\quad + pq(n - 1), \end{aligned} \quad (28)$$

where $SSA[B \times C]$, for example, denotes the sum of squares for treatment A that is indistinguishable from the $B \times C$ interaction. Treatment A and the $B \times C$ interaction are two labels for the same source of variation—they are *aliases*. Notice that the total sum of squares does not include the $A \times B \times C$ interaction. Three null hypotheses can be tested.

$$H_0: \mu_{1..} = \mu_{2..}$$

		Treat. comb.	
Group ₁	Participant ₁	$a_1b_1c_1$	$\bar{Y}_{.111}$
	Participant _{n}	$a_1b_1c_1$	
Group ₂	Participant _{$n+1$}	$a_1b_2c_2$	$\bar{Y}_{.122}$
	Participant _{$2n$}	$a_1b_2c_2$	
Group ₃	Participant _{$2n+1$}	$a_2b_1c_2$	$\bar{Y}_{.212}$
	Participant _{$3n$}	$a_2b_1c_2$	
Group ₄	Participant _{$3n+1$}	$a_2b_2c_1$	$\bar{Y}_{.221}$
	Participant _{$4n$}	$a_2b_2c_1$	

Figure 15 Layout for a three-treatment, fractional factorial design (CRFF- 2^{3-1}). A three-treatment, completely randomized factorial design with two levels of each treatment (CRF- 2^3 design) would have $2 \times 2 \times 2 = 8$ treatment combinations. The fractional factorial design contains only $1/2(8) = 4$ of the combinations

(treatment A population means are equal
or the $B \times C$ interaction is zero)

$$H_0: \mu_{.1.} = \mu_{.2.}$$

(treatment B population means are equal
or the $A \times C$ interaction is zero)

$$H_0: \mu_{..1} = \mu_{..2}$$

(treatment C population means are equal
or the $A \times B$ interaction is zero) (29)

The F statistics are

$$\begin{aligned} F &= \frac{SSA/(p-1)}{SSWCELL/pq(n-1)} = \frac{MSA}{MSWCELL} \\ F &= \frac{SSB/(q-1)}{SSWCELL/pq(n-1)} = \frac{MSB}{MSWCELL} \\ F &= \frac{SSC/(r-1)}{SSWCELL/pq(n-1)} = \frac{MSC}{MSWCELL}. \end{aligned} \quad (30)$$

In this example, treatments are aliased with interactions. Hence, if $F = MSA/MSWCELL$ is significant, a researcher does not know whether it is because treatment A is significant, or because the $B \times C$ interaction is significant, or both.

You may wonder why anyone would use such a design—after all, experiments are supposed to help us resolve ambiguity not create it. Fractional factorial designs are typically used in exploratory research situations where a researcher is interested in six or more treatments and can perform follow-up experiments if necessary. Suppose that a researcher wants to perform an experiment with six treatments each having two levels. A CRF-222222 design would have $2 \times 2 \times 2 \times 2 \times 2 \times 2 = 64$ treatment combinations. If two participants are randomly assigned to each combination, a total of $2 \times 64 = 128$ participants would be required. By using a one-fourth fractional factorial design, CRFF- 2^{6-2} design, the researcher can reduce the number of treatment combinations in the experiment from 64 to 16 and the number of participants from 128 to 32. If none of the F statistics are significant, the researcher has answered the research questions with one-fourth of the effort. If, however, some of the F statistics are significant, the researcher can perform several small follow-up experiments to determine what is significant.

In summary, the main advantage of a fractional factorial design is that it enables a researcher to efficiently investigate a large number of treatments in an initial experiment, with subsequent experiments designed to focus on the most promising lines of investigation or to clarify the interpretation of the original analysis. Many researchers would consider ambiguity in interpreting the outcome of the initial experiment a small price to pay for the reduction in experimental effort.

Hierarchical Designs

The multitreatment designs that have been discussed have had crossed treatments. Treatments A and B are crossed if each level of treatment B appears once with each level of treatment A and visa versa. Treatment B is *nested* in treatment A if each level of treatment B appears with only one level of treatment A . The nesting of treatment B in treatment A is denoted by $B(A)$ and is read, ‘ B within A ’. A hierarchical design (see **Hierarchical Models**) has at least one nested treatment; the remaining treatments are either nested or crossed.

Hierarchical designs are constructed from two or more or a combination of completely randomized and randomized block designs. A two-treatment, hierarchical design that is constructed from two completely randomized designs is denoted by CRH- $pq(A)$, where H indicates a hierarchical design and $pq(A)$ indicates that the design has p levels of treatment A and q levels of treatment $B(A)$ that are nested in treatment A . A comparison of nested and crossed treatments is shown in Figure 16.

Experiments with one or more nested treatments are well suited to research in education, industry, and the behavioral and medical sciences. Consider an example from the medical sciences. A researcher wants to compare the efficacy of a new drug, denoted

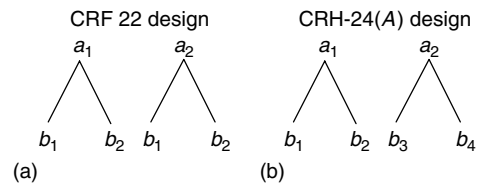


Figure 16 Figure (a) illustrates crossed treatments. In Figure (b), treatment $B(A)$ is nested in treatment A

by a_1 , with the currently used drug, denoted by a_2 . Four hospitals denoted by $b_1, \dots, b_4$, which is treatment $B(A)$, are available to participate in the experiment. Because expensive equipment is needed to monitor the side effects of the new drug, the researcher decided to use the new drug in two of the four hospitals and the current drug in the other two hospitals. The drugs are randomly assigned to the hospitals with the restriction that each drug is assigned to two hospitals. N patients are randomly assigned to the four drug-hospital combinations with the restriction that n patients are assigned to each combination. Figure 16 (b), shown earlier, illustrates the nesting of treatment $B(A)$ in treatment A . The layout for this CRH-24(A) design with two levels of treatment A and four levels of treatment $B(A)$ is shown in Figure 17.

The total sum of squares and total degrees of freedom for the CRH-24(A) design are partitioned as follows:

$$SSTOTAL = SSA + SSB(A) + SSWCELL$$

$$npq_{(j)} - 1 = (p - 1) + p(q_{(j)} - 1) + pq_{(j)}(n - 1), \quad (31)$$

where $q_{(j)}$ is the number of levels of treatment $B(A)$ that is nested in the j th level of treatment A . The design enables a researcher to test two null

hypotheses.

$$H_0: \mu_{1\cdot} = \mu_{2\cdot}$$

(treatment A population means are equal)

$$H_0: \mu_{11} = \mu_{12} \text{ or } \mu_{23} = \mu_{24}$$

(treatment $B(A)$ population means are equal)

(32)

If the second null hypothesis is rejected, the researcher can conclude that the dependent variable is not the same for the populations represented by hospitals b_1 and b_2 , or the dependent variable is not the same for the populations represented by hospitals b_3 and b_4 , or both. However, the test of treatment $B(A)$ does not address the question of whether $\mu_{11} = \mu_{23}$, for example, because hospitals b_1 and b_3 were assigned to different levels of treatment A . Also, because treatment $B(A)$ is nested in treatment A , it is not possible to test the $A \times B$ interaction. The F statistics are

$$F = \frac{SSA/(p - 1)}{SSWCELL/[pq_{(j)}(n - 1)]} = \frac{MSA}{MSWCELL}$$

$$F = \frac{SSB(A)/p(q_{(j)} - 1)}{SSWCELL/[pq_{(j)}(n - 1)]} = \frac{MSB(A)}{MSWCELL}. \quad (33)$$

		Treat. comb.	
Group ₁	Participant ₁	a_1b_1	$\bar{Y}_{\cdot 11}$
	Participant ₅	a_1b_1	
Group ₂	Participant ₆	a_1b_2	$\bar{Y}_{\cdot 12}$
	Participant ₁₀	a_1b_2	
Group ₃	Participant ₁₁	a_2b_3	$\bar{Y}_{\cdot 23}$
	Participant ₁₅	a_2b_3	
Group ₄	Participant ₁₆	a_2b_4	$\bar{Y}_{\cdot 24}$
	Participant ₂₀	a_2b_4	

Figure 17 Layout for a two-treatment, completely randomized hierarchical design, CRH-24(A) design, in which treatment $B(A)$ is nested in treatment A . The twenty participants were randomly assigned to four combinations of treatments A and $B(A)$ with the restriction that five participants were assigned to each treatment combination

As is often the case, the nested treatment in the drug example resembles a nuisance variable. The researcher probably would not conduct the experiment just to find out whether the dependent variable is different for the two hospitals assigned to drug a_1 or the two hospitals assigned to drug a_2 . The important question is whether the new drug is more effective than the currently used drug.

Hierarchical designs with three or more treatments can have both nested and crossed treatments. If at least one treatment is nested and two or more treatments are crossed, the design is a partial hierarchical design. For example, treatment $B(A)$ can be nested in treatment A and treatment C can be crossed with both treatments A and $B(A)$. This design is denoted by CRPH- $pq(A)r$, where PH indicates a partial hierarchical design. The nesting configuration for this design is shown in Figure 18.

Lack of space prevents me from describing other partial hierarchical designs with different combinations of crossed and nested treatments. The interested

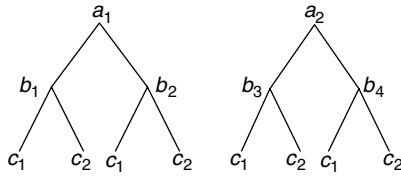


Figure 18 Nesting configuration of the three treatments in a CRPH-24(A)2 design. The four levels of treatment $B(A)$ are nested in the two levels of treatment A ; the two levels of treatment C are crossed with treatments A and $B(A)$

reader is referred to the extensive treatment of these designs in Kirk [10].

Analysis of Covariance

The discussion so far has focused on designs that use *experimental control* to reduce error variance and minimize the effects of nuisance variables. Experimental control can take different forms such as random assignment of participants to treatment levels, stratification of participants into homogeneous blocks, and refinement of techniques for measuring a dependent variable. **Analysis of covariance**, ANCOVA, is an alternative approach to reducing error variance and minimizing the effects of nuisance variables. The approach combines regression analysis with ANOVA and involves measuring one or more *concomitant variables* (also called *covariates*) in addition to the dependent variable. The concomitant variable represents a source of variation that was not controlled in the experiment and a source that is believed to affect the dependent variable. When this approach is used, the letters AC are appended to the designation for a design, for example, CRFAC-*pq*.

ANCOVA enables a researcher to (1) remove that portion of the dependent-variable error variance that is predictable from a knowledge of the concomitant variable thereby increasing power and (2) adjust the dependent-variable means so that they are free of the linear effects attributable to the concomitant variable thereby reducing bias. ANCOVA is often used in three kinds of research situations. One situation involves the use of intact groups with unequal concomitant-variable means and is common in educational and industrial research. The procedure statistically equates the intact groups so that their concomitant-variable means are equal. Unfortunately,

a researcher can never be sure that the concomitant-variable means that are adjusted represent the only nuisance variable or the most important nuisance variable on which the intact groups differ. Random assignment is the best safeguard against unanticipated nuisance variables. In the long run, over many replications of an experiment, random assignment will result in groups that are, at the time of assignment, similar on all nuisance variables.

ANCOVA also can be used to adjust concomitant-variable means when it becomes apparent at some time that although participants were randomly assigned to the treatment levels, the participants in the different groups were not equivalent on a relevant nuisance variable at the beginning of the experiment. Finally, ANCOVA can be used to adjust concomitant-variable means for differences in a relevant nuisance variable that develops during an experiment.

Statistical control and experimental control are not mutually exclusive approaches to reducing error variance and minimizing the effects of nuisance variables. It may be convenient to control some variables by experimental control and others by statistical control. In general, experimental control involves fewer assumptions than statistical control. However, experimental control requires more information about the participants before beginning an experiment. Once data collection has begun, it is too late to randomly assign participants to treatment levels or to form blocks of dependent participants. The advantage of statistical control is that it can be used after data collection has begun. Its disadvantage is that it involves a number of assumptions such as a linear relationship between the dependent and concomitant variables and equal within-groups regression coefficients that may prove untenable in a particular experiment.

Summary of ANOVA Nomenclature and Acronyms

The nomenclature and acronyms for ANOVA designs are summarized in Table 1. The classification of designs in Table 1 is based on (a) the number of treatments, (b) whether participants are assigned to relatively homogeneous blocks prior to random assignment, (c) the building block design, (d) presence or absence of confounding, (e) use of crossed or nested treatments, and (f) use of a covariate. The nomenclature owes much to Cochran and Cox [3] and Federer [5].

Table 1 Classification of ANOVA designs

ANOVA design	Acronym
I Designs with One Treatment	
A Treatment levels randomly assigned to experimental units	
1. Completely randomized design	CR- p
B Experimental units assigned to relatively homogeneous blocks or groups prior to random assignment	
1. Balanced incomplete block design	BIB- p
2. Cross-over design	CO- p
3. Generalized randomized block design	GRB- p
4. Graeco-Latin square design	GLS- p
5. Hyper-Graeco-Latin square design	HGLS- p
6. Latin square design	LS- p
7. Lattice balanced incomplete block design	LBIB- p
8. Lattice partially balanced incomplete block design	LPBIB- p
9. Lattice unbalanced incomplete block design	LUBIB- p
10. Partially balanced incomplete block design	PBIB- p
11. Randomized block design	RB- p
12. Youden square design	YBIB- p
II Designs with Two or More Treatments	
A Factorial designs: designs in which all treatments are crossed	
1. Designs without confounding	
a. Completely randomized factorial design	CRF- pq
b. Generalized randomized block factorial design	GRBF- pq
c. Randomized block factorial design	RB- pq
2. Design with group-treatment confounding	
a. Split-plot factorial design	SPF- $p \cdot q$
3. Designs with group-interaction confounding	
a. Latin square confounded factorial design	LSCF- p^k
b. Randomized block completely confounded factorial design	RBCF- p^k
c. Randomized block partially confounded factorial design	RBPF- p^k
4. Designs with treatment-interaction confounding	
a. Completely randomized fractional factorial design	CRFF- p^{k-i}
b. Graeco-Latin square fractional factorial design	GLSFF- p^k
c. Latin square fractional factorial design	LSFF- p^k
d. Randomized block fractional factorial design	RBFF- p^{k-i}
B Hierarchical designs: designs in which one or more treatments are nested	
1. Designs with complete nesting	
a. Completely randomized hierarchical design	CRH- $pq(A)$
b. Randomized block hierarchical design	RBH- $pq(A)$
2. Designs with partial nesting	
a. Completely randomized partial hierarchical design	CRPH- $pq(A)r$
b. Randomized block partial hierarchical design	RBPF- $pq(A)r$
c. Split-plot partial hierarchical design	SPPH- $p \cdot qr(B)$
III Designs with One or More Covariates	
A Designs can include a covariate in which case the letters AC are added to the acronym as in the following examples.	
1. Completely randomized analysis of covariance design	CRAC- p
2. Completely randomized factorial analysis of covariance design	CRFAC- pq
3. Latin square analysis of covariance design	LSAC- p
4. Randomized block analysis of covariance design	RBAC- p
5. Split-plot factorial analysis of covariance design	SPFAC- $p \cdot q$

A wide array of designs is available to researchers. Hence, it is important to clearly identify designs in research reports. One often sees statements such as ‘a two-treatment, factorial design was used’. It should be evident that a more precise description is required. This description could refer to 10 of the 11 factorial designs in Table 1.

References

- [1] Anderson, N.H. (2001). *Empirical Direction in Design and Analysis*, Lawrence Erlbaum, Mahwah.
- [2] Bogartz, R.S. (1994). *An Introduction to the Analysis of Variance*, Praeger, Westport.
- [3] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Edition, John Wiley, New York.
- [4] Cobb, G.W. (1998). *Introduction to Design and Analysis of Experiments*, Springer-Verlag, New York.
- [5] Federer, W.T. (1955). *Experimental Design: Theory and Application*, Macmillan, New York.
- [6] Harris, R.J. (1994). *ANOVA: An Analysis of Variance Primer*, F. E. Peacock, Itasca.
- [7] Hicks, C.R. & Turner Jr, K.V. (1999). *Fundamental Concepts in the Design of Experiments*, Oxford University Press, New York.
- [8] Jones, B. & Kenward, M.G. (2003). *Design and Analysis of Cross-over Trials*, 2nd Edition, Chapman & Hall, London.
- [9] Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*, 3rd Edition, Prentice-Hall, Englewood Cliffs.
- [10] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [11] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [12] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Design*, 3rd Edition, McGraw-Hill, New York.

(See also **Generalized Linear Mixed Models; Linear Multilevel Models**)

ROGER E. KIRK

Analysis of Variance: Multiple Regression Approaches

RICHARD J. HARRIS

Volume 1, pp. 83–93

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Analysis of Variance: Multiple Regression Approaches

Consider the following questions:

1. Is whether one has received the experimental treatment or the control ‘treatment’ predictive of one’s score on the dependent variable?
2. Do the experimental and control groups have different population means?
3. Is the group to which a participant was assigned predictive of that person’s score on the dependent variable?
4. Do the k groups have different population means?

Questions 1 and 2 appear to be mirror images of each other, essentially asking whether there is a relationship between the independent and dependent variables. The same can be said for questions 3 and 4. These appearances are *not* deceiving. As we shall see, there is a simple algebraic relationship between the correlation coefficient that is used to answer question 1 and the independent-means t ratio that is used to answer question 2. Similarly, there is a simple algebraic relationship between the multiple correlation coefficient that is used to answer question 3 and the overall F ratio that is used to answer question 4. At first glance, questions 1 and 3 appear to ask for more detailed analyses than do questions 2 and 4. That is, they ask for the prediction of individual participants’ scores, rather than simply comparing group means. This appearance *is* deceiving. The only information on which we can base our prediction of a given participant’s score is the group to which the participant was assigned or the group to which the participant naturally belongs. Hence, we must inevitably predict the same score for every individual in a given group and, if we are to minimize the average squared error of prediction, we must therefore focus on predicting the mean score for each group.

Consider two facts. First, for a two-group design, testing for the statistical significance of the correlation coefficient between a group-membership variable and a dependent variable yields the same conclusion as conducting an independent-means t Test of

the difference between the two means. Second, for a k -group design, testing for the statistical significance of the multiple correlation (see **R -squared**, **Adjusted R -squared**) between a set of $k - 1$ level-membership or contrast variables and the dependent variable yields the same conclusion as does the overall F ratio for the differences among the k means. These two facts do not imply, however, that there is nothing but mathematical elegance to be gained by comparing the two pairs of approaches. In particular, the correlation and regression approaches help us to appreciate that

1. large, highly significant t and F ratios may represent relationships that in fact account for rather small percentages of the total variance of a dependent variable;
2. the choice of alternative models for unequal- n factorial ANOVA designs (see **Type I**, **Type II and Type III Sums of Squares**; **Factorial Designs**) is not just a choice of error terms but also of the particular contrasts among the means that are being tested; and
3. while overall tests in unequal- n designs require the use of matrix algebra, single-degree-of-freedom tests of specific contrasts can be conducted by means of very simple algebraic formulae (see **Multiple Comparison Procedures**).

Let us demonstrate these facts and their consequences with a few examples.

Equivalence between Independent-means t and Dichotomy/Dependent Variable Correlation

Consider the data (Table 1) that are taken from Experiment 3 of Harris and Joyce [3].

From the above data, we compute $r_{XY} = -0.5396$. Applying the usual t Test for the significance of a correlation coefficient we obtain

$$\begin{aligned}
 t &= \frac{r_{XY}}{\sqrt{\frac{1 - r_{XY}^2}{N - 2}}} = \frac{-0.5396}{\sqrt{\frac{1 - 0.29117}{18}}} \\
 &= \frac{-0.5396}{0.1984} = -2.720
 \end{aligned} \tag{1}$$

2 Analysis of Variance: Multiple Regression Approaches

Table 1 Amount Allocated to 10-Interchange Partner as f(Allocation Task)

$X = \text{Allocn}$	$Y = \text{P10Outc}$	$X = \text{Allocn}$	$Y = \text{P10Outc}$
1	\$1.70	2	-\$1.50*
1	1.00	2	0.50
1	0.07	2	-1.50
1	0.00	2	-1.50
1	1.30	2	-1.00
1	3.00	2	2.10
1	0.50	2	2.00
1	2.00	2	-0.50
1	1.70	2	-1.50
1	0.50	2	0.00

Notes: $X = 1$ indicates that the group was asked to allocate the final shares of a prize directly. $X = 2$ indicates that they were asked how much of a room fee should be subtracted from each partner's individual contribution to determine his or her final share of the prize.

Source: Raw data supplied by first author of [3].

with 18 df , $p < 0.01$. Thus, we can be quite confident that having a higher score on X (i.e., being one of the groups in the expenses-allocation condition) is, in the population, associated with having a lower score on Y (i.e., with recommending a lower final outcome for the person in your group who worked on the most difficult problems).

Had we simply omitted the two columns of scores on X , the data would look just like the usual setup for a test of the difference between two independent means, with the left-hand column of Y scores providing group 1's recommendations and the right-hand column of Y scores giving group 2's scores on that dependent variable. Applying the usual formula for an independent-means t Test gives

$$\begin{aligned}
 t &= \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sum (Y_1 - \bar{Y}_1)^2 + \sum (Y_2 - \bar{Y}_2)^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\
 &= \frac{1.177 - (-0.290)}{\sqrt{\frac{8.1217 + 18.0690}{18} (1/10 + 1/10)}} \\
 &= \frac{1.467}{0.53945} = 2.719
 \end{aligned} \tag{2}$$

with $n_1 + n_2 - 2 = 18$ df , $p < 0.01$. Thus, we can be quite confident that, in the population, groups who determine final outcomes only indirectly by

allocation of shares of the room fee give lower outcomes to the partner who deals with the most difficult anagrams than do groups who determine final outcomes directly. That is, having a high score on X is associated with having a lower Y score – which is what we concluded on the basis of the test of our correlation coefficient.

The formula for computing the independent-means t from the correlation coefficient is just the formula given earlier for testing the statistical significance of a correlation coefficient. Inverting that relationship gives us

$$r_{XY} = \frac{t}{\sqrt{t^2 + df}}, \tag{3}$$

where df = the degrees of freedom for the t Test, namely, $n_1 + n_2 - 2$.

Thus, in the present example,

$$\begin{aligned}
 r_{XY} &= \frac{-2.719}{\sqrt{7.3930 + 18}} \\
 &= \frac{-2.719}{5.03915} = -0.5396.
 \end{aligned} \tag{4}$$

An Application of the Above Relationship

An important application of the above relationship is that it reminds us of the distinction between statistical significance – confidence that we've got the *sign* of an effect right – and substantive significance – the estimated *magnitude* of an effect (see **Effect Size Measures**). For example, ASRT's *Environmental Scan of the RadiationTherapist's Workplace* found that the mean preference for a great work environment over a great salary was significantly greater among the staff-therapist-sample respondents who still held a staff or senior-staff therapist title than among those who had between the time of their most recent certification renewal and the arrival of the questionnaire moved on to another position, primarily medical dosimetrist or a managerial position within the therapy suite. The t for the difference between the means was 3.391 with 1908 df , $p < 0.001$. We can be quite confident that the difference between the corresponding population means is in the same direction. However, using the above formula tells us that the correlation between the still-staff-therapist versus moved-on distinction is $r = 3.391/\sqrt{1919.5} = 0.0774$. Hence, the distinction accounts for $(0.0774)^2 = 0.6\%$, which is less

than one percent of the variation among the respondents in their work environment versus salary preferences.

As this example shows, it is instructive to convert ‘experimental’ statistics into ‘correlational’ statistics. That is, convert t ’s into the corresponding r^2 ’s. The resulting number can come as a surprise; many of the r^2 ’s for statistically significant differences in means will be humbly low.

Equivalence Between Analysis of Variance and Multiple Regression with Level-membership or Contrast Predictors

Consider the hypothetical, but not atypical, data shown in Table 2.

Note that within each college, the female faculty members’ mean salary exceeds the male faculty members’ mean salary by \$5000–\$10 000. On the other hand, the female faculty is concentrated in the low-paying College of Education, while a slight majority of the male faculty is in the high-paying College of Medicine. As a result, whether ‘on average’ female faculty are paid more or less than male faculty depends on what sort of mean we use to define ‘on average’. An examination of the *unweighted mean* salaries (cf. the ‘Unweighted mean’ row toward the bottom of the table) of the males and females in the three colleges (essentially a per-college mean), indicates that female faculty make, on average, \$6667 more per year than do male faculty. If instead we compute for each gender the *weighted mean* of the three college means, weighting by the number of faculty members of that gender in each college, we

find that the female faculty members make, on average, \$6890 *less* than do male faculty members (see **Type I, Type II and Type III Sums of Squares**). These results are the same as those we would have obtained had we simply computed the mean salary on a per-individual basis, ignoring the college in which the faculty member taught (see **Markov, Andrei Andreevich**).

For the present purposes, this ‘reversal paradox’ (see **Odds and Odds Ratios**) [6] helps sharpen the contrast among alternative ways of using **multiple linear regression** analysis (MRA) to analyze data that would normally be analyzed using ANOVA. This, in turn, sheds considerable light on the choice we must make among alternative models when carrying out a factorial ANOVA for unbalanced designs. In an unbalanced design, the percentage of the observations for a given level of factor A differs across the various levels of factor B. The choice we must make is usually thought of as bearing primarily on whether a given effect in the ANOVA design is statistically significant. However, the core message of this presentation is that it is also – and in my opinion, more importantly – a choice of what kinds of means are being compared in determining the significance of that effect. Specifically, a completely uncorrected model involves comparisons among the *weighted means* and is thus, for main effects, equivalent to carrying out a one-way ANOVA for a single factor, ignoring all other factors. Furthermore, the analysis makes no attempt to correct for confounds with those other factors. A completely uncorrected model is equivalent to testing a regression equation that includes only the contrast variables for the particular effect being tested. A fully corrected model

Table 2 Mean Faculty Salary* at Hypothetical U as f (College, Gender)

College	Gender						Unweighted mean	Weighted mean
	Males			Females				
	Mean ^a	Std. Dev.	<i>n</i>	Mean ^a	Std. Dev.	<i>n</i>		
Engineering	30	1.491	55	35	1.414	5	32.5	30.416
Medicine	50	1.423	80	60	1.451	20	55.0	52
Education	20	1.451	20	25	1.423	80	22.5	24
Unweighted mean	33.333			40			36.667	
Weighted mean	39.032			32.142			36.026	

^aSalaries expressed in thousands of dollars.

Source: Adapted from [1] (Example 4.5) by permission of author/copyright holder.

4 Analysis of Variance: Multiple Regression Approaches

Table 3 Raw Data, Hypothetical U Faculty Salaries

Group	College	Gender	Sal	nij
Engnr-M	1	1	30.000	25
	1	1	28.000	15
	1	1	32.000	15
Engnr-F	1	2	33.000	1
	1	2	35.000	3
	1	2	37.000	1
Medcn-M	2	1	48.000	20
	2	1	50.000	40
	2	1	52.000	20
Medcn-F	2	2	58.000	5
	2	2	60.000	10
	2	2	62.000	5
Educn-M	3	1	18.000	5
	3	1	20.000	10
	3	1	22.000	5
Educn-F	3	2	23.000	20
	3	2	25.000	40
	3	2	27.000	20

involves comparisons among the *unweighted* means. A fully corrected model is equivalent to testing each effect on the basis of the increment to R^2 that results from adding the contrast variables representing that effect last, after the contrast variables for all other effects have been entered. And in-between models, where any given effect is corrected for confounds with from zero to all other effects, involve contrasts that are unlikely to be interesting and correspond to questions that the researcher wants answered. I will note one exception to this general condemnation of in-between models – though only as an intermediate

step along the way to a model that corrects only for interactions of a given order or lower.

You can replicate the analyses I am about to use to illustrate the above points by entering the following variable names and values into an SPSS™ data editor (aka .sav file). One advantage of using hypothetical data is that we can use lots of identical scores and thus condense the size of our data file by employing the ‘Weight by’ function in SPSS (Table 3).

One-way, Independent-means ANOVA

First, I will conduct a one-way ANOVA of the effects of College, ignoring for now information about the gender of each faculty member. Submitting the following SPSS commands

```
Title Faculty Salary example .
Weight by nij .
Subtitle Oneway for college effect .
Manova sal by college (1,3) /
  Print = cellinfo (means) signif
    (univ) design (solution) /
  Design /
  Contrast (college) = special
    (1 1 1, -1 2 -1, 1 0 -1) /
Design = college (1), college (2).
Weight off .
```

yields (in part) the output as shown in Tables 4 and 5.

Notice that the means are identical to those listed in the ‘Weighted mean’ column of the data table shown previously.

Table 4 Cell means and standard deviations

Variable.. SAL		Salary (thousands of dollars)		
FACTOR	CODE	Mean	Std. Dev.	N
COLLEGE	Engineering	30.417	2.028	60
COLLEGE	Medicine	52.000	4.264	100
COLLEGE	Education	24.000	2.462	100
For entire sample		36.250	13.107	260

Table 5 Tests of significance for SAL using unique sums of squares; analysis of variance – design 1

Source of Variation	SS	DF	MS	F	Sig of F
WITHIN CELLS	2642.58	257	10.28		
COLLEGE	41854.17	2	20927.08	2035.23	0.000

To accomplish the same overall significance test using MRA, we need a set of two predictor variables (more generally, $k - 1$, where k = number of independent groups) that together completely determine the college in which a given faculty member works. There are many alternative sets of predictor variables (see **Regression Model Coding for the Analysis of Variance and Type I, Type II and Type III Sums of Squares**), but it is generally most useful to construct *contrast variables*. This is accomplished by choosing a set of $k - 1$ interesting or relevant contrasts among our k means. We then set each case's score on a given contrast equal to the contrast coefficient that has been assigned to the group within which that case falls. Let us use the contrast between the high-paying College of Medicine and the other two colleges as our first contrast (labeled 'medvoth' in our SPSS data file) and the contrast between the Colleges of Engineering and Education (labeled 'engveduc') as our second contrast. This yields the expanded .sav file as shown in Table 6.

Recall that contrast coefficients are defined only up to a multiplicative constant. For example, $\mu_{\text{Med}} - 0.5\mu_{\text{Engn}} - 0.5\mu_{\text{Educ}} = 0$ if and only if $2\mu_{\text{Med}} - \mu_{\text{Engn}} - \mu_{\text{Educ}} = 0$. The 'extra' three columns in the expanded .sav file above give contrasts for the Gender main effect and the interaction between Gender and each of the two College contrasts; more about these later. We then run an MRA of Salary predicted from

MedvOth and EngvEduc, for example, by submitting the following SPSS commands.

```
Weight off .
Subtitle One-way for College and
      College-First Sequential .
Weight by nij .
Regression variables = sal mfcontr
      medvoth engveduc gbcoll1 gbcoll2 /
Statistics = defaults cha /
Dep = sal / enter medvoth engveduc /
      enter gbcoll1 gbcoll2 /
      enter mfcontr/.
Weight off .
```

Only the first step of this stepwise regression analysis – that in which only the two College contrasts have been entered – is relevant to our one-way ANOVA. I will discuss the remaining two steps shortly.

The resulting SPSS run yields, in part, the output as shown in Tables 7 and 8.

The test for the statistical significance of R^2 consists of comparing $F = R^2/(k - 1)/(1 - R^2)/$

Table 7 Model Summary

Model	R	R^2	Adjusted R^2
1	0.970 ^a	0.941	0.940

^aPredictors: (Constant), ENGVEDUC, MEDVOTH.

Table 6 Scores on Contrast Variables

Group	College	Gender	Sal	nij	mfcontr	medvoth	engveduc	gbcoll1	gbcoll2
Engnr-M	1	1	30.000	25	1	-1	1	-1	1
	1	1	28.000	15	1	-1	1	-1	1
	1	1	32.000	15	1	-1	1	-1	1
Engnr-F	1	2	33.000	1	-1	-1	1	1	-1
	1	2	35.000	3	-1	-1	1	1	-1
	1	2	37.000	1	-1	-1	1	1	-1
Medcn-M	2	1	48.000	20	1	2	0	2	0
	2	1	50.000	40	1	2	0	2	0
	2	1	52.000	20	1	2	0	2	0
Medcn-F	2	2	58.000	5	-1	2	0	-2	0
	2	2	60.000	10	-1	2	0	-2	0
	2	2	62.000	5	-1	2	0	-2	0
Educn-M	3	1	18.000	5	1	-1	-1	-1	-1
	3	1	20.000	10	1	-1	-1	-1	-1
	3	1	22.000	5	1	-1	-1	-1	-1
Educn-F	3	2	23.000	20	-1	-1	-1	1	1
	3	2	25.000	40	-1	-1	-1	1	1
	3	2	27.000	20	-1	-1	-1	1	1

6 Analysis of Variance: Multiple Regression Approaches

Table 8 ANOVA^a

Model		Sum of Squares	<i>df</i>	Mean Square	<i>F</i>	Sig.
1	Regression	41854.167	2	20927.083	2035.228	0.000 ^b
	Residual	2642.583	257	10.282		
	Total	44496.750	259			

^aDependent Variable: Salary (thousands of dollars). ^bPredictors: (Constant), ENGVEDUC, MEDVOTH.

Table 9 Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		<i>t</i>	Sig.
		B	Std. Error	Beta			
1	(Constant)	35.472	0.205			173.311	0.000
	MEDVOTH	8.264	0.138	0.922		59.887	0.000
	ENGVEDUC	3.208	0.262	0.189		12.254	0.000

^aDependent Variable: Salary (thousands of dollars).

Table 10 Tests of Significance for SAL using UNIQUE sums of squares; analysis of variance – design 2

Source of Variation	SS	DF	MS	<i>F</i>	Sig of <i>F</i>
WITHIN + RESIDUAL	2642.58	257	10.28		
COLLEGE(1)	36877.60	1	36877.60	3586.47	0.000
COLLEGE(2)	1544.01	1	1544.01	150.16	0.000

$(N - k)$ to the critical value for an F with $k - 1$ and $N - k$ *df*, yielding the F of 2035.228 reported in the above table, which matches the overall F for College from our earlier ANOVA table.

But, what has been gained for the effort involved in defining contrast variables? The MRA output continued with the listing of regression coefficients (Table 9) and tests of the statistical significance thereof.

We want to compare these MRA-derived t Tests for the two contrasts to the corresponding ANOVA output (Table 10) generated by our Contrast subcommand in conjunction with the expanded design statement naming the two contrasts.

Recall that the square of a t is an F with 1 *df* in the numerator. We see that the t of 59.887 for the significance of the difference in mean salary between the College of Medicine and the average of the other two colleges corresponds to an F of 3586.45. This value is equal, within round-off error, to the ANOVA-derived value. Also, the square of the t for the Engineering versus Education contrast, $(12.254)^2 =$

150.16, matches equally closely the ANOVA-derived F . Notice, too, that the unstandardized regression coefficients for the two contrasts are directly proportional to signs and magnitudes of the corresponding contrasts: 3.208 (the regression coefficient for EngvEduc) is exactly half the difference between those two means, and 8.264 (the coefficient for MedvOth) exactly equals one-sixth of $2(\text{mean for Medicine}) - (\text{mean for Engineering}) - (\text{mean for Education})$. The divisor in each case is the sum of the squared contrast coefficients.

Factorial ANOVA via MRA

We will add to the data file three more contrast variables: MFContr to represent the single-*df* Gender effect and GbColl1 and GbColl2 to represent the interaction between Gender and each of the previously selected College contrasts. We need these variables to be able to run the MRA-based analysis. More importantly, they are important in interpreting the particular patterns of differences among means

that are responsible for statistically significant effects. See the article in this encyclopedia on **regression model coding in ANOVA** for details of how to compute coefficients for an interaction contrast by cross multiplying the coefficients for one contrast for each of the two factors involved in the interaction. We can then test each of the two main effects and the interaction effect by testing the statistical significance of the increment to R^2 that results from adding the contrast variables representing that effect to the regression equation. This can be done by computing

$$F_{\text{incr}} = \frac{\frac{(\text{Increase in } R^2)}{(\text{number of predictors added})}}{\frac{(R^2 \text{ for full model})}{(N - \text{total \# of predictors} - 1)}} \quad (5)$$

or by adding 'Cha' to the statistics requested in the SPSS Regression command.

For equal- n designs

1. all sets of mutually orthogonal contrasts are also uncorrelated;
2. the increment to R^2 from adding any such set of contrasts is the same, no matter at what point they are added to the regression equation; and

3. the F s for all effects can be computed via simple algebraic formulae; see [1], any other ANOVA text, or the article on factorial ANOVA in this encyclopedia.

However, when n s are unequal (unless row- and column-proportional),

1. Orthogonal contrasts (sum of cross-products of coefficients equal zero) are, in general, not uncorrelated;
2. The increment to R^2 from adding any such set of contrasts depends on what other contrast variables are already in the regression equation; and
3. Computing the F for any multiple- df effect requires the use of matrix algebra.

To illustrate this context dependence, consider the following additional output (Table 11) from our earlier stepwise-MRA run. There, we entered the College contrasts first, followed by the interaction contrasts and then by the Gender main-effect contrast, together with a second stepwise-MRA run in which the order of entry of these effects is reversed (Table 11, 12).

Table 11 Model Summary (College, then CxG interaction, then Gender)

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics					
					R^2 Change	F Change	$df1$	$df2$	Sig. F Change	
1	0.970 ^a	0.941	0.940	3.206622	0.941	2035.228	2	257	0.000	← College
2	0.981 ^b	0.962	0.961	2.577096	0.021	71.447	2	255	0.000	← Interaction
3	0.994 ^c	0.988	0.988	1.441784	0.026	560.706	1	254	0.000	← Gender added

^aPredictors: (Constant), ENGVEDUC, MEDVOTH.

^bPredictors: (Constant), ENGVEDUC, MEDVOTH, GBCOLL2, GBCOLL1.

^cPredictors: (Constant), ENGVEDUC, MEDVOTH, GBCOLL2, GBCOLL1, MFCONTR.

Table 12 Model Summary (Gender, then C x G interaction, then College)

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics					
					R^2 Change	F Change	$df1$	$df2$	Sig. F Change	
1	0.258 ^a	0.067	0.063	12.68669	0.067	18.459	1	258	0.000	
2	0.463 ^b	0.214	0.205	11.68506	0.148	24.063	2	256	0.000	
3	0.994 ^c	0.988	0.988	1.44178	0.774	8280.598	2	254	0.000	

^aPredictors: (Constant), MFCONTR.

^bPredictors: (Constant), MFCONTR, GBCOLL1, GBCOLL2.

^cPredictors: (Constant), MFCONTR, GBCOLL1, GBCOLL2, MEDVOTH, ENGVEDUC.

8 Analysis of Variance: Multiple Regression Approaches

Table 13 Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients Beta	<i>t</i>	Sig.
		<i>B</i>	Std. Error			
1 (Gender only)	(Constant)	35.588	0.802		44.387	0.000
	MFCONTR	3.445	0.802	0.258	4.296	0.000
2 (Gender and G x C)	(Constant)	37.967	0.966		39.305	0.000
	MFCONTR	1.648	0.793	0.124	2.079	0.039
	GBCOLL1	1.786	0.526	0.188	3.393	0.001
	GBCOLL2	−6.918	1.179	−0.349	−5.869	0.000
3 (Full model)	(Constant)	36.667	0.141		260.472	0.000
	MFCONTR	−3.333	0.141	−0.250	−23.679	0.000
	GBCOLL1	−0.833	0.088	−0.088	−9.521	0.000
	GBCOLL2	3.654E-15	0.191	0.000	0.000	1.000
	MEDVOTH	9.167	0.088	1.023	104.731	0.000
	ENGVEDUC	5.000	0.191	0.294	26.183	0.000

^aDependent Variable: Salary (thousands of dollars).

Order of entry of an effect can affect not only the magnitude of the *F* for statistical significance but even our estimate of the direction of that effect, as shown in the regression coefficients for Gender in the various stages of the ‘Gender, C x G, College’ MRA (Table 13).

Notice that the Gender contrast is positive, indicative of higher salaries for males, when Gender is the first effect entered into the equation, but negative when it is entered last. Notice, also that the test for significance of a regression coefficient in the full model is logically and arithmetically identical to the test of the increment to R^2 when that contrast variable is the last one added to the regression equation. This is of course due to the fact that different contrasts are being tested in those two cases: When Gender is ‘first in,’ the contrast being tested is the difference between the weighted means, and the *B* coefficient for MFCONTR equals half the difference (\$6890) between the mean of the 155 males’ salaries and the mean of the 105 females’ salaries. When Gender is ‘last in,’ the contrast being tested is the difference between the unweighted means, and the *B* coefficient for MFCONTR equals half the difference (−\$6667) between the mean of the three college means for males and the mean of the three college means for females. Each of these comparisons is the right answer to a different question.

The difference between weighted means – what we get when we test an effect when entered first –

is more likely to be relevant when the unequal *ns* are a reflection of preexisting differences in representation of the various levels of our factor in the population. Though even in such cases, including the present example, we will probably also want to know what the average effect is within, that is, controlling for the levels of the other factor. Where the factors are manipulated variables, we are much more likely to be interested in the differences among unweighted means, because these comparisons remove any confounds among the various factors.

But, what is being tested when an effect is neither first nor last into the equation? A little known, or seldom remarked upon, aspect of MRA is that one can, with the help of a bit of matrix algebra, express each sample regression coefficient as a linear combination of the various subjects’ scores on the dependent variable. See section 2.2.4 of Harris [2] for the details. But when we are using MRA to analyze data from an independent-means design, factorial, or otherwise, every subject in a particular group who receives a particular combination of one level of each of the factors in the design has exactly the same set of scores on the predictor variables. Hence, all of those subjects’ scores on *Y* must be given the same weight in estimating our regression coefficient. Thus, the linear combination of the individual *Y* scores that is used to estimate the regression coefficient must perforce also be a linear combination of contrasts among the means and therefore also a contrast among

Table 14 Solution Matrix for Between-Subjects Design

1-COLLEGE		2-GENDER		PARAMETER			
FACTOR							
1	2	1	2	3	4	5	6
1	1	-0.707	0.866	-0.500	0.707	0.866	-0.500
1	2	-0.707	0.866	-0.500	-0.707	-0.866	0.500
2	1	-0.707	0.000	1.000	0.707	0.000	1.000
2	2	-0.707	0.000	1.000	-0.707	0.000	-1.000
3	1	-0.707	-0.866	-0.500	0.707	-0.866	-0.500
3	2	-0.707	-0.866	-0.500	-0.707	0.866	0.500

those means. SPSS's MANOVA command gives, if requested to do so by including 'Design (Solution)' in the list of requested statistics, the contrast coefficients for the contrasts it actually tests in any given analysis. For instance, for the full-model analysis of the faculty salary data, the solution matrix is as shown in Table 14.

Thus, the full-model analysis tests the contrasts we specified, applied to the unweighted means. For example, the Gender effect in column 4 compares the average of the three male means to the average of the three female means. Consider another example. We specify 'Method = Sequential' and indicate by specifying 'Design = College, Gender, Gender by College' that we want College to be tested first (uncorrected for confounds with Gender or $G \times C$) and Gender to be tested second (corrected for confounds with College but not with $G \times C$). The column of the solution matrix (not reproduced here) corresponding to the Gender effect tells us that the contrast actually being tested is

$$\begin{aligned} &0.758\mu_{\text{Engnr-M}} + 2.645\mu_{\text{Med-M}} \\ &+ 2.645\mu_{\text{Educ-M}} - 0.758\mu_{\text{Engnr-F}} \\ &- 2.645\mu_{\text{Med-F}} - 2.645\mu_{\text{Educ-F}}. \end{aligned}$$

On the other hand, if the interaction effect is entered first, followed by Gender, the contrast being tested is

$$\begin{aligned} &4.248\mu_{\text{Engnr-M}} + 3.036\mu_{\text{Med-M}} \\ &+ 0.087\mu_{\text{Educ-M}} - 0.666\mu_{\text{Engnr-F}} \\ &- 1.878\mu_{\text{Med-F}} - 4.827\mu_{\text{Educ-F}}. \end{aligned}$$

Finally, whenever Gender is entered first, the contrast we are testing is

$$2.807\mu_{\text{Engnr-M}} + 4.083\mu_{\text{Med-M}}$$

$$\begin{aligned} &+ 1.021\mu_{\text{Educ-M}} - 0.377\mu_{\text{Engnr-F}} \\ &- 1.507\mu_{\text{Med-F}} - 6.028\mu_{\text{Educ-F}} \\ &= 7.91[(55\mu_{\text{Engnr-M}} + 80\mu_{\text{Med-M}} \\ &+ 20\mu_{\text{Educ-M}})/155 - (5\mu_{\text{Engnr-F}} \\ &+ 20\mu_{\text{Med-F}} + 80\mu_{\text{Educ-F}})/105], \quad (6) \end{aligned}$$

that is, the difference between the weighted means for males versus females.

Kirk's [4] and Maxwell and DeLaney's [5] texts provide explanations for some of the contrasts that arise when testing partially corrected, 'in-between,' effects. For example, effects involving contrasts among means weighted by the harmonic mean of the cell sizes for the various levels of the other factor. However, the chances that such contrasts would be of any interest to a researcher seem to the author to be remote. It seems advisable, therefore, to choose between conducting a full-model analysis, the default in SPSS' MANOVA, or testing each effect uncorrected for any other effects. This is one of the options provided by SAS as *Type I Sums of Squares*. This also can be accomplished in SPSS by carrying out one run for each effect that is to be tested and specifying that effect as the first to be entered into the equation.

If you do decide to conduct a sequential analysis of an unequal- n factorial design, you will need to be careful about what means to compare when describing your results. If you ask SPSS to print out marginal means by using the Omeans subcommand, as in 'Omeans = Tables Gender, College,' there's no need to print means for the highest-order interaction. This follows because it duplicates the output of 'Print = Cellinfo (Means)'. SPSS will report both the weighted and the unweighted means for each selected effect, as shown in Table 15.

Table 15 Analysis of variance design 1

Combined Observed Means for GENDER			
Variable.. SAL			
GENDER			
Males	WGT.		39.03226
	UNWGT.		33.33333
Females	WGT.		32.14286
	UNWGT.		40.00000
Combined Observed Means for COLLEGE			
Variable.. SAL			
COLLEGE			
Engineer	WGT.		30.41667
	UNWGT.		32.50000
Medicine	WGT.		52.00000
	UNWGT.		55.00000
Educatio	WGT.		24.00000
	UNWGT.		22.50000

For whichever effect is tested first in your sequential analysis, you should compare the weighted means labeled 'WGT,' because those are what the significance test actually tested. The effect that is tested last should be described in terms of the unweighted means labeled 'UNWGT'. And describing any effect that is tested in an in-between position requires that you (a) describe what contrasts were actually tested in deriving the overall significance test for the effect and the tests of follow-up contrasts, using the entries in the solution matrix to guide you, and (b) use the coefficients in that solution matrix to compute with a calculator the direction and magnitude of each component contrast, because neither the weighted nor the unweighted means provided by SPSS were the ones used to estimate this in-between contrast.

There is one exception to the above abjuration of in-between contrasts that is likely to be useful and is usually referred to as the 'experimental' order. In this case, each main effect is corrected for the other main effects but not for any interactions. Two-way interactions are corrected for all main effects and the other two-way interactions, but not for 3-way or higher-order interactions, and so on. This order of testing holds out the promise of finding that none of the interactions higher than a given order, for example, no three- or four-way interactions are large or statistically significant and thus being able to simplify the model used to explain responses. However, once all the effects that are to be retained in the simplified model have

been determined, each retained effect should then be retested, correcting each for all other retained effects.

Hand Computation of Full-model Contrasts

Our exploration of unequal- n factorial designs has relied on the use of computer programs such as SPSS. However, another little-remarked-upon aspect of such designs is that the full-model (fully corrected) test of any single- df contrast can be conducted via the straightforward (if sometimes tedious) formula,

$$F_{\text{contr}} = \frac{SS_{\text{contr}}}{MS_w}, \text{ where}$$

$$SS_{\text{contr}} = \frac{\left(\sum_{j=1}^{n_{\text{groups}}} c_j \bar{Y}_j \right)^2}{\sum c_j^2 / n_j}. \quad (7)$$

The important thing to keep in mind about this formula is that a contrast coefficient must be applied to and a c_j^2/n_j term computed for each individual cell in the design, that is, each combination of a level of each of the factors. Thus, for example, SS_{contr} for the main effect of gender in the faculty salary example would be computed as

$$\frac{[(1)\bar{Y}_{\text{Engnr-M}} + (1)\bar{Y}_{\text{Med-M}} + (1)\bar{Y}_{\text{Educ-M}} + (-1)\bar{Y}_{\text{Engnr-F}} + (-1)\bar{Y}_{\text{Med-F}} + (-1)\bar{Y}_{\text{Educ-F}}]^2}{1/55 + 1/80 + 1/20 + 1/5 + 1/20 + 1/80}.$$

Significance tests for multiple- df effects such as the College main effect require the use of matrix algebra or a computer program.

References

- [1] Harris, R.J. (1994). *ANOVA: An Analysis of Variance Primer*, F. E. Peacock, Itasca.
- [2] Harris, R.J. (2001). *A Primer of Multivariate Statistics*, 3rd Edition, Lawrence Erlbaum Associates, Mahwah.
- [3] Harris, R.J. & Joyce, M.A. (1980). What's fair? It depends on how you phrase the question, *Journal of Personality and Social Psychology* **38**, 165–179.
- [4] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Monterey.

- [5] Maxwell, S.E. & Delaney, H.D. (2003). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Lawrence Erlbaum Associates, Mahwah. (See also **Analysis of Variance: Cell Means Approach**)
- [6] Messick, D.M. & Van de Geer, J.P. (1981). A reversal paradox, *Psychological Bulletin* **90**, 582–593.

RICHARD J. HARRIS

Ansari–Bradley Test

CLIFFORD E. LUNNEBORG

Volume 1, pp. 93–94

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ansari–Bradley Test

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_m)$ be independent random samples from two distributions having known medians or a common median but potentially different variances, $\text{var}(X)$ and $\text{var}(Y)$.

The parameter of interest is the ratio of variances, $\gamma = \text{var}(X)/\text{var}(Y)$. The usual hypothesis to be nullified is that $\gamma = 1$. The substantive (alternative) hypothesis is either that $\gamma < 1$, that $\gamma > 1$, or that γ differs from 1.

For the Ansari–Bradley test [1], the combined set of $(n + m)$ observations first is ordered from smallest to largest after having been adjusted for different population medians. Then, ranks are assigned to the observations as follows. The largest and smallest observations are assigned ranks of 1. The second-largest and second-smallest observations are assigned ranks of 2. This process is continued until all observations have been assigned a rank.

The test statistic is the sum of the ranks assigned to the observations in the smaller of the two samples, $g = \text{Sum}(R_j)$, $j = 1, \dots, m$. By convention the smaller sample is identified as the sample from the Y distribution and, hence, the variance in this population is the denominator in the variance ratio γ .

Naively, the statistic g will be small if the dispersion in the Y population is great and large if the population dispersion is limited. Thus, larger values of g are consistent with $\gamma > 1$ and smaller values of g are consistent with $\gamma < 1$.

Tail probabilities for the null distribution of g have been tabulated [3] for a range of values of m and n where n is no smaller than m . A normal approximation to the null distribution is used for larger samples, for example, when $(n + m)$ is greater than 20 [3].

Example A random sample of 20 tenth-graders is randomly divided into two samples, each of size 10.

For the students in one of the samples, mathematics achievement is assessed using Edition 2001 of a standard test. Students in the second sample are assessed with Edition 2003 of the test. Scaled scores on the two editions are known to have the same median. However, there is suggestive evidence in earlier studies that the variability in scores on Edition 2003 may be greater.

The sampled test scores are given in the first two rows of Table 1. Their Ansari–Bradley ranks appear in the third and fourth rows.

The tied observations (91s, 94s, 102s) were awarded the average rank for each set of ties. The sum of ranks for the Edition 2003 sample is in the predicted direction: that is, $47.5 < 62.5$. Under the null hypothesis, the tabulated probability of a rank sum of 47 or smaller is 0.10 and that of a rank sum of 48 or smaller is 0.13. We would be unlikely to reject the null hypothesis on this evidence.

The `exactRankTests` package in the statistical programming language *R* (www.R-project.org) includes a function `ansari.exact` to carry out the test and, optionally, to construct a confidence interval for γ . The algorithm used in that function yields a P value of 0.14 for these data.

Summary

The validity of the Ansari–Bradley test [1] requires that the two population medians be known or that they be known to be equal. The results of the test can be misleading if, as is most often the case, the populations differ in location by an unknown amount. An alternative test based on the squared ranks of the absolute deviations of observations about their (estimated) population means does not require equivalence or knowledge of population medians and is described in [2]. This test, the *Conover test*, is included in the *StatXact* program (www.cytel.com) and see **Exact Methods for Categorical Data**.

Table 1 Ansari–Bradley illustration

	1	2	3	4	5	6	7	8	9	10	Sum
x : 2001 sample	82	91	91	92	94	94	102	102	103	106	
y : 2003 sample	86	89	90	94	94	97	104	108	110	114	
x : Ranks	1	5.5	5.5	7	9.25	9.25	7.5	7.5	6	4	62.5
y : Ranks	2	3	4	9.25	9.25	9	5	3	2	1	47.5

2 Ansari–Bradley Test

References

- [1] Ansari, A.R. & Bradley, R.A. (1960). Rank-sum tests for dispersion, *Annals of Mathematical Statistics* **31**, 1174–1189.
- [2] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [3] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.

CLIFFORD E. LUNNEBORG

Arbuthnot, John

ROBERT B. FAUX

Volume 1, pp. 94–95

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Arbuthnot, John

Born: April 29, 1667, in Kincardineshire, Scotland.

Died: February 27, 1735, in London, England.

John Arbuthnot was born in the parish of Arbuthnott, Kincardineshire, Scotland. His parents were Alexander Arbuthnott, an Anglican clergyman, and Margaret Lammy Arbuthnott. As you can see, the traditional spelling of the name included an additional 't'. Arbuthnot pursued his education in England and Scotland and began his writing career while studying in London. In 1696, he received his Doctor of Medicine degree from St Andrews. He was married, and fathered four children [1, 3].

Arbuthnot became well-known in London for his skill as a physician. In 1704, Arbuthnot was elected a Fellow of the Royal Society. In 1705, he was appointed Physician Extraordinary to Queen Anne and in 1709, became her Physician Ordinary. In 1710, Arbuthnot was elected a fellow of the Royal College of Physicians. However, it is for his satirical writings that Arbuthnot is best known. He counted among his friends Alexander Pope, Jonathan Swift, John Gay, and Thomas Parnell. In 1714, these gentlemen established the Scriblerus Club. The remit of the Club was to satirize pedantry and bad poetry. In addition to his literary endeavors, Arbuthnot, over a period of 40 years, wrote eight works of a scientific nature. Three of these works dealt with mathematics in some form [2, 3].

Mathematical Works

Laws of Chance

Arbuthnot's first book, entitled *Of the Laws of Chance*, was published anonymously in 1692; it was a translation of Christiaan Huygens's *De ratiociniis in ludo Aleae* [2, 3]. Huygens intentionally left two problems unsolved in his book, and in his translation, Arbuthnot offered solutions to the problems. His solutions were later replicated by James Bernoulli [2]. However, the primary intent of the book was to expose the general reader to the uses of probability in games of chance as well as in other endeavors such as politics (see **Probability: Foundations** of).

This book is often considered to be of less importance than the two subsequent books Arbuthnot was to devote to mathematical topics [2, 3].

Usefulness of Mathematics

In 1701, Arbuthnot published a second treatise on mathematics, again anonymously, entitled *An Essay on the Usefulness of Mathematical Learning*, subtitled *In a Letter from a Gentleman in the City to His Friend in Oxford* [3]. In this work, Arbuthnot reflected upon the intellectual and practical benefits of learning mathematics. As mathematics was neglected by most students of the time, who believed that it was too difficult a subject, Arbuthnot argued that mathematicians had provided ample simplifications and examples of calculations that would allow anyone to perform them. Arbuthnot ends his treatise by suggesting that the best ways to teach mathematics are by practical demonstrations and progressing from simple to complex problems [1].

Arguing for Divine Providence

In 1710, Arbuthnot published his third mathematical treatise. Entitled *An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes*, this treatise represents Arbuthnot's efforts to demonstrate the usefulness of mathematics as well as the existence of God [2, 3]. Using probability formulas, Arbuthnot argued that the existence of God was revealed by the prevalence of male births found in the official statistical summaries of births and deaths. Such dominance was not due to chance but was a reflection of God's desire to ensure the continuation of the human race. Given that more males than females were likely to suffer from disease and death, such dominance would ensure that there would be a sufficient number of males who would marry and father children.

These three works did not significantly contribute to the development of statistical analysis but, rather, reflect Arbuthnot's ever-searching mind [2, 3].

References

- [1] Aitken, G.A. (1892). *The Life and Works of John Arbuthnot*, Clarendon Press, London.
- [2] Beattie, L.M. (1935). *John Arbuthnot: Mathematician and satirist*, Harvard University Press, Cambridge.
- [3] Steensma, R.C. (1979). *Dr. John Arbuthnot*, Twayne, Boston.

ROBERT B. FAUX

Area Sampling

JERRY WELKENHUYSEN-GYBELS AND DIRK HEERWEGH

Volume 1, pp. 95–96

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Area Sampling

In an area sample, the primary sampling units are well-defined fractions of the earth's surface [2]. The sampling frame, also referred to as the area frame, is in fact a map that has been subdivided into a certain number of mutually exclusive and exhaustive areas. Of course, the actual sampling frame need not be in the form of a map. In practice, the sampling frame will usually consist of a list of the areas from which the sample is to be drawn.

Any partition of the area frame into area segments will yield unbiased estimates of the population parameters of interest. However, the efficiency of the estimators can differ greatly between partitions. More specifically, the design of an efficient area sample requires that areas are as equal in size as possible [2].

Area samples can be used for several purposes. In agricultural and forestry research, they are often employed to study the characteristics of the land covered by the sampled area. In this context, area samples are, for example, used to study the number of acres in certain crops, the number of acres covered by forests, or the number of acres under urban development. Area samples are also used in behavioral research [3] when no complete or up-to-date sampling frame is available for the observational units (households, individuals, businesses, etc.) that are of actual interest. In such a case, information is collected from all observational units within the sampled areas. Area samples have, for example, been used for this purpose in the context of the United States census as well as in other countries that do not maintain current lists of residents. However, even if current population registers are available, area sampling is sometimes used in the first stage of two-stage sampling procedure for practical reasons. For example, the Belgian General Election studies [1] use a two-stage sample from the population to collect information about political and related attitudes. In the

first stage, a random sample of villages is drawn, and then within the sampled villages, random samples of individuals are drawn. The latter sampling design is used to reduce the interviewer workload by concentrating respondents in certain areas and hence reducing travel distances.

A very important issue with respect to the use of area sampling is the accurate definition of the primary sampling units, that is the areas, as well as the development of a set of rules that associates the elements of the population under study with the areas. In behavioral research, administrative boundaries (villages, counties, etc.) are often useful in determining the boundaries of the primary sampling units, provided the size of these administrative areas is suitable for the purpose of the research. Individuals and households can be associated with the area segments through their primary residence. The latter is, however, not always straightforward because some individuals or households might have multiple residences.

References

- [1] Carton, A., Swyngedouw, M., Billiet, J. & Beerten, R. (1993). *Source Book of the Voters Study in Connection with the 1991 General Election*, Department sociologie/Sociologisch Onderzoeksinstituut K.U.Leuven, Interuniversitair Steunpunt Politieke-Opinie Onderzoek, Leuven.
- [2] Fuller, W.A. (1989). Area sampling, in *Encyclopedia of Statistical Sciences*, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 397–402.
- [3] Kish, L. (1965). *Survey sampling*, John Wiley & Sons, New York.

(See also **Survey Sampling Procedures**)

JERRY WELKENHUYSEN-GYBELS AND
DIRK HEERWEGH

Arithmetic Mean

DAVID CLARK-CARTER

Volume 1, pp. 96–97

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Arithmetic Mean

The arithmetic mean, usually referred to as the mean, is the most commonly used measure of central tendency or location. It is often denoted by $\bar{X}$ or M (the latter symbol is recommended by the American Psychological Association [1]). The mean is defined as the sum of a set of scores divided by the number of scores, that is, for a set of values, $X_1, X_2, \dots, X_n$ ($i = 1, 2, \dots, n$),

$$\bar{X} = \frac{\sum_i X_i}{n}. \quad (1)$$

Thus, for example, the arithmetic mean of 15 and 10 is $25/2 = 12.5$.

One advantage that the mean has over some other measures of central tendency is that all the members of the set of scores contribute equally to its calculation. This can also be a drawback, however, in that the mean can be seriously affected by extreme scores. As a simple illustration, consider the five values 5, 10, 15, 20, and 25, for which both the mean and median are 15. When a much larger value is added to the set, say 129, the mean becomes 34 while the median is 17.5. If, instead of

129 the new number had been 27, then the mean would have been 17, though the median would still be 17.5. However, the problem of outliers can be ameliorated by more robust versions of the mean such as **trimmed means** and Winsorized means (see **Winsorized Robust Measures**) [2], or by using the **harmonic** or **geometric means**.

Another disadvantage of the arithmetic mean is that, with discrete numbers, it may not appear to be a sensible figure given the nature of the variable of interest – for example, if the mean number of children in a class comes out at 35.7.

Despite these shortcomings, the mean has desirable properties, which account for its widespread use for describing and comparing samples both as a summary measure and in traditional inferential methods, such as t Tests.

References

- [1] American Psychological Association. (2001). *Publication Manual of the American Psychological Association*, 5th Edition, American Psychological Association, Washington.
- [2] Wilcox, R.R. (1996). *Statistics for the Social Sciences*, Academic Press, San Diego.

DAVID CLARK-CARTER

Ascertainment Corrections

MICHAEL C. NEALE

Volume 1, pp. 97–99

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ascertainment Corrections

Statistical Theory

The aim of genetic epidemiologists is to characterize the population of interest, which is straightforward when data from a random sample of the population are available. However, by either accident or design, it may be that a random sample is not available. Two examples will clarify these situations. First, suppose an investigator attempts to obtain magnetic resonance imaging scans to assess body circumference. Subjects with very large circumference will not fit into the scanner, so their data will not be available, which is a case of data systematically missing by accident. Second, consider a twin study of depression in which the only pairs available are those in which at least one twin has major depressive disorder, because ascertainment has been through hospital registration. Here, the data are missing by design. Both cases may require correction for ascertainment. At first sight, these corrections may seem technically challenging, but for the most part they are based on simple principles, which involve correctly specifying the probability density function (pdf) for the sample as well as for the population. Sometimes it is necessary to provide estimates of population parameters in order to compute the appropriate correction.

Maximum likelihood estimation typically proceeds by estimating the parameters of the distribution to find the values that maximize the joint likelihood of the observed data points. For example, if it is assumed that a set of scores were drawn from a population with a normal distribution, the mean μ and the variance σ^2 may be estimated. Of primary importance here is that the normal distribution is a pdf, and that the sum over all possible observed points, which we write as the integral from $-\infty$ to $+\infty$, equals unity, that is,

$$\int_{-\infty}^{\infty} \phi(x) dx = 1, \quad (1)$$

where

$$\phi(x) = \frac{e^{-\frac{(x-\mu)^2}{\sigma^2}}}{\sqrt{2\pi}}. \quad (2)$$

In the event that only persons with a score of t or above are available (perhaps due to preselection of

the sample or some malfunction of equipment such that those with scores below t had their data deleted), then the sample mean and variance will be biased estimates of the population statistics. Similarly, maximum likelihood estimates of the population mean and variance obtained by maximizing the likelihood will be biased in the absence of correction. Essentially, the likelihood must be renormalized so that the probability of all possible outcomes (the integral of possible scores from $-\infty$ to $+\infty$) equals unity. For a sample with scores restricted to be above threshold t , the ascertainment correction equals the proportion of the distribution that remains, which is:

$$\int_t^{\infty} \phi(x) dx. \quad (3)$$

so the likelihood of an observed score $x > t$ becomes:

$$\frac{\phi(x)}{\int_t^{\infty} \phi(x) dx}. \quad (4)$$

An ascertainment correction of this type can be made easily within a program such as Mx [3] using its `\mnor` function in the formula for the weight. Note that knowledge of the threshold for ascertainment, t , is required to define the correction (*see Software for Behavioral Genetics*).

Three complications to the above situation are regularly encountered in **twin studies** and family studies (*see Family History Versus Family Study Methods in Genetics*). First, selection of families may be based on either one or both members of a pair of relatives. Second, the analysis of binary or ordinal data requires somewhat special treatment. Third, the variable that is used to select the twins is sometimes not the one that is the focus of the analysis.

Consider the analysis of data collected from pairs of twins, under the assumption that the distribution of the pairs of scores is bivariate normal in the population. In the absence of ascertainment, the likelihood of a pair of scores is simply:

$$\phi(x_1, x_2), \quad (5)$$

where $\phi(x_1, x_2)$ is the multivariate normal probability density function (*see Catalogue of Probability Density Functions*) given by

$$|2\pi \Sigma|^{-2m/2} \exp \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_i)' \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_i) \right], \quad (6)$$

2 Ascertainment Corrections

where m is the number of variables, Σ is their population covariance matrix; μ_i is their (column) vector of population means; and $|\Sigma|$ and Σ^{-1} respectively denote the determinant and inverse of the matrix Σ . If pairs are ascertained such that twin 1's score is above threshold, $x_1 > t$, then the ascertainment correction is identical to that in 3 above. When pairs are ascertained if and only if they are both above threshold t , then the ascertainment correction is given by the double integral:

$$\int_t^\infty \int_t^\infty \phi(x_1, x_2) dx_2 dx_1. \quad (7)$$

Revision of these formulas to cases in which twins are selected for being discordant, for example, where twin 1 is above threshold t and twin 2 is below threshold $-u$, would be achieved by changing the limits of the integrals. Of particular note in situations of this sort is that the population threshold t must be estimated from other sources, since the estimate of the correlation between relatives depends heavily on this value.

Correction for the third complication mentioned above, where the variable of interest is not the same as the variable used for ascertainment, is also straightforward. The ascertainment correction remains unchanged, but the likelihood must be written as the joint likelihood of the ascertainment variable and the variable of interest, corrected for ascertainment. If the variable of interest is denoted by y_1 and y_2 for twin 1 and twin 2, and the variable on which ascertainment is based is x_1 in twin 1 and x_2 in twin 2, then the likelihood may be written as

$$\frac{\phi(x_1, x_2, y_1, y_2)}{\int_t^\infty \int_t^\infty \phi(x_1, x_2) dx_2 dx_1} \quad (8)$$

when ascertainment is for pairs concordant for being above threshold t . The most important thing to note in this case is that the correct expression for the likelihood involves the joint likelihood of both the ascertainment variable and the variable of interest. If the ascertainment variables x_1 and x_2 are both independent of both y_1 and y_2 , it would not be necessary to correct for ascertainment, and uncorrected univariate analysis of y_1 and y_2 would yield the same results.

Binary Data

A popular approach to the analysis of binary data collected from pairs of relatives is to assume that there is an underlying bivariate normal distribution of liability in the population (*see Liability Threshold Models*). The binary observed variable arises as a sort of measurement artifact in that those above threshold t get a score of 1 and those below get a score of 0. The aim of the approach is to estimate familial resemblance as a **tetrachoric correlation**, which is the correlation between the relatives' underlying liability distributions. Here we can classify sampling scenarios by reference to the probability that an individual is ascertained given that they are affected. This probability, often referred to as π [2], identifies two special situations at its extremes. Any study that has obtained at least one pair cannot have an ascertainment probability $\pi = 0$, but we can use the limit as $\pi \rightarrow 0$ for cases in which it is very unlikely that a subject is ascertained. In this case, it would be extremely unlikely to ascertain both members of the pair, and the scheme is known as *single ascertainment*. The appropriate correction for ascertainment here is simply the probability that one member of the pair is ascertained, that is, (3). This situation might be encountered if patients were ascertained through a clinic setting where it is very unlikely that their relative attends the same clinic, and the relatives are obtained for study purely through information provided by the patient.

At the other extreme, the probability of ascertainment given affected status is unity, which is known as *complete ascertainment*. One circumstance where this may be encountered is where twin pairs are ascertained through records of all hospitals in a country, and where affected status inevitably leads to hospitalization. In this case, the only pairs that are missing from the sample are those in which neither relative is affected. The ascertainment correction would therefore be

$$1 - \int_{-\infty}^t \int_{-\infty}^t \phi(x_1, x_2) dx_2 dx_1. \quad (9)$$

which is equal to the sum of the probabilities of observing the three remaining outcomes for a relative pair:

$$\int_{-\infty}^t \int_t^\infty \phi(x_1, x_2) dx_2 dx_1$$

$$\begin{aligned}
 & + \int_t^\infty \int_{-\infty}^t \phi(x_1, x_2) dx_2 dx_1 \\
 & + \int_t^\infty \int_t^\infty \phi(x_1, x_2) dx_2 dx_1. \quad (10)
 \end{aligned}$$

The situation in which sampling is at neither of these extremes, $0 < \pi < 1$, is known as *incomplete ascertainment* [1]. Treatment of this scenario is more complicated because it depends on whether the pair has been ascertained through one or both members of the pair. For singly ascertained pairs it is

$$\frac{\pi}{2\pi \int_{-\infty}^t \phi(x_1) dx_1 - \pi^2 \int_t^\infty \int_t^\infty \phi(x_1, x_2) dx_2 dx_1} \quad (11)$$

and for pairs in which both members were ascertained it is

$$\frac{\pi(2 - \pi)}{2\pi \int_{-\infty}^t \phi(x_1) dx_1 - \pi^2 \int_t^\infty \int_t^\infty \phi(x_1, x_2) dx_2 dx_1}. \quad (12)$$

An application using all of these corrections was described by Sullivan et al. in their **meta-analysis** of twin studies of schizophrenia [4].

Acknowledgment

Michael C. Neale is primarily supported from PHS grants MH-65322 and DA018673.

References

- [1] Allen, G. & Hrubec, Z. (1979). Twin concordance a more general model, *Acta Geneticae Medicae et Gemellologiae (Roma)* **28**(1), 3–13.
- [2] Morton, N.E. (1982). *Outline of Genetic Epidemiology*, Karger, New York.
- [3] Neale, M., Boker, S., Xie, G. & Maes, H. (2003). *Mx: Statistical Modeling*, 6th Edition, Box 980126, Department of Psychiatry, Virginia Commonwealth University, Richmond.
- [4] Sullivan, P.F., Kendler, K.S. & Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies, *Archives of general psychiatry* **60**(12), 1187–1192.

(See also **Missing Data**)

MICHAEL C. NEALE

Assortative Mating

SCOTT L. HERSHBERGER

Volume 1, pp. 100–102

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Assortative Mating

Assortative mating is a departure from random mating in which like individuals preferentially mate with each other (positive assortative mating, or homogamy) or unlike individuals preferentially mate with each other (negative assortative mating, or disassortative mating). Cross-characteristic assortative mating is also possible, wherein individuals having a certain level of phenotype (observable characteristic) A mate with others having a certain level of phenotype B.

Since the early twentieth century, the degree of assortative mating has been expressed by the correlation r between the phenotypic values of mated individuals (*see Correlation Issues in Genetics Research*). In the earliest example of positive assortative mating, Pearson and Lee [5] calculated a ‘significant’ r of 0.093 ± 0.047 for stature on the basis of Table 1 provided by Galton [2]:

In addition, as shown in Table 2, Pearson and Lee [5] found in their own data positive direct and cross-assortative mating for stature, arm span, and forearm length:

Table 1 Heights of husbands and wives

		Husband			Totals
		Short	Medium	Tall	
Wife	Short	9	28	14	51
	Medium	25	51	28	104
	Tall	12	20	18	50
		46	99	60	205

Table 2 Assortative mating. Data based on 1000 to 1050 cases of husband and wife

	Husband’s character	Wife’s character	Correlation and probable error
Direct	Stature	Stature	0.2804 ± 0.0189
	Span	Span	0.1989 ± 0.0204
	Forearm	Forearm	0.1977 ± 0.0205
Cross	Stature	Span	0.1820 ± 0.0201
	Stature	Forearm	0.1403 ± 0.0204
	Span	Stature	0.2033 ± 0.0199
	Span	Forearm	0.1533 ± 0.0203
	Forearm	Stature	0.1784 ± 0.0201
	Forearm	Span	0.1545 ± 0.0203

In addition to the present-day research that continues to provide evidence for positive assortative mating for anthropometric characteristics, considerable evidence is also available for race, socioeconomic status, age, intellectual ability, education, physical attractiveness, occupation, and to a lesser extent, for personality and attitude variables [3]. Among anthropometric characteristics, mate correlations for stature are generally the highest; among intellectual ability characteristics, verbal ability, and among personality variables, extraversion.

It is worth noting that the motivation for measuring assortative mating by Galton, **Pearson**, and their contemporaries was to assess its effect on a characteristic’s phenotypic variance in the population. The question of greatest interest was what would happen to the phenotypic variance of a characteristic, and other characteristics ‘incidentally’ correlated with it, if mate selection occurred on the basis of the characteristic’s perceived value. The general finding of this research was that even with relatively small correlations between mates, the characteristic’s phenotypic variance would increase dramatically among offspring if the correlation was positive and decrease dramatically if the correlation was negative. This was even true for the variance of characteristics moderately associated with the selected characteristic [4].

Although the degree of assortative mating r between the phenotypic values of mated individuals is observable, the genetic consequences depend on the correlation m between the **Genotype** of the mates. To determine the association between r and m , what governs the choice of mates must be known. Mate choice can be based on phenotypic similarity (phenotypic homogamy) or environmental similarity (social homogamy), or both. Phenotypic similarity means that the mates are selected on the basis of their phenotypic values for a characteristic. When mate choice is based on the phenotype, the effect of positive assortative mating is to increase population **additive genetic variance (heritability)**; the effect of negative assortative mating is to decrease it. For example, positive assortative mating’s effect of increasing a characteristic’s heritability is illustrated in Figure 1. In this figure, m is the genetic correlation between mates, b is the path between a parent’s genotype and the gametes produced by the parent, and a is the path between the gametes and the genotype of an offspring. Under conditions of random mating, $m = 0$, and the correlation between either parent and

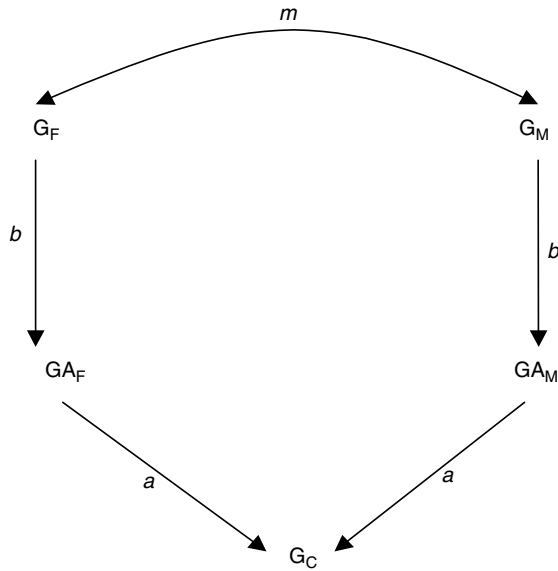


Figure 1 Correlation between family members under assortative mating; G_F = father genotype; G_M = mother genotype; G_C = child genotype; G_{A_F} = father gametes; G_{A_M} = mother gametes; m = genetic correlation between father and mother; b = the path from father or mother genotype to father or mother gametes; and a = the path from father or mother gametes to child genotype

offspring is $a \times b$. When m is greater than zero, the genetic correlation between a single parent and child, r_{op} , increases by a factor of $(1 + m)$:

$$\begin{aligned} r_{op} &= (ab + abm) \\ &= ab(1 + m). \end{aligned} \quad (1)$$

Thus, the genetic correlation between parent and offspring increases whenever m , the genetic correlation between mates, is nonzero. Further details can be found in [1].

On the other hand, if assortative mating occurs *only* through environmental similarity, such as social class, $m = 0$ and there is effect on the characteristic's heritability. However, in modern times when marriages are less likely to be arranged, it is extremely unlikely that selection is based solely on characteristics of the individuals' environment, without consideration given to phenotypic characteristics. In addition, heritability will still increase if differences on the environmental characteristic are in part due to genetic differences among individuals. For example,

the significant heritability of social class would argue for assortative mating arising from both phenotypic and social homogamy.

Determining the impact of assortative mating on genetic variance is critical when attempting to estimate the heritability of a characteristic. When human twins are used to estimate heritability, assortative mating is assumed to be negligible; if this assumption is not made, assortative mating must be explicitly factored into the estimation of the heritability. Two types of twins exist. Monozygotic (identical) twins result from the splitting of a single, fertilized ovum, and are genetically identical; dizygotic (fraternal) twins result from the fertilization of two separate ova, and are no more genetically alike than full siblings. If heredity influences a characteristic, identical twins (who have identical genotypes) should be more alike than fraternal twins (who share on the average 50% of their genes). If the identical twins are no more alike than fraternal twins, the heritability of the characteristic is essentially zero. If assortative mating is assumed to be negligible, heritability is calculated by subtracting the fraternal twin correlation from the identical twin correlation and doubling the difference. If this assumption is incorrect, the identical twin correlation is unaffected, but the fraternal twin correlation is increased. The result will be that the correlation between fraternal twins will be closer to the correlation between identical twins, thereby spuriously lowering the heritability estimate.

References

- [1] Cavalli-Sforza, L.L. & Bodmer, W.F. (1999). *The Genetics of Human Populations*, Dover, Mineola.
- [2] Galton, F. (1886). Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute* **15**, 246–263.
- [3] Mascie-Taylor, C.G.N. (1995). Human assortative mating: evidence and genetic implications, in *Human populations. Diversity and Adaptations*, A.J. Boyce & V. Reynolds, eds, Oxford University Press, Oxford, pp. 86–105.
- [4] Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs, *Philosophical Transactions of the Royal Society A* **200**, 1–66.
- [5] Pearson, K. & Lee, A. (1903). In the laws of inheritance in man. I. Inheritance of physical characters, *Biometrika* **2**, 357–462.

SCOTT L. HERSHBERGER

Asymptotic Relative Efficiency

CLIFFORD E. LUNNEBORG

Volume 1, pp. 102–102

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Asymptotic Relative Efficiency

When deciding among estimators or test procedures, one naturally will prefer the estimator or test procedure that requires the least number of observations to ensure it is sufficiently close to the true population value (*see Estimation*). The *asymptotic relative efficiency* (ARE) or Pitman efficiency of an estimation or testing procedure is defined as ‘the limit with increasing sample size of the ratio of the number of observations required for each of two consistent statistical procedures to achieve the same degree of accuracy’ [1, p. 42].

For example, when the observations are drawn from a Normal distribution, the **Hodges-Lehmann estimator** of central location has an ARE of 0.96 relative to the arithmetic mean [2, p. 246]. In the same circumstances, the sign test has an ARE with respect to the t Test of only 0.67. In other words, the sign test requires almost ‘50% more observations than the t Test to achieve the same power’ [2, p. 176].

The ARE and, thus, the appropriate test or estimator strongly depends upon the distribution from which the observations are drawn. For example, the ARE of

the Wilcoxon test with respect to the t Test is 0.95 when the observations are drawn from a Normal distribution, but with distributions that are more heavily weighted in the tails than the Normal, the **Wilcoxon test** could have an ARE with respect to the t of 10 or more [2, p. 176].

It is important, therefore, to have a good idea of the kind of distribution sampled. Further, the ARE summarizes large sample results and may not be a trustworthy guide when sample sizes are only small to moderate. Consequently, albeit ARE is an important factor, it should not be the only one to be considered in selecting an estimator or test procedure.

Acknowledgments

The author gratefully acknowledges the assistance of Phillip Good in the preparation of this article.

References

- [1] Good, P. & Hardin, J. (2003). *Common Errors in Statistics*, Wiley-Interscience, Hoboken.
- [2] Lehmann, E.L. (1999). *Elements of Large Sample Theory*, Springer, New York.

CLIFFORD E. LUNNEBORG

Attitude Scaling

MATTHEW S. JOHNSON AND BRIAN W. JUNKER

Volume 1, pp. 102–110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Attitude Scaling

Introduction

A large part of social psychology focuses on the study of attitudes. As [1] said, ‘The concept of attitude is probably the most distinctive and indispensable concept in contemporary American social psychology’. Although the statement was made in the 1930s, the statement rings true today. Attitude research covers a wide variety of topics, such as the study of job performance or absenteeism in factories and how they relate to the attitudes of the factory workers. Political studies often attempt to explain political behaviors such as voting by the attitude of the voting individual. Marketing researchers often wish to study consumers’ attitudes to predict their behavior [15].

Attitudes are simply one of many hypothetical constructs used in the social sciences. They are not observable, and hence are not easily quantified. If attitudes were easily quantified, studying the relationships between the attitudes and dependent variables of interest becomes simple.

In order to scale attitudes, researchers assume attitudes are a measurable quantity. Thurstone’s 1928 paper ‘Attitudes can be measured’ claims just that. In order to do so, the range of attitudes is generally assumed to lie on a single bipolar continuum, or scale. This underlying scale can represent attitudes such as the *liberal-conservative* scale if politics is being studied, or the *risk willingness* scale if investment strategies are the subject of interest. The number of attitudes one might study is infinite. Some of the more popular studied attitudes include social or political attitudes, prejudices, interpersonal attraction, self-esteem, and personal values [14].

Some recent attitudinal studies using attitude scaling methodologies include the following examples:

- *Attitude measurement applied to smoking cessation.* Noël [35] hypothesizes the existence of a unidimensional latent construct representing the level of change a smoker who is attempting to quit has reached. Noël calls this construct the *change maturity index* of the respondent. The author uses a set of 40 statements representing various levels of maturity of cessation to measure the index. ‘Warnings about health hazards of smoking move me emotionally’, and ‘Instead

of smoking I engage in some physical activity’ are two sample statements from the survey given to the smokers in this study.

- *The inventory of interpersonal problems.* Kim and Pilkonis [26] note that one of the best indicators of personality disorders is chronic difficulties in interpersonal relationships. The Inventory of Interpersonal Problems [IIP; 21] is one instrument suggested for studying interpersonal problems. The IIP consists of 127 statements derived from admission interviews with psychiatric patients. Kim and Pilkonis [26] hypothesize these statements measure five latent attitudes: *interpersonal sensitivity*; *interpersonal ambivalence*; *aggression*; *need for social approval*, and *lack of sociability*.
- *Political agency.* Muhlberger [34] studies a latent construct called *political agency* by constructing a set of eight opinion statements that tap varying levels of political agency. The low end of the political agency construct represents the attitudes of individuals following politics for external reasons [others expect them to follow]. The upper end corresponds to ‘intrinsically valuable’ reasons for following politics.

Assuming the attitude of interest lies on a unidimensional, bipolar scale, attitude scaling techniques estimate the position of individuals on that scale. Scaling is achieved by developing or selecting a number of stimuli, or items, which measure varying levels of the attitude being studied. Thurstone [48] advocates the use of opinions, the verbal expressions of attitude. These items, or opinions, are then presented to the respondents and information is collected using some predetermined mode. Researchers use this information to make inferences about the respondents’ attitudes.

Thurstone Scaling

One of the earliest data collection methods for scaling attitudes is **Thurstone’s Law of Comparative Judgment** [LCJ; 47]. The LCJ, which has its roots in psychophysical scaling, measures the items in the same way that psychophysical techniques scale stimuli such as tones on a physiological scale such as loudness. However, this only locates the items on the scale, and a second stage of measurement is needed to measure respondents.

2 Attitude Scaling

Thurstone's method begins by developing a large number of items, or statements, that represent particular attitudes toward the topic of study (e.g., politics; education). Once the set of items is constructed, the items are presented to a group of judges who sort the items. After the items have been scaled from the sorting data, they can be administered to new respondents in a second, separate data collection stage.

Stage One: Locating Items

Method of Paired Comparisons. In the method of paired comparison each of J survey items is paired with each of the other items for a total of $\binom{J}{2}$ item pairs and judges are asked which of the two items in each of the $\binom{J}{2}$ item pairs is located higher on the attitude scale. For example, scaling 10 former US Presidents on the liberalism-conservatism scale would be accomplished by asking the judges to examine each of the $\binom{10}{2} = 45$ pairs of presidents and determine which was more conservative. For example judges might be asked: 'Which US President had the more conservative social policy: Reagan or Clinton'? The judges' attitudes should not affect their responses in any way.

The assumption central to Thurstone's theory is that each survey item produces a psychological sensation in each judge. The sensation is assumed to follow a normal distribution with mean located at the location of the item on the attitude scale. If the psychological sensation produced by item j is larger than the sensation produced by item k for a particular judge, that judge would determine item j is located above item k .

Under Thurstone's Case V Model, for example, the probability that a judge rates item j above item k is represented by $\Phi(\beta_j - \beta_k)$, where Φ is the standard normal cumulative distribution function, and β_j and β_k are the locations of items j and k respectively. Therefore, by using the $J \times J$ matrix, where element (j, k) is the proportion of judges rating item j above item k , the locations of the items are estimated. Mosteller [33] develops a goodness-of-fit statistic for this scaling method.

A similar method for locating items was suggested by [9] and [29]. Their treatment of the paired comparison problem assumes that the probability of choosing item j over item k follows $\Psi(\beta_j - \beta_k)$ where Ψ is the *expit*, or inverse logit function

$(\exp(\cdot)/1 + \exp(\cdot))$. This model for paired comparisons is called the **Bradley-Terry model**. A review of the extensions of paired comparisons models is presented in [8].

The drawback of the paired comparisons item scaling method is that it requires a huge number of comparisons. For example locating the last forty presidents on the liberalism-conservatism scale would require each judge to make $\binom{40}{2} = 780$ paired comparisons.

Method of Equal-appearing Intervals. Because the method of paired comparisons requires so many judgments, Thurstone [49] developed the method of equal-appearing intervals. In this method the judges are asked to separate the items into some fixed number of rating intervals according to where the judge believes the items are located on the latent scale. Assuming the rating intervals are of equal width, the intervals are assigned consecutive scores (e.g., 1–11), and the scale value assigned to each survey item is estimated by the mean or median score received by the item.

Method of Successive Intervals. Thurstone considered the method of equal-appearing intervals as a way to approximate the method of paired comparisons. Realizing that the method of equal-appearing intervals and the method of paired comparisons did not yield perfectly linear results, the method of successive intervals was developed. This method also asks judges to sort the items into some number of interval categories. However, the intervals are not assumed to be of equal width. Once the judgment data is collected, interval widths and item locations are estimated from the data.

Stage Two: Locating Respondents

After the survey items have been located on the latent scale according to one of the three procedures discussed above, the items are used to measure the attitudes of the respondents. The procedure is quite simple. A subset of items (from the item pool measured by the judges) is chosen so that the items are more or less uniformly distributed across the latent scale. Respondents receive the items, one at a time, and are asked whether or not they endorse (e.g., like/dislike) each item. Thurstone's

scaling method assumes that respondents endorse only those items that are located near the respondent; this assumption implies a *unimodal* response function. The method then estimates the location of the respondent with the *Thurstone estimator* which is the average or median location of the endorsed items.

To scale respondents on a social liberalism-conservatism attitude scale, a political scientist might ask the respondents: ‘For each of the following Presidents please mark whether you agreed with the President’s social policy (1 = Agree, 0 = Disagree)’. A respondent might disagree because they feel the President’s social policy was too liberal, or because the policy was too conservative. The scale location of a respondent who agreed with the policy of Clinton and Carter, but disagreed with the remaining Presidents, would be estimated by $\tilde{\theta}_1 = (72.0 + 67.0)/2 = 69.5$. Similarly, the scale position of a second respondent who agreed with the policies of Carter, Ford, and Bush would be estimated as $\tilde{\theta}_2 = (67.0 + 39.3 + 32.8)/3 = 46.4$. The first respondent is more liberal than the second respondent.

Unfolding Response Models

Coombs [11] describes a procedure called *unfolding* which simultaneously scales items and subjects on the same linear scale using only Thurstone’s second stage of data collection. The method assumes there exists some range of attitudes around each item such that all respondents within that range necessarily endorse that item; outside of that range the respondents necessarily disagree with that item (see **Multidimensional Unfolding**).

Coombs’s model is an example of an unfolding response model. Unfolding response models assume that the *item response function*, the probability a respondent located at θ_i endorses an item located at β_j , is a unimodal function which reaches a maximum at β_j . Coombs’s model assumes a deterministic response function:

$$P_j(\theta_i) \equiv P\{X_{ij} = 1|\theta_i\} = \begin{cases} 1 & \text{if } \theta \in (\beta_j - \delta_j, \beta_j + \delta_j) \\ 0 & \text{if } \theta \notin (\beta_j - \delta_j, \beta_j + \delta_j) \end{cases} \quad (1)$$

The parameter δ_j is called the *latitude of acceptance* for item j .

By being able to locate items and respondents simultaneously, it became unnecessary to pre-measure the opinions as required by Thurstone’s scaling method. However, Coombs’s deterministic model is quite restrictive because it does not allow any response pattern to contain the triplet $\{1, 0, 1\}$. For example, if former US Presidents Clinton, Carter, Ford, Bush, and Reagan are ordered from most liberal to most conservative, then a person who endorses the social policy of Presidents Ford and Reagan, but not Bush, violates the model.

Since Coombs’s deterministic unfolding response model, a number of probabilistic models have been developed [e.g., 3, 12, 20, 30]. One of the earliest probabilistic unfolding response models used for scaling in attitude studies is the squared logistic model [3]. The model assumes the logit of the item response function is quadratic in the respondent’s location on the attitude scale. Specifically the model assumes $P_j(\theta) = \Psi(-(\theta - \beta_j)^2)$, which reaches a maximum value of $1/2$ when $\theta = \beta_j$.

The squared logistic model is too restrictive for many attitudinal surveys because the maximal endorsement probability is fixed at $1/2$. The PAR-ELLA [20] and the hyperbolic cosine models [4, 51] overcome this limitation by adding another parameter similar to the latitude of acceptance in Coombs’s model [30]. Unfolding response models for polytomous response scales have also been utilized in attitudinal studies [41]. Noël [35], for example, utilizes a polytomous unfolding response model to scale the attitudes of smokers as they approach cessation.

The locations of items and respondents are estimated using one of a number of estimation algorithms for unfolding responses models. These include joint **maximum likelihood** [3, 4], marginal maximum likelihood [20, 51] and **Markov Chain Monte Carlo** techniques [22, 24]. Post [37] develops a non-parametric definition of the unfolding response model which allows for the consistent estimation of the rank order of item locations along the attitude scale [22]. Assuming Post’s nonparametric unfolding model and that the rank order of item locations is known, [22] shows that the Thurstone estimator (i.e., the average location of the endorsed items) does in fact consistently estimate respondents by their attitudes.

Guttman Scaling

Guttman [18] suggests another method for the scaling of respondents. The main difference between Thurstone's scaling method and Guttman's is in the type of questions used to scale the respondents' attitudes. Guttman's key assumption is that individuals necessarily agree with all items located below their own position on the attitude scale and necessarily disagree with all items above.

Unlike Thurstone's scaling assumption, which implies a unimodal response function, Guttman's assumption implies a monotone response function. In fact, the response function can be parameterized as the step function:

$$P_j(\theta) = \begin{cases} 1 & \text{if } \theta > \beta_j \\ 0 & \text{if } \theta \leq \beta_j. \end{cases} \quad (2)$$

So questions that are valid for Thurstone scaling are not for Guttman scaling. For example a question valid for Thurstone scaling, such as, 'Did you agree with President Reagan's social policy'? might be altered to ask, 'Do you feel that President Reagan's social policy was too conservative?' to use in Guttman scaling.

Once a large number of items have been developed judges are utilized to sort the items. The researcher then performs a *scalogram analysis* to select the set of items that are most likely to conform to Guttman's deterministic assumption.

Guttman's deterministic assumption restricts the number of possible response patterns. If all J items in an attitudinal survey are Guttman items, then at most $J + 1$ response patterns should be observed. These $J + 1$ response patterns rank order survey respondents along the attitude scale. Respondents answering 'No' to all items (00...0) are positioned below respondents who endorse the lowest item (10...0), who lie below respondents who endorse the two lowest items (110...0), and so forth.

Goodman's Partially Ordered Items

Goodman [17] calls a set of strictly ordered items, as assumed in Guttman scaling, a *uniform scale* because there is only one order of items. However in some cases, it is not plausible to assume that items are strictly ordered. Assume that two orderings of past US Presidents from most conservative to most liberal are plausible,

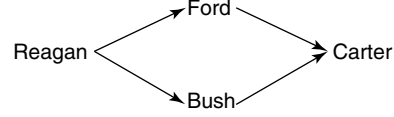


Figure 1 An example of a bifurm scale or belief poset

for example, $\text{Reagan} \rightarrow \text{Ford} \rightarrow \text{Bush} \rightarrow \text{Carter}$ and $\text{Reagan} \rightarrow \text{Bush} \rightarrow \text{Ford} \rightarrow \text{Carter}$. Because there are two plausible orderings of the items, Goodman calls the resulting scale a *bifurm scale*. Wiley and Martin [53] represent this *partially ordered set* of beliefs, or belief *poset* as in Figure 1.

Items are only partially ordered (Reagan is less liberal than both Ford and Bush, and Ford and Bush are less liberal than Carter), hence, unlike Guttman scaling, there is no longer a strict ordering of subjects by the response patterns. In particular, subjects with belief state 1100 (i.e., Reagan and Ford are too conservative) and 1010 (i.e., Reagan and Bush are too conservative) cannot be ordered with respect to one another.

In general, if J items make up a bifurm scale, then $J + 2$ response patterns are possible, as compared to $J + 1$ response patterns under Guttman scaling. Although the bifurm scale can be easily generalized to multifurm scales (e.g., trifurm scales), the limitations placed on the response patterns often prove to be restrictive for many applications.

Monotone Item Response Models

Probabilistic *item response theory* models, a class of generalized mixed effect models, surmount the restrictions of Guttman's and Goodman's deterministic models by adding a random component to the model. The **Rasch model** [39], one example of such a model, assumes the logit of the item response function is equal to the difference between the respondent's location and the item's location (i.e., $\log\{P_j(\theta)/(1 - P_j(\theta))\} = \theta - \beta_j$). The normal-ogive model assumes $P_j(\theta) = \Phi(\theta - \beta)$, where $\Phi(\cdot)$ is the normal cumulative distribution function [28]. The Rasch and normal-ogive models have also been generalized to applications where the latent attitude is assumed to be multidimensional ($\theta \in \mathbb{R}^k$) [6, 32, 40].

Respondents and items are located on the attitude scale using some estimation procedure for item response models. These include joint maximum

likelihood, conditional maximum likelihood [50], marginal maximum likelihood, or empirical Bayes [7, 46], and fully Bayesian methods using Markov chain Monte Carlo techniques [23, 36] (*see Markov Chain Monte Carlo and Bayesian Statistics*). With the recent studies of the nonparametric properties of monotone item response models [19, 25, 45], nonparametric estimation procedures have been suggested [13, 38]. In fact [45] shows under minor regularity conditions that the scale scores, found by summing the responses of an individual over all items, consistently estimate the rank order of respondents along the attitude scale.

Likert Scaling & Polytomous Response Functions

Likert scaling [27], like Thurstone and Guttman scaling, uses a panel of expert judges to locate the items on the attitude scale. However, Likert scaling uses a polytomous response scale (e.g., strongly disagree = 0; disagree = 1; neutral = 2; agree = 3; strongly agree = 4) rather than a dichotomous response scale (disagree = 0; agree = 1). Typically an odd number, usually five or seven, response categories are used, with a middle ‘neutral’ or ‘undecided category’; however, the use of an even number of response categories is equally valid.

The central assumption in Likert scaling is that the respondents located high on the attitude scale are more likely to use high response categories than are individuals located on the low end. If the widths of the response categories are constant across items, then the respondents can be rank ordered along the attitude scale by simply summing their responses across all items. Junker [25] describes conditions under which the Likert scale scores consistently estimate the rank order of respondents along the attitude scale.

Classical Likert scale scores can be thought of as a nonparametric procedure for rank ordering the respondents along the attitude scale. Item response theory models offer a parametric alternative. The Rasch and normal-ogive models have been generalized for use with polytomous item responses. The partial credit model [31] generalizes the Rasch model by assuming the adjacent category logits are equal to the difference between the location of the respondent and the location of the item-category on the attitude

scale:

$$P\{X_j = t | X_j \in \{t-1, t\}\} = \theta - \beta_{jt}. \quad (3)$$

The location of item j is the average of the item’s K item-category locations (i.e., $\beta_j = (1/K) \sum \beta_{jt}$). Johnson, Cohen, and Junker [23] use the PCM to study the attitudes of Research and Development directors towards a number of mechanisms for the appropriation of returns on the company’s innovations.

Ranking Methods

Coombs [10] develops an alternative to the procedures of Thurstone, Guttman, and Likert for the location of items and respondents that is based on respondents comparing items. The data collection method, often called an *unfolding* data method, asks the respondents to rank order items according to their preference. To use this procedure to locate past Presidents and respondents on the social liberalism-conservatism scale the researcher could ask respondents to rank order the Presidents according to how much the respondent agreed with the social policy of the Presidents.

Like the scaling methods discussed earlier, Coombs relies on the assumption that both respondents and survey items are located on the same single scale, and hypothesizes that respondents are more likely to endorse items located near their position on the latent scale. Hence subject i prefers item j to item k if and only if $|\theta_i - \beta_j| < |\theta_i - \beta_k|$, where θ_i is the location of individual i on the latent scale and β_j is the location of item j on the latent scale.

Consider the five items labeled A, B, C, D, and E, and a subject located at θ along the attitude scale (Coombs calls this the *J-scale* for *joint* scale) in Figure 2. Imagine a hinge at the location of the respondent’s attitude and *fold* the scale at that point. This folding results in the rank order of the items for that respondent. The subject located at θ would rank order the items (B, A, C, D, E). Coombs refers to the rank orders given by individuals as *I-scales*.

Coombs designed a method for estimating the rank orders of both the survey items and the survey respondents by assuming that no extraneous variation is present. The assumption of error-free data is a

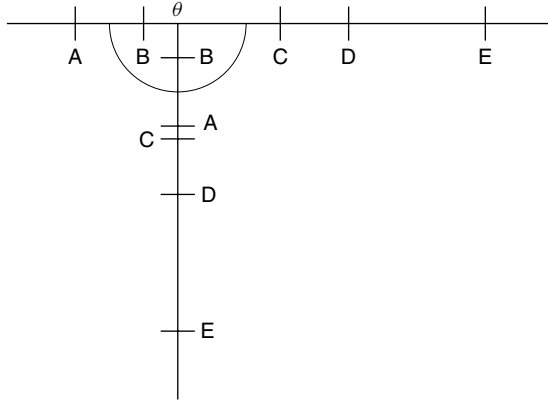


Figure 2 The relationship between items' and a subject's locations and the observed ranking of the items in Coombs' (1950) unfolding model

serious limitation, and it is not surprising that few real data sets can be analyzed using the method Coombs suggests.

A method that is closely related to Coombs's ranking method is the *pairwise preference* method. The method, like the paired comparison item scaling method, pairs each of the J survey items with the other $J - 1$ items, but unlike the paired comparison method the respondents' personal attitudes affect their responses. For each item pair the respondents are asked to select which of the pair of items they prefer. For example, 'Whose social policy did you prefer, Bush's or Reagan's'?

Assuming pairwise preference is error-free, a respondent whose attitude is located at θ_i on the latent scale prefers item j to item k whenever he or she is located below the midpoint of two items (assuming $\beta_j < \beta_k$). Coombs [11] describes a method to construct a complete rank ordering of all J items from the set of $\binom{J}{2}$ pairwise preference comparisons and estimates the rank order of respondents.

Bechtel [5] introduces a stochastic model for the analysis of pairwise preference data that assumes the location of respondents' attitudes are normally distributed along the attitude scale. Sixtl [44] generalizes this model to a model which assumes a general distribution $F(\theta)$ for the locations of respondents on the attitude scale. The probability that a randomly selected respondent prefers item j to item k closes $P(j \text{ preferred to } k) = \int_{-\infty}^{\infty} I\{t < (\beta_j + \beta_k)/2\} dF(t) = F(\beta_j + \beta_k)/2$. Let $\hat{P}(j \text{ to } k)$

denote the sample proportion of respondents who prefer item j to item k . Sixtl estimates the midpoints between survey items with $F^{-1}(\hat{P}(j \text{ to } k))$, and an *ad hoc* procedure is used to estimate the respondents' locations on the attitude scale.

Summary

The early scaling methods of Thurstone, Guttman, Likert, and Coombs give researchers in the behavioral sciences a way to quantify, or measure, the seemingly unmeasurable construct we call attitude. This measure of attitude allowed researchers to examine how behavior varied according to differences in attitudes. However, these techniques are often too restrictive in their applicability. Modern attitude scaling techniques based on item response theory models overcome many of the limitations of the early scaling methods, but that is not to say they cannot be improved upon.

The increasing use of computer-based attitudinal surveys offer a number of ways to improve on the current attitude scaling methodologies. Typically attitudinal surveys have all respondents answering all survey items, but an adaptive survey may prove more powerful. Computerized adaptive assessments which select items that provide the most information about an individual's attitude (or ability) have been used extensively in educational testing [e.g., 52] and will likely receive more attention in attitudinal surveys. Roberts, Yin, and Laughlin, for example, introduce an adaptive procedure for unfolding response models [42].

The attitude scaling methods described here use discrete responses to measure respondents on the attitude scale. Another direction in which attitude scaling may be improved is with the use of continuous response scales. Continuous response scales typically ask respondents to mark their response on a line segment that runs from 'Complete Disagreement' to 'Complete Agreement'. Their response is then recorded as the proportion of the line that lies below the mark. Because a respondent can mark any point on the line, each response will likely contain more information about the attitude being studied than do discrete responses [2].

Continuous response formats are not a new development. In fact Freyd [16] discussed their use before Thurstone's Law of Comparative Judgment, but continuous response scales were once difficult to implement because each response had to be measured

by hand. Modern computer programming languages make continuous response scales more tractable. There are several options available for the analysis of multivariate continuous responses, including **factor analysis**, multivariate regression models (*see Multivariate Multiple Regression*), and generalizations of item response models to continuous response formats [43].

References

- [1] Allport, G.W. (1935). Attitudes, in *Handbook of Social Psychology*, C. Murchinson, ed., Clark University Press, Worcester, pp. 798–844.
- [2] Alwin, D.F. (1997). Feeling thermometers versus 7-point scales: which are better? *Sociological Methods & Research* **25**, 318–340.
- [3] Andrich, D. (1988). The application of an unfolding model of the PIRT type for the measurement of attitude, *Applied Psychological Measurement* **12**, 33–51.
- [4] Andrich, D. & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses, *Applied Psychological Measurement* **17**, 253–276.
- [5] Bechtel, G.G. (1968). Folded and unfolded scaling from preferential paired comparisons, *Journal of Mathematical Psychology* **5**, 333–357.
- [6] Bèguin, A.A. & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models, *Psychometrika* **66**, 471–488.
- [7] Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm, *Psychometrika* **46**, 443–459.
- [8] Bradley, R.A. (1976). Science, statistics, and paired comparisons, *Biometrika* **32**, 213–232.
- [9] Bradley, R.A. & Terry, M.E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons, *Biometrika* **39**, 324–345.
- [10] Coombs, C.H. (1950). Psychological scaling without a unit of measurement, *Psychological Review* **57**, 145–158.
- [11] Coombs, C.H. (1964). *A Theory of Data*, Wiley, New York.
- [12] Davison, M. (1977). On a metric, unidimensional unfolding models for attitudinal and development data, *Psychometrika* **42**, 523–548.
- [13] Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimates, *Psychometrika* **47**, 7–28.
- [14] Eagly, A.H. & Chaiken, S. (1993). *The Psychology of Attitudes*, Harcourt Brace Jovanovich.
- [15] Fishbein, M. & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*, Addison-Wesley.
- [16] Freyd, M. (1923). The graphic rating scale, *Journal of Educational Psychology* **14**, 83–102.
- [17] Goodman, L.A. (1975). A new model for scaling response patterns. *Journal of the American Statistical Society*, **70**; Reprinted in *Analyzing Qualitative/Categorical Data*, edited by Jay. Magdison. Cambridge, MA: Abt Books, 1978.
- [18] Guttman, L. (1950). The basis for scalogram analysis, *Measurement and Prediction, Studies in Social Psychology in World War II*, Vol. IV, University Press, Princeton, pp. 60–90.
- [19] Hemker, B.T., Sijtsma, K., Molenaar, I.W. & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models, *Psychometrika* **62**, 331–347.
- [20] Hoijtink, H. (1990). A latent trait model for dichotomous choice data, *Psychometrika* **55**, 641–656.
- [21] Horowitz, L.M., Rosenberg, S.E., Baer, B.A., Ureño, G. & Villaseñor, V.S. (1988). Inventory of interpersonal problems: psychometric properties and clinical applications, *Journal of Consulting and Clinical Psychology* **56**, 885–892.
- [22] Johnson, M.S. (2001). *Parametric and non-parametric extensions to unfolding response models*, PhD thesis, Carnegie Mellon University, Pittsburgh.
- [23] Johnson, M.S., Cohen, W.M. & Junker, B.W. (1999). Measuring appropriability in research and development with item response models, Technical report No. 690, Carnegie Mellon Department of Statistics.
- [24] Johnson, M.S. & Junker, B.W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models, *Journal of Educational and Behavioral Statistics* **28**(3), 195–230.
- [25] Junker, B.W. (1991). Essential independence and likelihood-based ability estimation for polytomous items, *Psychometrika* **56**, 255–278.
- [26] Kim, Y. & Pilkonis, P.A. (1999). Selecting the most informative items in the IIP scales for personality disorders: an application of item response theory, *Journal of Personality Disorders* **13**, 157–174.
- [27] Likert, R.A. (1932). A technique for the measurement of attitudes, *Archives of Psychology* **140**, 5–53.
- [28] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [29] Luce, R.D. (1959). *Individual Choice Behavior*, Wiley, New York.
- [30] Luo, G. (1998). A general formulation for unidimensional latent trait unfolding models: making explicit the latitude of acceptance, *Journal of Mathematical Psychology* **42**, 400–417.
- [31] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.
- [32] McDonald, R.P. (1997). Normal-ogive multidimensional model, in *Handbook of Modern Item Response Theory*, Chap. 15, W.J. van der Linden & R.K. Hambleton, eds, Springer-Verlag, New York, pp. 257–269.
- [33] Mosteller, F. (1951). Remarks on the method of paired comparisons: III. a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed, *Psychometrika* **16**, 207–218.

-
- [34] Muhlberger, P. (1999). A general unfolding, non-folding scaling model and algorithm, in *Presented at the 1999 American Political Science Association Annual Meeting*, Atlanta.
- [35] Noël, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: application to smoking cessation, *Psychological Methods* **4**(2), 173–191.
- [36] Patz, R. & Junker, B. (1999). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics* **24**, 342–366.
- [37] Post, W.J. (1992). *Nonparametric Unfolding Models: A Latent Structure Approach*, M&T Series, DSWO Press, Leiden.
- [38] Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika* **56**, 611–630.
- [39] Rasch, G. (1961). On general laws and the meaning of measurement in psychology, in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Statistical Laboratory, University of California, June 20–July 30, 1960, Proceedings published in 1961 by University of California Press.
- [40] Reckase, M.D. (1997). Loglinear multidimensional model for dichotomous item response data, in *Handbook of Modern Item Response Theory*, Chap. 16, W.J. van der Linden & R.K. Hambleton, eds, Springer-Verlag, New York, pp. 271–286.
- [41] Roberts, J.S., Donoghue, J.R. & Laughlin, J.E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory, *Applied Psychological Measurement* **24**(1), 3–32.
- [42] Roberts, J.S., Lin, Y. & Laughlin, J. (2001). Computerized adaptive testing with the generalized graded unfolding model, *Applied Psychological Measurement* **25**, 177–196.
- [43] Samejima, F. (1973). Homogeneous case of the continuous response model, *Psychometrika* **38**(2), 203–219.
- [44] Sixtl, F. (1973). Probabilistic unfolding, *Psychometrika* **38**(2), 235–248.
- [45] Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* **55**, 293–325.
- [46] Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model, *Psychometrika* **47**, 175–186.
- [47] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 278–286.
- [48] Thurstone, L.L. (1928). Attitudes can be measured, *American Journal of Sociology* **33**, 529–554.
- [49] Thurstone, L.L. & Chave, E.J. (1929). *The Measurement of Attitude*, University of Chicago Press, Chicago.
- [50] Verhelst, N.D. & Glas, C.A.W. (1995). Rasch models: foundations, recent developments, and applications, *The One Parameter Logistic Model*, Chap. 12, Springer-Verlag, New York.
- [51] Verhelst, N.D. & Verstralen, H.H.F.M. (1993). A stochastic unfolding model derived from the partial credit model, *Kwantitative Methoden* **42**, 73–92.
- [52] Wainer, H., ed. (2000). *Computerized Adaptive Testing: A Primer*, 2nd edition, Lawrence Erlbaum.
- [53] Wiley, J.A. & Martin, J.L. (1999). Algebraic representations of beliefs and attitudes: partial order models for item responses, *Sociological Methodology* **29**, 113–146.

(See also **Multidimensional Scaling; Unidimensional Scaling**)

MATTHEW S. JOHNSON AND
BRIAN W. JUNKER

Attrition

WILLIAM R. SHADISH AND JASON K. LUELLEN

Volume 1, pp. 110–111

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Attrition

Attrition refers to a loss of response from study participants. With measurement attrition, participants fail to complete outcome measures ranging from a single item to all measures. With treatment attrition, participants drop out of treatment regardless of whether they are measured (*see Dropouts in Longitudinal Studies: Methods of Analysis*).

Attrition can be random or systematic. Attrition is problematic for at least three reasons. First, all attrition reduces (*see Power*) to detect a statistically significant result. Second, systematic attrition can reduce the generalizability of results to the population from which the original sample was drawn. Third, systematic differential attrition (e.g., attrition that is correlated with a treatment condition in an experiment) leaves different kinds of participants in one condition versus another, which can bias estimates of treatment effects. For example, Stanton and Shadish [18] found that addicts with the worst prognosis are more likely to withdraw from discussion groups than from family therapy. If a study suggests that addicts respond better to discussion groups than family therapy, it may simply be because the worst addicts stayed in family therapy. Random differential attrition does not cause such a bias, but it can be difficult to establish whether attrition is random.

Shadish, Cook, and Campbell [16] discuss practical options for preventing and minimizing attrition. A key point is that the researcher should avoid measurement attrition even when treatment attrition occurs. A participant who did not receive the assigned treatment can be included in the analysis provided the participant completed the measures; but a participant with measurement attrition cannot be included in the analysis.

Because attrition is a potential threat to estimates of treatment effects in experiments, the researcher should analyze attrition thoroughly to understand the extent of the threat. Such analyses include simple descriptive statistics of the overall attrition rate, attrition rates broken down by group, overall differences between those who completed the study and those who did not, differences between those who completed the study and those who did not broken down by group, and differences between those who stayed in treatment and those who stayed in the control [7]. More detailed analyses examining whether different

groups or measures have different patterns of attrition may be undertaken with computer programs such as [19] and [11].

Alternatively, the researcher can try to compensate for attrition when estimating effects. The two general approaches are to estimate effects by imputing values for the **missing data** and to estimate effects without directly imputing missing data (*see Dropouts in Longitudinal Studies: Methods of Analysis; Missing Data*). Several approaches to imputing missing data exist [6, 13, 9, 10] (*see* [4] and [14] for accessible overviews). The simplest and least satisfactory approach is mean substitution; the best approach is usually **multiple imputation** with **maximum likelihood estimation**. *See* [5] for a review of computer programs that offer multiple imputation **Maximum Likelihood Estimation**.

When estimating effects without imputing missing data, multigroup **structural equation modeling** approaches may be useful [1–3, 12]. Other methods come from economists and involve modeling the dropout process itself (e.g., [8], and [20]). *See* [16] for a more detailed review of approaches to accounting for attrition when estimating effects. Regardless of whether imputing missing data or not, the researcher will generally want to conduct a variety of analyses under different assumptions about the nature of attrition and offer a range of estimates of effect (e.g., [15]). This can result in sets of treatment estimates under different assumptions that can be made to bracket the true effect under some conditions [17].

References

- [1] Allison, P.D. (1987). Estimation of linear models with incomplete data, in *Social Methodology*, C. Clogg, ed., Jossey-Bass, San Francisco, pp. 71–103.
- [2] Arbuckle, J.J. (1997). *Amos User's Guide, Version 3.6*, Small Waters Corporation, Chicago.
- [3] Bentler, P.M. (1995). *EQS: Structural Equations Program Manual*, Multivariate Software, Encin, CA.
- [4] Honaker, J., Joseph, A., King, G., Scheve, K. & Singh, N. (1999). *Amelia: A Program for Missing Data [Computer software]*, Harvard University, Cambridge, Retrieved from <http://gking.harvard.edu/stats.shtml#amelia>
- [5] Horton, N.J. & Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables, *The American Statistician* **55**, 244–255.
- [6] Jenrich, R.I. & Schluter, M.D. (1986). Unbalanced repeated measures models with structured covariance matrices, *Biometrics* **42**, 805–820.

- [7] Lazar, I. & Darlington, R. (1982). Lasting effects of early education, *Monographs of the Society for Research in Child Development* **47**, (2–3, Serial No. 195).
- [8] Leaf, R.C., DiGiuseppe, R., Mass, R. & Alington, D.E. (1993). Statistical methods for analyses of incomplete service records: Concurrent use of longitudinal and cross-sectional data, *Journal of Consulting and Clinical Psychology* **61**, 495–505.
- [9] Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition, Wiley, New York.
- [10] Little, R.J. & Schenker, N. (1995). Missing data, in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, G. Arminger, C.C. Clogg & M.E. Sobel, eds, Plenum Press, New York, pp. 39–75.
- [11] Marcantonio, R.J. (1998). *ESTIMATE: Statistical Software to Estimate the Impact of Missing Data [Computer software]*, Statistical Research Associates, Lake in the Hills.
- [12] Muthen, B.O., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random, *Psychometrika* **52**, 431–462.
- [13] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- [14] Schafer, J.L. & Graham, J.W. (2002). Missing data: our view of the state of the art, *Psychological Methods* **7**, 147–177.
- [15] Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association* **94**, 1096–1120.
- [16] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
- [17] Shadish, W.R., Hu, X., Glaser, R.R., Kownacki, R.J. & Wong, T. (1998). A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes, *Psychological Methods* **3**, 3–22.
- [18] Stanton, M.D. & Shadish, W.R. (1997). Outcome, attrition and family-couples treatment for drug abuse: a meta-analysis and review of the controlled, comparative studies, *Psychological Bulletin* **122**, 170–191.
- [19] Statistical Solutions. (2001). *SOLAS 3.0 for Missing Data Analysis [Computer software]*. (Available from Statistical Solutions, Stonehill Corporate Center, Suite 104, 999 Broadway, Saugus 01906).
- [20] Welch, W.P., Frank, R.G. & Costello, A.J. (1983). Missing data in psychiatric research: a solution, *Psychological Bulletin* **94**, 177–180.

WILLIAM R. SHADISH AND JASON K. LUELLEN

Average Deviation

DAVID C. HOWELL

Volume 1, pp. 111–112

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Average Deviation

The average deviation (AD) is a reasonably robust measure of scale, usually defined as

$$AD = \frac{\sum |X_i - m|}{n}, \quad (1)$$

where m is some measure of location, usually the mean, but occasionally the median. As defined, the AD is simply the average distance of observations from the center of the distribution. Taking deviations about the median rather than the mean minimizes AD.

The average deviation is often referred to as the mean absolute deviation and abbreviated M. A. D., but that notation is best used to refer to the **median absolute deviation**.

For a normally distributed variable, the average deviation is equal to 0.7979 times the standard deviation (SD). (For distributions of the same shape, the two estimators are always linearly related.) The AD becomes increasingly smaller than the standard deviation for distributions with thicker tails. This is due to the fact that the standard deviation is based on squared deviations, which are disproportionately influenced by tail values.

We often measure the relative qualities of two estimators in terms of their **asymptotic relative efficiency**, which is the ratio of the variances of those estimators over repeated sampling. For a normal distribution, the relative efficiency of the AD in comparison with the sample standard deviation is given by

$$\text{Relative efficiency} = \frac{\text{variance}(SD)}{\text{variance}(AD)} = 0.88 \quad (2)$$

This can be interpreted to mean that you would need to base your estimate of scale using AD as your estimator on 100 cases to have the same standard

error as a standard deviation based on 88 cases. In this case, the standard deviation is 12% more efficient. However, that advantage of the SD over the AD quickly disappears for distributions with more observations in the tails.

Tukey [2] illustrated this for a mixture of samples from two normal distributions with equal means; one distribution with $\sigma = 1.0$ and the other with $\sigma = 3.0$. (You might think of the latter as a distribution of careless responses.) The composite is called a *contaminated distribution*, a type of **finite mixture distribution**. When 99% of the observations were drawn from the first distribution and 1% were drawn from the second (and were not necessarily outliers), the relative efficiency of AD relative to SD jumped from 0.88 for the normal distribution to 1.44 for the contaminated distribution. In other words, the AD was nearly 50% more efficient than the SD for even a slightly contaminated distribution. And this was in a situation in which you would need many thousands of observations for the differences in the two distributions to be obvious to the eye. With only two observations out of 1000 drawn from the ‘contaminating’ distribution, the AD and SD were equally efficient. Tukey long advocated the use of what he called robust statistics, and had relatively little regard for the variance and standard deviation as estimators for data from most research studies [1].

References

- [1] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Linear Regression*. Addison-Wesley, Reading.
- [2] Tukey, J.W. (1960). A survey of sampling from contaminated distributions, in *Contributions to Probability and Statistics*, I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B. Mann, eds, Stanford University Press, Stanford.

DAVID C. HOWELL

Axes in Multivariate Analysis

WOJTEK J. KRZANOWSKI

Volume 1, pp. 112–114

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Axes in Multivariate Analysis

Right from the start of development of the topic, multivariate analysis has been closely associated with multidimensional geometry. In one of the pioneering papers, **Pearson** [3] took a geometrical standpoint, using the representation of p variables measured on a sample of n objects as a set of n points in p -dimensional space, and went on to introduce **principal component analysis** as a technique for successively identifying the r -dimensional subspaces of closest fit to these points for $r = 1, 2, \dots, p-1$. This technique is now a cornerstone of descriptive multivariate analysis, and the corresponding geometrical representation of the data is at the heart of many other analytical techniques.

So let us suppose that the p variables $X_1, X_2, \dots, X_p$ have been measured on n sample individuals, and that $\mathbf{x}_i^t = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the vector of P values observed on the i th individual for $i = 1, 2, \dots, n$ (vectors conventionally being interpreted as *column* vectors, hence the transpose superscript ‘t’ when the vector is written as a row). Moreover, we need to assume that all variables are *quantitative*. Then, the above representation of the sample as n points in p dimensions is obtained directly by associating each variable X_j with one of a set of p orthogonal axes in this space and assigning the observed value x_i to the point with coordinates $(x_{i1}, x_{i2}, \dots, x_{ip})$ on these axes.

Of course, this geometrical representation is essentially an idealized model, as we can never actually see it when p is greater than three. Hence, the motivation for Pearson in 1901, and for many researchers since then, has been to identify low-dimensional subspaces of the full p -dimensional space into which the data can be projected in order to highlight ‘interesting’ features in such a way that they can be plotted and seen. Now a one-dimensional subspace is just a line in the original space, a two-dimensional subspace can be characterized by any pair of lines at right angles (i.e., orthogonal) to each other, a three-dimensional subspace by any three mutually orthogonal lines, and so on. Thus, the search for an r -dimensional subspace in the original space resolves itself into a search for r mutually orthogonal lines in the space.

Moreover, if we can quantify the aspect of ‘interestingness’ that we wish to capture in the subspace by some numerical index or function, then the problem becomes one of seeking r mutually orthogonal lines that optimize this index or function. The final step is to realize that if the axes of the original space correspond to the variables $X_1, X_2, \dots, X_p$, then multidimensional geometry tells us that any line Y in the space can be expressed as a linear combination $Y = a_1X_1 + a_2X_2 + \dots + a_pX_p$ for suitable values of the coefficients $a_1, a_2, \dots, a_p$. Moreover, if $Z = b_1X_1 + b_2X_2 + \dots + b_pX_p$ is another line in the space, then Y and Z are orthogonal if and only if

$$a_1b_1 + a_2b_2 + \dots + a_pb_p = \sum_{i=1}^p a_ib_i = 0 \quad (1)$$

Thus, when Pearson looked for an r -dimensional subspace ‘closest’ to the original data points, he in effect looked for r mutually orthogonal combinations like Y and Z that defined this subspace. These he called the *principal components* of the data. **Hotelling** subsequently [1] showed that these components were the mutually orthogonal linear combinations of the original variables that maximized the sample variance among all linear combinations, an operationally better criterion to work with. These components can then be treated as the *axes* of the subspace, and an approximate representation of the data is given by plotting the *component scores* against these axes. Since the axes are related to the original variables, it is often possible to interpret them in terms of the substantive application and this will help in any interpretation of the plot of the scores.

As an example, consider a study conducted into differences in texture as perceived in the mouth among 24 meat samples. A panel of trained assessors scored each of the meat samples on a scale of 0–10 for each of 31 descriptors of texture, so that in our notation above we have $n = 24$ and $p = 31$. The resulting data can thus be visualized as 24 points (representing the meat samples) in 31 dimensions. Clearly, this is impossible to represent physically, but if we conduct a principal component analysis of the data we find that the first two components together account for about two-thirds of the total variance among the meat samples that was present in the original 31 variables. Thus, a simple scatter plot of the scores, using components 1 and 2 as axes, will give a good two-dimensional approximation

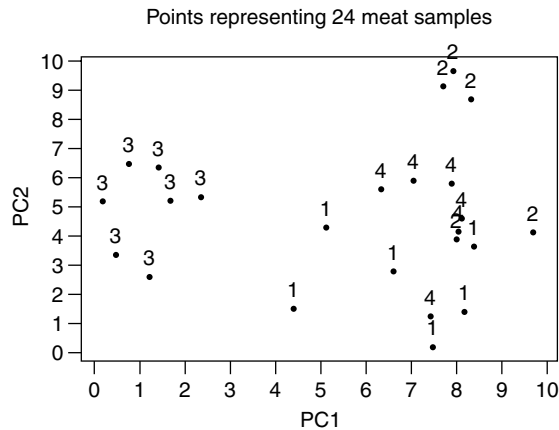


Figure 1 A scatter plot of the scores for each of 24 meat samples on the first two principal components

to the true 31-dimensional configuration of points. This plot is shown in Figure 1 below. The 24 meat samples were of four types: reformed meats (type 1), sausages (type 2), whole meats (type 3), and beef burgers (type 4). The points in the diagram are labelled by type, and it is immediately evident that whole meats are recognizably different from the other types. While the other three types do show some evidence of systematic differences, there are, nevertheless, considerable overlaps among them. This simple graphical presentation has thus provided some valuable insights into the data.

In recent years, a variety of techniques producing subspaces into which to project the data in order to optimize criteria other than variance has been developed under the general heading of **projection pursuit**, but many of the variants obtain the data projection in a subspace by direct computational means, without the intermediate step of obtaining linear combinations to act as axes. In these cases, therefore, any substantive interpretation has to be based on the plot alone. Other popular techniques such as *canonical variate analysis* (see **Canonical Correlation Analysis**) produce linear combinations but nonorthogonal ones. These are therefore *oblique* axes in the original space, and if used as orthogonal axes against which to plot scores they produce

a deformation of the original space. In the case of canonical variables, such a deformation is justified because it converts **Mahalanobis distance** in the original space into Euclidean distance in the subspace [2], and the latter is more readily interpretable. In some techniques, such as **factor analysis**, the linear combinations are derived implicitly from a statistical model. They can still be viewed as defining axes and subspaces of the original space, but direct projection of points into these subspaces may not necessarily coincide with derived scores (that are often estimated in some way from the model). In any of these cases, projecting the original axes into the subspace produced by the technique will show the inclination of the subspace to the original axes and will help in the interpretation of the data. Such projection of axes into subspaces underlies the ideas of **biplots**.

The above techniques have required quantitative data. Data sets containing qualitative, nominal, or ordinal variables will not permit direct representation as points in space with coordinates given by variable values. It is, nevertheless, possible to construct a representation using techniques such as **multidimensional scaling** or **correspondence analysis**, and then to derive approximating subspaces for this representation. However, such representations no longer associate variables with coordinate axes, so there are no underlying linear combinations of variables to link to the axes in the approximating subspaces.

References

- [1] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441.
- [2] Krzanowski, W.J. (2000). *Principles of Multivariate Analysis: A User's Perspective*, University Press, Oxford.
- [3] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series* **2**, 559–572.

WOJTEK J. KRZANOWSKI

Bagging

ADELE CUTLER, CHRIS CORCORAN AND LESLIE TOONE

Volume 1, pp. 115–117

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bagging

Bagging was introduced by Leo Breiman in 1996 [1] and is an example of a general class of methods called *ensembles*, in which we combine procedures to give greater predictive accuracy. Bagging stands for ‘bootstrap aggregating’ and means that we apply a regression or classification procedure to bootstrap samples from the original data and aggregate the results.

In bagging, we first take a *bootstrap sample* (see **Bootstrap Inference**) from the data by randomly sampling cases *with* replacement, until the bootstrap sample has the same number of cases as the original data. Some of the original cases will not appear in the bootstrap sample. Others will appear once, twice, or even more often.

To bag a **nonlinear regression** procedure, we take many bootstrap samples (a thousand is not uncommon) and fit a regression model to each bootstrap sample. We combine the fitted models by averaging their predicted response values. Figure 1 illustrates bagging a regression tree. In Figure 1(a), we show the data and the underlying function; in Figure 1(b), the result of fitting a single regression tree (see **Tree Models**); in Figure 1(c), the results of fitting regression trees to 10 bootstrap samples; and in Figure 1(d), the average over 100 such trees. Bagging gives a smoother, more accurate fit than the single regression tree.

For classification, we fit a classifier to each bootstrap sample and combine by choosing the most frequently predicted class, which is sometimes called *voting*. For example, if 55% of the fitted classifiers predict ‘class 1’ and the other 45% predict ‘class 2’, the bagged classifier predicts ‘class 1’.

The classification or regression procedure that we apply to each bootstrap sample is called the *base learner*. Breiman [1] suggests that bagging can substantially increase the predictive accuracy of an *unstable* base learner, that is, one for which small changes in the dataset can result in large changes in the predictions. Examples of unstable methods include trees, stepwise regression, and neural nets, all of which are also *strong learners* – methods that perform much better than random guessing. However, bagging can also turn a very inaccurate method

(a so-called *weak learner*) into a highly accurate method, provided the method is sufficiently unstable. To introduce more instability, the base learner itself may incorporate randomness. Dietterich [4] introduced a method in which the base learners are trees with splits chosen at random from amongst the best 20 splits at each node. **Random forests** [2] can be thought of as bagged trees for which nodes are split by randomly selecting a number of variables, and choosing the best split amongst these variables.

There are several other examples of combining random predictors, such as [5, 6], some of which bypass the bootstrap altogether and simply apply the base learner to the entire dataset, relying on the randomness of the base learner. However, the bootstrap has useful by-products because roughly one-third of the observations are omitted from each bootstrap sample. We can use these so-called *out-of-bag* cases to get good estimates of prediction error [2], eliminating the need for formal cross-validation. They also give measures of variable importance; consider randomly permuting the values of a variable in the out-of-bag cases and comparing the resulting prediction error to the original prediction error on the original out-of-bag data. If the variable is important, the prediction error for the permuted data will be much higher; if it is unimportant, it will be relatively unchanged.

For very large problems (thousands of independent variables), bagging can be prohibitively slow, in which case random forests [3] are more suitable. Bagging trees is a special case of random forests, so the methods for interpreting and understanding random forests are also applicable to bagged trees. Another related method is **boosting**.

We illustrate bagging using a dataset from the Cache County Study on Memory in Aging [7]. The dataset comprised 645 participants aged 65 or older for whom a clinical assessment of Alzheimer’s disease was determined by a panel of experts. We included only subjects who were assessed as non-impaired (having no cognitive impairment) (class 1; 465 people) or having only Alzheimer’s disease (class 2; 180 people). The subjects also completed several neuropsychological tests, and our goal was to investigate how well we could predict the presence of Alzheimer’s disease using the neuropsychological test results.

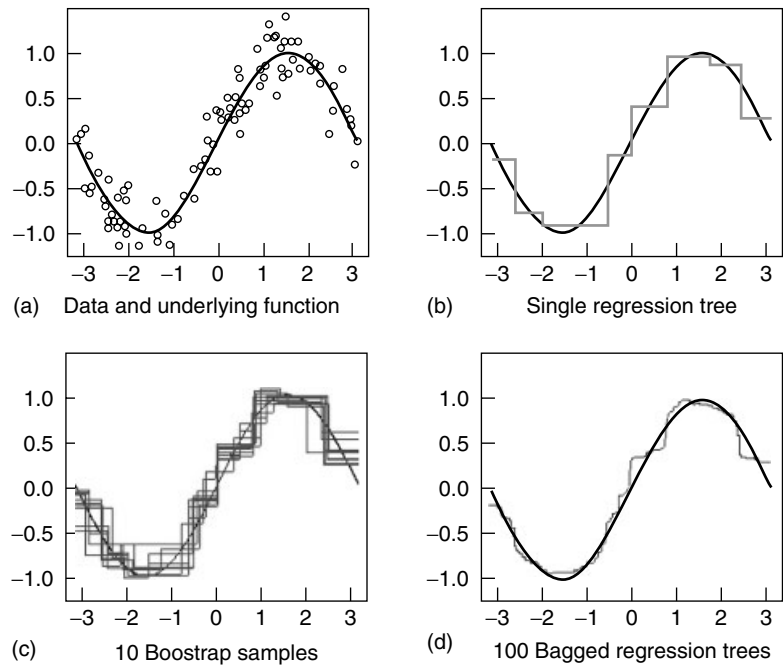


Figure 1 Bagging regression trees

Table 1 LDA, error rate 4.7%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	91	2
	Alzheimer's	4	32

Table 2 Logistic regression, error rate 3.9%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	90	3
	Alzheimer's	2	34

We imputed missing values using the class-wise medians. Then, we divided the data into a training set of 516 subjects and a test set of 129 subjects. Using the R package (see **Software for Statistical Analyses**), we fit linear **discriminant analysis** (LDA),

Table 3 Classification tree, error rate 10.1%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	88	5
	Alzheimer's	8	28

Table 4 Bagged classification tree, error rate 3.1%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	90	3
	Alzheimer's	1	35

logistic regression, classification trees (see **Classification and Regression Trees**), bagged classification trees, boosted classification trees, and random forests to the training set and used the resulting classifiers to predict the disease status for the test set. We found the results in Tables 1–6.

Table 5 Boosted classification tree, error rate 3.1%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	90	3
	Alzheimer's	1	35

Table 6 Random forests, error rate 2.3%

		Predicted class	
		Non-impaired	Alzheimer's
True class	Non-impaired	91	2
	Alzheimer's	1	35

We note that even though an individual classification tree had the worst performance, when such trees were combined using bagging, boosting, or random forests, they produced a classifier that was more accurate than the standard techniques of LDA and logistic regression.

References

- [1] Breiman, L. (1996a). Bagging predictors, *Machine Learning* **26**(2), 123–140.
- [2] Breiman, L. (1996b). Out-of-bag estimation, University of California, Technical report, University of California, Department of statistics, <ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z>
- [3] Breiman, L. (2001). Random forests, *Machine Learning* **45**(1), 5–32.
- [4] Dietterich, T.G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization, *Machine Learning* **40**(2), 139–158.
- [5] Ho, T.K. (1998). The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 832–844.
- [6] Ji, C. & Ma, S. (1997). Combinations of weak classifiers, *IEEE Transactions on Neural Networks* **8**(1), 32–42.
- [7] Tschanz, J.T., Welsh-Bohmer, K.A., Skoog, I., Norton, M.C., Wyse, B.W., Nickles, R. & Breitner, J.C.S. (2000). Dementia diagnoses from clinical and neuropsychological data compared: the cache county study, *Neurology* **54**, 1290–1296.

ADELE CUTLER, CHRIS CORCORAN AND
LESLIE TOONE

Balanced Incomplete Block Designs

JOHN W. COTTON

Volume 1, pp. 118–125

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Balanced Incomplete Block Designs

Agricultural scientists investigating the effects of different fertilizers on crop yields, for example, can split a growing area into different regions called *blocks*. Then they divide each block into plots (cells or units) to receive different treatments. By randomizing the assignment of treatments to the different plots, the experimenters reduce or eliminate bias and facilitate the statistical analysis of yields in different plots and blocks. Similarly, experimenters performing repeated measurements designs (*see Repeated Measures Analysis of Variance; Longitudinal Data Analysis*) often use balancing procedures to equalize the contribution of special effects, such as individual differences, upon the average performance under each treatment. For example, they can employ a **Latin square design** in order to have each person serve once on every occasion and once in every treatment. But if it is necessary to limit the number of observations per person, an incomplete design with fewer observations than treatments is employed. The design becomes a *balanced incomplete design* by adding restrictions, which will be described below. This article examines ways to construct or select a balanced incomplete design with a given number of blocks and treatments. It also presents alternative data analysis procedures that are appropriate for such designs.

One formal structure for a BIBD is as follows. Let there be b blocks (or individual persons) forming the rows of a rectangle, k columns in the rectangle (defining cells in a block), and t treatments, each replicated r times. The total number of observations in the design is $N = rt = bk$. There are other constraints: $b \geq t > k$, preventing any block from including every treatment. The design is balanced in that the number of times, λ , that two treatments such as A and C appear in some block is a constant for all possible pairs of treatments. For an observation Y_{ij} in Treatment i and Block j , the basic model for this design is

$$Y_{ij} = \mu + \tau_i + \beta_j + e_{ij}, \quad (1)$$

where μ is the population mean, τ_i is the effect of Treatment i , β_j is the effect of contribution j , and e_{ij} is error for that observation that is distributed as

$N(0, \sigma^2)$. We adopt the convention that

$$\sum_i \tau_j = \sum_j \beta_j = 0. \quad (2)$$

An Example of a BIBD Experiment

John [6] presents a standard analysis of a dishwashing experiment with the dependent variable being the number of plates washed before foam disappeared, discussing it again in [7, pp. 221–234]. The treatments are nine different detergents. A block consists of three different workers, each with a basin in which to wash with a different detergent in the range from A to J omitting I. The response measure is the number of plates washed at a constant speed before the foam from the detergent disappears. The experiment has $t = 9$ treatments, $k = 3$ plots (cells per block), $r = 4$ replications of each treatment, $b = 12$ blocks, and $\lambda = 1$ cases in which each pair of treatments such as A and C appear together in any block. $N = 4(9) = 12(3) = 36$, which is consistent with the fact that with nine treatments there are $t!/(t-2)!2! = 9!/(7!2!) = 36$ possible treatment pairs. There is another relation between λ and other design features: $\lambda(t-1) = r(k-1)$. Table 1(a) presents this design in a compact form together with data to be discussed below. Table 1(b) presents the same information in a display with the rows as treatments and the columns as blocks, which necessitates blank cells for treatment and block combinations that do not appear.

Intrablock and Interblock Information Assessed for the Example Experiment

In a balanced incomplete blocks design, we have two basic choices as to how to compare the effects

Table 1(a) A balanced incomplete block design with twelve blocks and nine treatments (Design and Data from John [6, p. 52])

		BLOCK						
1	A (19)	B (17)	C (11)	7	A(20)	E(26)	J(31)	
2	D (6)	E (26)	F (23)	8	B(16)	F(23)	G(21)	
3	G (21)	H (19)	J (28)	9	C(13)	D(7)	H(20)	
4	A (20)	D (7)	G (20)	10	A(20)	F(24)	H(19)	
5	B (17)	E (26)	H (19)	11	B(17)	D(6)	J(29)	
6	C (15)	F (23)	J (31)	12	C(14)	E(24)	G(21)	

2 Balanced Incomplete Block Designs

Table 1(b) Design redisplayed to focus on responses to specific treatments

	BLOCK												
Treat	1	2	3	4	5	6	7	8	9	10	11	12	Mean
A	19			20			20			20			19.8
B	17				17			16			17		16.8
C	11					15			13			14	13.2
D		6		7					7		6		6.5
E		26			26		26					24	25.5
F		23				23		23		24			23.2
G			21	20				21				21	20.8
H			19		19				20	19			19.2
J			28			31	31				29		29.8
Mean	15.7	18.3	22.7	15.7	20.7	23.0	25.7	20.0	13.3	21.0	17.3	19.7	19.42

of two or more different treatments. First, consider the problem of comparing responses to Treatments A and B in Table 1. The most common **analysis of variance** method assesses the effects of two design features, blocks, and treatments. Block 1 has a score of 19 under A and a score of 17 under B. The difference, +2, can be called an *intra-block comparison*. Because the [6] data set has $\lambda = 1$, there is only one A – B comparison in our research example. Otherwise we could average all such A – B differences in making an intra-block comparison of these two treatments. A comparison of the performance in different blocks allows the comparison of raw averages for the different blocks, regardless of the treatments involved. However, an overall test of intra-block effects in an analysis of variance must extract block effects first and then extract treatment effects adjusted for block effects. This is called an *intra-block analysis*. Equations 3 through 10 below are consistent with [6, p. 52]. Let

$$C = \frac{\left(\sum_i \sum_j Y_{ij} \right)^2}{N}, T_i = \sum_j Y_{ij}, B_j = \sum_i Y_{ij} \quad (3)$$

and

$$Q_i = kT_i - \sum_j n_{ij} B_j. \quad (4)$$

Because $n_{ij} = 1$ rather than 0 only if Treatment i is present in Block j , Q_i adjusts kT_i by subtracting the total of all block totals for all blocks containing Treatment i . The estimated Treatment i effect is

$$\hat{\tau}_i = \frac{Q_i}{t\lambda}, \quad (5)$$

and the adjusted (*adj*) average for that treatment is

$$\bar{Y}_{i(\text{adj})} = \bar{Y} + \hat{\tau}_i. \quad (6)$$

The sum of squares breakdown for an intra-block analysis is as follows.

$$SS_{\text{Total}} = \sum \sum Y_{ij}^2 - C, \quad (7)$$

$$SS_{\text{Blocks}} = \frac{\sum B_j^2}{k} - C, \quad (8)$$

$$SS_{\text{Treatments(adj)}} = \frac{\sum Q_i^2}{\lambda kt}, \quad (9)$$

and

$$SS_{\text{Residual}} = SS_{\text{Total}} - SS_{\text{Blocks}} - SS_{\text{Treatments(adj)}}. \quad (10)$$

The degree of freedom values are

$$df_{\text{Total}} = N - 1, df_{\text{Blocks}} = b - 1,$$

$$df_{\text{Treatments}} = t - 1, \quad \text{and}$$

$$df_{\text{Residual}} = N - t - b + 1, \text{ respectively.}$$

Alternatively, for an interblock analysis, we can find all treatment averages regardless of the blocks in which they are located. Table 1(b) shows these means for our example data. It is intuitive that we need to compute differences such as $\bar{Y}_A - \bar{Y}_B$ and $\bar{Y}_A - \bar{Y}_C$ on the basis of several observations rather than on one for each treatment, thus computing an *interblock comparison* for any two treatments. In practice, we can either proceed with a general method [7, p. 225] of finding a total sum of squares because of regression

and extracting new squares from it as needed, or by computing the values of SS_{Total} and SS_{Residual} from the intrablock analysis and then finding the unadjusted value of the sum of squares for treatments with the obvious formula:

$$SS_{\text{Treatments}} = \frac{\sum T_i^2}{r} - C. \quad (11)$$

and finding the adjusted sum of squares for blocks by subtraction:

$$SS_{\text{Blocks(adj)}} = SS_{\text{Total}} - SS_{\text{Treatments}} - SS_{\text{Residual}}. \quad (12)$$

Table 2 presents three versions of the analysis of variance for the [6] BIBD example, going beyond calculation of mean squares to include F tests not reported earlier. The first two analyses correspond to the two just described. The so-called Type I analyses extract one effect first and then adjust the sum of squares for the next variable by excluding effects of the first. The first example of a Type I analysis is consistent with a previously published [6, p.54] *intrablock analysis*, in which block effects are extracted first followed by the extraction of treatment effects adjusted for block effects. Here treatment effects and block effects both have very small P values ($p < 0.0001$).

The second Type I analysis reverses the order of extracting the sums of squares. Treatment effects are computed first and then block effects adjusted for treatment effects are measured. Therefore, this is an *interblock analysis*. For this analysis, the treatment effects of the P value is very small ($p < 0.0001$), but the adjusted block effects do not even come close to the 0.05 level.

The third analysis in Table 2 is called a *Type II analysis*. Here an adjusted sum of squares is used

from each effect that has been assessed. Accordingly, for treatments and blocks that are the only factors of interest, this analysis is implied by its predecessors using $SS_{\text{Treatments(adj)}}$ from the intrablock analysis and $SS_{\text{Blocks(adj)}}$ from the interblock analysis. All Table 2 entries not previously computed [6] could have been obtained with a hand calculator using the formulas above or simpler ones such as $F = MS_{\text{Effects}}/MS_{\text{Error}}$. Actually, they were calculated using a SAS Proc GLM Type I or II analysis.

The model of (1) assumes that both treatment and block effects are fixed (*see Fixed and Random Effects*). An alternate mixed effects model used in the SAS program just mentioned is

$$Y_{ij} = \mu + \tau_i + b_j + e_{ij}, \quad (13)$$

where the Roman symbol b_j implies that the effect for the j th block is random rather than fixed as with the Greek β_j of (1). This new model assumes that the blocks have been selected at random from a population of blocks. Most authors performing interblock analyses employ the model in (13).

All sums of squares, mean squares, and F statistics obtained with these two models are identical, but expected mean squares for them differ. One reason to consider using an intrablock analysis is that the average mean square in a BIBD is not contaminated by block effects. With an intrablock analysis of the current data using (13), this expected mean square is $\sigma^2 + \text{a function of treatment effects, } \tau_i^2$. In contrast, with an interblock analysis, the expected mean square is $\sigma^2 + 0.75\sigma_b^2 + \text{a function of treatment effects}$. With the latter expected mean square, a significantly large F for treatments might be due to block effects, σ_b^2 , rather than treatment effects. A comparable problem arises in testing block effects in an intrablock analysis, where the expected mean square for blocks is $\sigma^2 + 3\sigma_b^2 + \text{a function of treatment effects}$. If a

Table 2 An expanded analysis of Table 1 Data

Source	df	Type I SS (Intrablock Analysis)	MS	F	p
Blocks	11	412.75	37.52	45.53	<0.0001
Treatments(Adjusted)	8	1086.815	135.85	164.85	<0.0001
Source	df	Type I SS (Interblock Analysis, New Order)	MS	F	p
Treatments	8	1489.50	186.19	225.94	<0.0001
Blocks(Adjusted)	11	10.065	0.91	1.11	0.4127
Source	df	Type II SS	MS	F	p
Blocks(Adjusted)	11	10.065	0.91	1.11	0.4127
Treatments(Adjusted)	8	1086.815	135.85	164.85	<0.0001
Error	16	13.185	0.82	—	—

4 Balanced Incomplete Block Designs

test for block effects is not of interest to the experimenter, an intrablock analysis is fine – one merely fails to report a test for block effects. However, a Type II analysis usually has the advantage of yielding uncontaminated expected mean squares (and therefore uncontaminated F values) for both treatment and block effects.

Evaluating Differences Between Two Treatment Effects and Other Contrasts of Effects

An overall assessment of treatment effects in the dishwashing experiment [6] can be supplemented by a comparison of the adjusted average numbers of dishes washed with one detergent and some other detergent. Alternatively, one can make more complicated analyses such as comparing the adjusted average of A and B with the adjusted average for C (see **Multiple Comparison Procedures**). Consider a comparison between A and B. From (6), we know how to compute $\bar{Y}_{A(\text{adj})}$ and $\bar{Y}_{B(\text{adj})}$. We need to know the variance (V) and standard error ($s.e.$) of each adjusted treatment mean and of the difference between two independent adjusted means. Knowing from [1, p. 275] that

$$V(\bar{Y}_{i(\text{adj})}) = \frac{k\sigma^2}{\lambda t}, \quad (14)$$

for each treatment is helpful, but we must estimate σ^2 with MS_{Error} from Table 2 and then use the relation

$$V(\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})}) = \frac{2k\sigma^2}{\lambda t}, \quad (15)$$

leading to

$$s.e.(\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})}) = \sqrt{\frac{2k\sigma^2}{\lambda t}}. \quad (16)$$

Now a standard t Test with $df = N - t - b + 1$ is

$$t = \left(\frac{\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})}}{s.e.(\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})})} \right). \quad (17)$$

For the dishwashing experiment, (4–6) and (14–17) yield

$$T_A = 79, \quad T_B = 67,$$

$$B_A = 234, \quad B_B = 221,$$

$$Q_A = 3, \quad Q_B = -20,$$

$$\hat{\tau}_A = 0.333, \quad \hat{\tau}_B = -2.222,$$

$$\bar{Y}_{A(\text{adj})} = 19.75, \quad \bar{Y}_{B(\text{adj})} = 17.19,$$

$$V(\bar{Y}_{A(\text{adj})}) = 0.2733 = V(\bar{Y}_{B(\text{adj})}),$$

$$V(\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})}) = 0.5467,$$

$$\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})} = 2.556,$$

$$s.e.(\bar{Y}_{A(\text{adj})} - \bar{Y}_{B(\text{adj})}) = 0.74,$$

and

$$t = 3.46, \quad (18)$$

with $df = 36 - 9 - 12 + 1 = 16$, so that a two-tailed test has $p < 0.01$. So performance under Detergent A is significantly better than under Detergent B.

More complicated contrast analyses use standard methods, essentially the same as those with means from a one-way analysis of variance. The principal difference in the BIBD case is that means, variances, and standard errors reflect the adjustments given above.

Evaluation of Possible Polynomial Trends in Treatment Effects for BIBD Data

A further analysis [6] of the dishwashing data moved from the usual study of qualitative treatment variations to quantitative treatment variations. The nine detergents studied included a standard detergent (Control), four with a first base detergent with 0 to 3 amounts of an additive, and four with a second base detergent with 0 to 3 amounts of an additive. Eight contrasts among the nine treatments were evaluated: Linear, quadratic, and cubic components for Base 1; linear, quadratic, and cubic components for Base 2; Base 1 versus Base 2; and Control versus Bases 1 and 2 combined. The resulting fixed effects ANOVA found significant linear and quadratic effects of additive amounts for Base 1, significant linear effects of additive amounts for Base 2, significant superiority of Base 2 over Base 1, and significant superiority of Control over the two averages of Base 1 and 2. As expected, the linear effects were increasing the number of plates washed with increasing amounts of additive. Also see [6] for formulas for this contrast analysis. See also [4, Ch. 5] and various sources such

as manuals for statistical computing packages for a general treatment of polynomial fitting and significance testing.

Bayesian Analysis of BIBD Data

Bayesian analysis adds to conventional statistics a set of assumptions about probable outcomes, thus combining observed data and the investigator's expectations about results (*see Bayesian Methods for Categorical Data*). These expectations are summarized in a so-called prior distribution with specific or even very vague indications of a central tendency measure and variability measure related to those beliefs. The prior distribution plus the observed data and classical assumptions about the data are combined to yield a posterior distribution assigning probabilities to parameters of the model. Box and Tiao [2, pp. 396–416] present a Bayesian method of analysis for BIBD data sets originally assumed to satisfy a mixed model for analysis of variance. Beginning with an assumption of a noninformative prior distribution with equal probabilities of all possible σ_b^2 (block effect variance) and σ_e^2 (error variance) values, they prove an equation defining the posterior distribution of the parameters $\tau_i - \bar{\tau}$. This posterior distribution is a product of three factors: (a) a multivariate t distribution centered at the mean of intrablock treatment effects, (b) a multivariate t distribution centered at the mean of interblock treatment effects, and (c) an incomplete beta integral with an upper limit related to the treatment vector of parameters $\tau_i - \bar{\tau}$. Combining the first two factors permits giving a combined estimate of treatment effects simultaneously reflecting intrablock and interblock effects. Applications of approximation procedures are shown in their [2, pp. 415–417]. Their Table 7.4.5 includes numerical results of this initially daunting analysis method as applied to a set of simulated data for a three-treatment, fifteen-block BIBD experiment. Also see [7, pp. 235–238] for a non-Bayesian combination of intrablock and interblock treatment effects.

Youden Squares as a Device for Increasing the Number of Effects Studied in a BIBD Model

A Youden square is a BIBD with the same number of treatments as blocks. One advantage of using

such a design is that it permits (but does not require) the inclusion and assessment of an additional experimental variable with as many values as the number of plots per block. Normally each value of the added (auxiliary) variable occurs exactly once with each value of the main treatment variable. This is an orthogonal relationship between two variables. Box, Hunter, and Hunter [1, p. 260, pp. 276–279] describe the use of a Youden square design in a so-called wear testing experiment, in which a machine simultaneously measures the amount of wear in $k = 4$ different pieces of cloth after an emery paper has been rubbed against each for 1000 revolutions of the machine. The observed wear is the weight loss (number of 1 milligram units) in a given piece of cloth. Their example presumes $t = 7$ kinds of cloth (treatments) and $b = 7$ testing runs (blocks). An added variable, position of the emery paper rubbing a cloth, had four options with each appearing with one of the four cloths of a block. In discussing this experiment, the authors expand (1) above to include an ' α_l ' effect with this general purpose

$$Y_{ij} = \mu + \alpha_l + \tau_i + \beta_j + e_{ij}. \quad (19)$$

In their case, this extra effect is called an $l = 7$ blocks (positions) effect because it is generated by the emery paper positions within blocks. Table 3 combines information from [1, p. 260, p. 277] to display wear, treatments (from A to G), and blocks (position) in each plot of each block in the experiment. In order to facilitate a later analysis in Table 5, I have reordered cells within the different blocks of [1, p. 277] in a nonunique way ensuring that each treatment appears exactly once in each column of Table 3. A Type I analysis of variance [1, p. 279] assesses

Table 3 Design and data from a Youden square experiment with an extra independent variable [1, p. 260, p. 277]

BLOCK	PLOT				Total
	1	2	3	4	
1	B α 627	D β 248	F γ 563	G δ 252	1690
2	A α 344	C β 233	G γ 226	F δ 442	1245
3	C α 251	G β 297	D γ 211	E δ 160	919
4	G α 300	E γ 195	B δ 537	A β 337	1369
5	F α 595	B γ 520	E β 199	C δ 278	1592
6	E α 185	F β 606	A γ 369	D δ 196	1356
7	D α 273	A β 396	C γ 240	B δ 602	1511

6 Balanced Incomplete Block Designs

Table 4 A SAS PROC MIXED analysis of Table 1 BIBD data, extracting only treatment effects (Blocks are random and serve as a control factor.)

Source	$df_{\text{numerator}}$	$df_{\text{denominator}}$	$SS_{\text{numerator}}$	F	p
Treatments	8	16	Not shown	220.57	<0.001

these effects in the following order: blocks, blocks (positions), and treatments. Compared to a comparable analysis [1, p. 278] without blocks (positions), the authors find identical sums of squares for treatments and blocks as before. This is a consequence of the orthogonality between treatments and blocks (positions) in the wear testing experiment. Because the sum of squares for the residual error is reduced by the amount of the sum of squares for blocks (position), the F for treatments is increased in the expanded analysis.

Modern Mixed Model Analysis of BIBD Data

In the early analysis of variance work, equations like (1) and (13) above were universally employed, using what is called the **generalized linear model** (GLM) regardless of whether random effects other than error were assumed. More modern work such as [9, p. 139] replaces (13), for example, with

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (20)$$

where $\mathbf{X}$ is a design matrix for fixed effects, $\mathbf{B}$ is a vector of fixed parameters, $\mathbf{Z}$ is the design matrix for random effects, $\mathbf{u}$ is a vector of random elements, and $\mathbf{e}$ is the vector of error elements. The new kind of analysis estimates the size of random effects but does not test them for significance. Thus, random block effects become a covariate used in assessing treatment effects rather than effects to be tested for significance themselves. We believe, like others such as Lunneborg [10], that more attention needs to be given to the question of whether so-called random effects have indeed been randomly drawn from a specific population. Strict justification of using a *mixed model* analysis or even a random or mixed model GLM analysis seems to require an affirmative answer to that question or some indication of the test's robustness to its failure.

Table 4 summarizes a reanalysis of Table 1 dishwashing data using SAS PROC MIXED. Because

blocks are now treated as a random covariate, the only significance test is for treatments, yielding an $F = 220.57$, almost identical to the 225.94 for the comparable interblock analysis test in Table 2.

Using Youden Squares to Examine Period Effects in a BIBD Experiment Using Mixed Models

Psychologists may desire to use a Youden square or other modified BIBDs design permitting an assessment of the effects of additional features, such as occasions (periods, stages, time, or trial number). Accordingly, Table 5 presents a PROC MIXED analysis of the Table 3 wear testing experiment data with Plots 1 through 4 now being interpreted as stages, and blocks (positions) are ignored. This new analysis shows significant ($p < 0.0001$) effects of treatments but not of stages, even at the 0.05 level. The program employed uses the Kenward–Roger degrees of freedom method of PROC MIXED, as described in [8]. Behavioral scientists using repeated measures in Youden squares and other BIBDs also may want to sacrifice the advantages of orthogonality, such as higher efficiency, in order to analyze both period effects and carryover effects (residual effects of prior treatments on current behavior) as is done with other designs [8]. This also permits the selection from a larger number of BIBDs of a given size, which can facilitate use of randomization tests.

Table 5 A SAS PROC MIXED reanalysis of the Table 3 Youden square example data (Assume that plots (columns) are stages whose effects are to be assessed)

	$df_{\text{numerator}}$	$df_{\text{denominator}}$	F	p
Stage	3	12	2.45	0.1138
Treatment	6	13.5	74.61	<0.0001
Covariance parameter	Estimate			
Block	367.98			
Residual	1140.68			

Assessing Treatment Effects in the Light of Observed Covariate Scores

Littell, Milliken, Stroup, and Wolfinger [9, pp. 187–201] provide extended examples of SAS PROC MIXED analysis of BIBD data for which measures of a covariate also are available. Educators also will be interested in their [9, pp. 201–218] PROC MIXED **analyses of covariance** for data from two split-plot experiments on the effectiveness of different teaching methods with years of teacher's experience as a covariate in one study and pupil IQ as a covariate in the other.

Construction of Balanced Incomplete Block Designs

We delayed this topic until now in order to take into account variations of BIBD discussed above.

Finding all possible designs for a given set of b , k , r , t , and λ is a problem in algebraic **combinatorics**. Cox and Reid [3, pp. 72–73] provide a recent summary of some possible designs in the range $k = 2$ to 4, $t = 3$ to 16, $b = 3$ to 35, and $r = 2$ to 10. See also [7, pp. 221–223, pp. 268–287] for theory and an introduction to early and relevant journal articles, as well as [1, pp. 269–275; 5, p. 74, Table XVII].

Randomization of blocks or possibly also of plots in a BIBD is in principle good experimental design practice [3, pp. 79–80, pp. 252–253]. We now consider the number of possible BIBD designs in the case of very small experiments with $t = 3$ treatments, $b = 3$ blocks, and $k = 2$ plots per block, controlling the set from which a random experiment must be selected.

Case 1. Here is a tiny BIBD with three two-treatment blocks containing Treatments A B, B C, and C A, respectively. Clearly there are $3! = 6$ possible permutations of the three blocks independently of position ordering in the blocks themselves. So a reasonable selection of a BIBD of this structure would choose randomly from 6 options.

Case 2. The examples in Case 1 above are also Youden squares because $t = b$. Suppose we have an auxiliary variable for plots in a block like the 'blocks (position)' of Table 3 or the stage number. Let its values be α and β . Given the same structure of $t = b = 3$, $k = 2$, $r = 2$, and $\lambda = 1$ as before, there are three pairs of ordinary letters that must partially

define our three blocks: A B, A C, and B C. With each pair there are two orderings of the Greek letters. So we can have $(A\alpha B\beta)$ or $(A\beta B\alpha)$ for an A B block. If we do not require orthogonality of treatments and the auxiliary variable, there are $2^3 = 8$ orderings of Greek pairs for the three blocks. But the three blocks may be permuted in six ways, yielding a total of 48 possible designs for this specific Youden square example. With this many options for a simple design or, even better with larger designs with many options, random selection of a specific set of t blocks has the further possible advantage of permitting use of randomization theory in significance testing.

Case 3. Let us modify Case 2 to require orthogonality of treatments and the auxiliary variable. With that constraint, there are two canonical Youden squares for the $t = b = 3$, $k = 2$, $r = 2$, and $\lambda = 1$ case: $(A\alpha B\beta, B\alpha C\beta, \text{ and } C\alpha A\beta)$ or $(B\alpha A\beta, C\alpha B\beta, \text{ and } A\alpha C\beta)$. Each canonical set has 6 permutations of its 3 blocks, yielding a total of 12 possible Youden square designs from which to sample.

Miscellaneous Topics: Efficiency and Resolvable Designs

Efficiency relates most clearly to the variance of a contrast such as between two treatment effects, $\tau_A - \tau_B$. The reference variance is the variance between such effects in a complete design such as a two independent groups' t Test or a two-group comparison in a standard randomized block design with every treatment present in every block. Cox and Reid [3, pp. 78–79] define the efficiency factor of a BIBD as:

$$\varepsilon = \frac{t(k-1)}{(t-1)k}, \quad (21)$$

being less than 1 for all incomplete designs with at least $t = 2$ treatments and $t > k$ plots (units) in a block. But ε is not enough to define efficiency. Quite possibly, the error variance of scores for the different units of a block of t units, σ_t^2 , is different from the error variance of scores from a block with k units, σ_k^2 . Therefore, the efficiency of a BIBD compared to an ordinary randomized block design takes into account these variances as well as ε :

$$\text{Efficiency} = \varepsilon \frac{\sigma_t^2}{\sigma_k^2}. \quad (22)$$

8 **Balanced Incomplete Block Designs**

A *resolvable BIBD* is one in which separate complete analyses of each replication of the data are possible, permitting comparison of the r replicates. In each block of the experiment, each treatment appears exactly once. An analysis of variance for a resolvable design may split its sum of squares for blocks into a sum of squares for replicates and a sum of squares for blocks within replicates [3, pp. 73–74; 7, pp. 226–227].

References

- [1] Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*, Wiley, New York.
- [2] Box, G.E.P. & Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading.
- [3] Cox, D.R. & Reid, N. (2000). *The Theory of the Design of Experiments*, Chapman & Hall/CRC, Boca Raton.
- [4] Draper, N.R. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Edition, Wiley, New York.
- [5] Fisher, R.A. & Yates, F. (1953). *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th Edition, Hafner, New York.
- [6] John, P.W.M. (1961). An application of a balanced incomplete block design, *Technometrics* **3**, 51–54.
- [7] John, P.W.M. (1971). *Statistical Design and Analysis of Experiments*, Macmillan, New York.
- [8] Jones, B. & Kenward, M.K. (2003). *Design and Analysis of Cross-over Trials*, 2nd Edition, Chapman & Hall/CRC, London.
- [9] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS® System for Mixed Models*, SAS Institute Inc., Cary.
- [10] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury, Pacific Grove.

JOHN W. COTTON

Bar Chart

BRIAN S. EVERITT

Volume 1, pp. 125–126

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bar Chart

A bar chart is a graphical display of data that have been classified into a number of categories. Equal-width rectangular bars are used to represent each category, with the heights of the bars being proportional to the observed frequency in the corresponding category. An example is shown in Figure 1 for the age of marriage of a sample of women in Guatemala.

An extension of the simple bar chart is the component bar chart in which particular lengths of each bar are differentiated to represent a number of frequencies associated with each category forming the chart. Shading or color can be used to enhance the display. An example is given in Figure 2; here the numbers of patients in the four categories of a response variable for two treatments (BP and CP) are displayed.

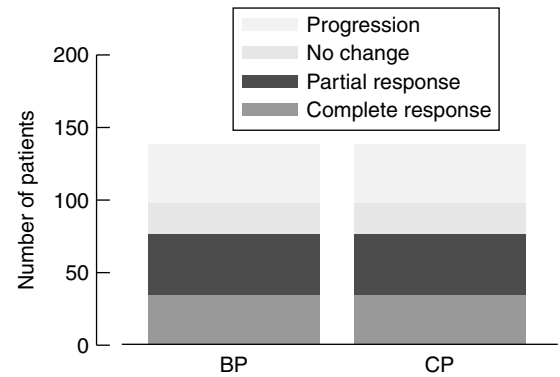


Figure 2 Component bar chart for response to treatment

Data represented by a bar chart could also be shown as a **dot chart** or a **pie chart**. A bar chart is the categorical data counterpart to the **histogram**.

BRIAN S. EVERITT

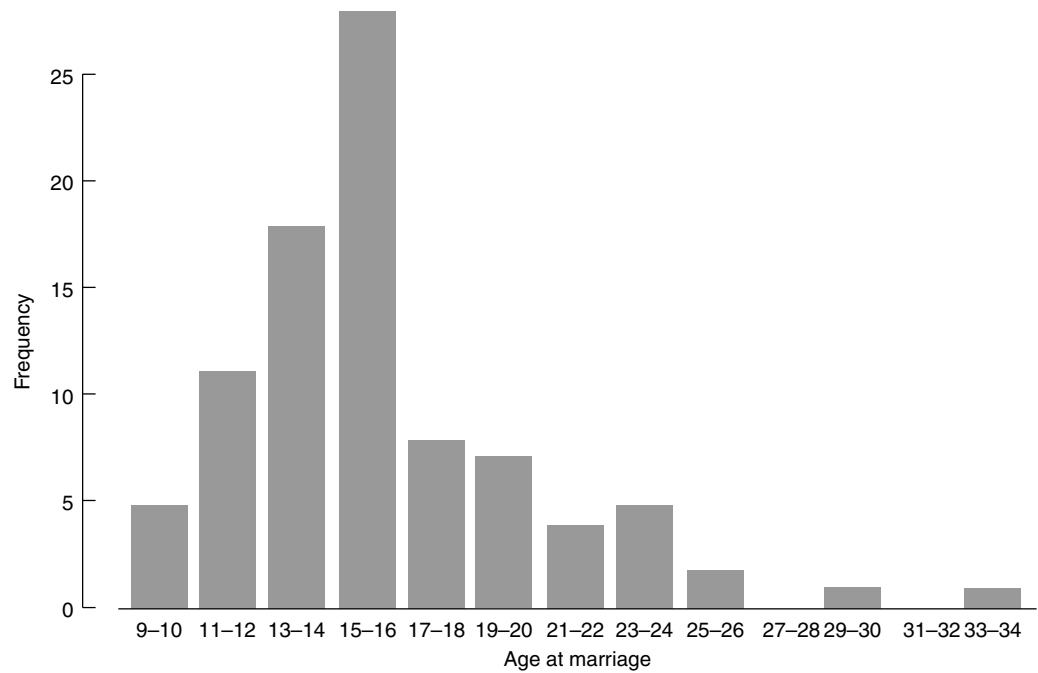


Figure 1 Bar chart for age of marriage of women in Guatemala

Battery Reduction

JOSEPH M. MASSARO

Volume 1, pp. 126–129

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Battery Reduction

Often a researcher is in the position where he has n variables of interest under investigation, but desires to reduce the number for analysis or for later data collection. Specifically, a researcher may desire to select a subset of m variables from the original n variables that reproduce as much of the information as possible, contained in the original n variables. In other words, he may desire to find the subset of m variables which accounts for a large proportion of the variance of the original n variables. For example, if he has a long questionnaire measuring the effect of a given treatment on the day-to-day activities of a certain population of patients, there may be concern about the burden such a questionnaire places upon the patient. So there is a need to try to reduce the size of the questionnaire (or reduce the battery of questions) without substantially reducing the information obtained from the full questionnaire. To accomplish this, he can perform battery reduction using the data collected from patients who completed the full battery of questions at some time in the past.

There are a number of procedures for performing battery reduction. In the following, we illustrate the concept using Gram–Schmidt transformations. Cureton & D’Agostino [1, Chapter 12] contains complete details of this procedure. Also, D’Agostino et al. [2] have developed a macro in SAS that carries out this procedure and is available from ralph@math.bu.edu.

Assume that the n variables on which we would like to perform battery reduction are denoted $X_1, \dots, X_n$. Assume also that these n variables are standardized with mean zero and variance unity. Then the total variance explained by $X_1, \dots, X_n$, is n , the number of variables. To find the subset of m variables which will explain as much as possible the variance of $X_1, \dots, X_n$, we first perform a **principal component analysis** and decide upon the m components to be retained. These are the components that account for the salient variance in the original data set. The SAS [3] procedure PRINCOMP can be used to perform principal components analysis. The SAS [3] procedure FACTOR can also be employed (*see Software for Statistical Analyses*). Both procedures automatically standardize the variables before employing principal components. Note also that the

above-mentioned battery reduction macro created by D’Agostino et al. [2] automatically standardizes the variables and creates these components as part of its battery reduction.

Once m is determined, let $\mathbf{A}$ denote the $n \times m$ matrix in which the columns contain the correlations of $X_i, i = 1, \dots, n$, to the m principal components. Symbolically $\mathbf{A}$ is

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}. \quad (1)$$

The j th column contains the correlations of the original variables X_i to the j th component, and the sum of the squares of all the $a_{ij}, i = 1, \dots, n, j = 1, \dots, m$, of $\mathbf{A}$ equals the amount of the total variance of the original n variables that is explained by the m retained components. We refer to this as *salient variance*. In principal components analysis, $\mathbf{A}$ is referred to as the *initial component matrix*. It is also often referred to as the *initial factor matrix*. The elements of $\mathbf{A}$ are called the *loadings*. The sum of the squares of the loadings of the i th row of $\mathbf{A}$ equals the proportion of variance of $X_i, i = 1, \dots, n$, explained by the m principal components. This is called the communality of X_i , symbolized as h_i^2 .

Now, to find the subset of m variables which explains, as much as possible, the salient variance of the original n variables, we can employ the Gram–Schmidt orthogonal rotations to the $n \times m$ initial component matrix $\mathbf{A}$. The goal of the Gram–Schmidt rotation in battery reduction is to rotate $\mathbf{A}$ into a new $n \times m$ component matrix, where the variable accounting for the largest proportion of the salient variance (call this ‘variable 1’) has a nonzero loading on the first component, but zero loadings on the remaining $m-1$ components; the variable accounting for the largest proportion of residual variance (‘variable 2’), where residual variance is the portion of the salient variance which is not accounted for by the variable 1, has a nonzero loading on the first two components, but zero loadings on the remaining $m-2$ components; the variable accounting for the largest proportion of second-residual variance (‘variable 3’) has a nonzero loading on the first three components, but zero loadings on

2 Battery Reduction

the remaining $m-3$ components, and so on, until the variable accounting for the largest proportion of the $(m-1)$ th residual variance ('variable m ') is found. Variables 1 through m are then the variables which reproduce, as much as possible, the variance retained by the m principal components, and so also the salient variance contained in the original n variables. In the vocabulary of principal components analysis, variable 1 is the first *transformed component*, variable 2 is the second, and so on. To determine how much of the original variance of all n variables is explained by the m transformed components, we simply compute the sum of squares of all the loadings in the final $n \times m$ Gram-Schmidt rotated matrix (this should be close to the sum of squares of the elements of the $n \times m$ initial component matrix $\mathbf{A}$). The following example will illustrate the use of the Gram-Schmidt process in battery reduction.

In the Framingham Heart Study, a 10-question depression scale was administered (so $n = 10$), where the responses were No or Yes to the following (the corresponding name to which each question will hereafter be referred is enclosed in parentheses):

1. I felt everything I did was an effort (EFFORT).
2. My sleep was restless (RESTLESS).
3. I felt depressed (DEPRESS).
4. I was happy (HAPPY).
5. I felt lonely (LONELY).
6. People were unfriendly (UNFRIEND).
7. I enjoyed life (ENJOYLIF).
8. I felt sad (FELTSAD).
9. I felt that people disliked me (DISLIKED).
10. I could not get going (GETGOING).

A Yes was scored as 1 and No as 0 except for questions 4 and 7, where this scoring was reversed so that a score of 1 would indicate depression for all questions.

After performing a principal components analysis on this data, there were three components with variances greater than unity. The variances of these three components were 3.357, 1.290, and 1.022 for a percentage variance explained equal to $100 \times (3.357 + 1.290 + 1.022)/10 = 56.69\%$. Thus, using the Kaiser rule for selecting the number of retained components [1], we set m equal to 3 for this example. The 10×3 initial component matrix $\mathbf{A}$ is in Table 1.

Table 1 Initial component matrix $\mathbf{A}$ for Framingham Heart Study depression questionnaire

	a_1	a_2	a_3	h^2
EFFORT	0.60	0.15	0.41	0.55
RESTLESS	0.39	0.07	0.55	0.46
DEPRESS	0.77	-0.13	-0.10	0.62
HAPPY	0.70	-0.23	-0.06	0.55
LONELY	0.64	-0.23	-0.21	0.51
UNFRIEND	0.35	0.68	-0.33	0.69
ENJOYLIF	0.52	-0.27	-0.27	0.42
FELTSAD	0.71	-0.22	-0.20	0.59
DISLIKED	0.34	0.72	-0.22	0.68
GETGOING	0.58	0.20	0.47	0.60

Note: $h^2 = a_1^2 + a_2^2 + a_3^2$ is the communality.

Now, to use Gram-Schmidt transformations to determine the three *variables* which explain the largest portion of the salient variance from the original 10 variables, we do the following:

1. Find, from $\mathbf{A}$ in Table 1, the variable which explains the largest proportion of salient variance from the original 10 variables. This is the variable UNFRIEND, with a sum of squares of loadings (communality) across the three components equal to $0.35^2 + 0.68^2 + (-0.33)^2 = 0.69$.
2. Take the loadings of UNFRIEND from Table 1 (0.35, 0.68, -0.33) and normalize them (i.e., divide each element by the square root of the sum of the squares of all three elements). This yields the normalized loadings: 0.42, 0.82, -0.40.
3. Create a 3×3 ($m \times m$) matrix $\mathbf{Y}_1$, which, in the Gram-Schmidt process, is given by

$$\mathbf{Y}_1 = \begin{bmatrix} a & b & c \\ k_2 & -ab/k_2 & -ac/k_2 \\ 0 & c/k_2 & -b/k_2 \end{bmatrix}, \quad (2)$$

where $a = 0.42$, $b = 0.82$, $c = -0.40$ (the normalized row of UNFRIEND from $\mathbf{A}$), and $k_2 = (1 - a^2)^{1/2}$. Thus,

$$\mathbf{Y}_1 = \begin{bmatrix} 0.42 & 0.82 & -0.40 \\ 0.91 & -0.38 & 0.18 \\ 0 & -0.44 & -0.90 \end{bmatrix}. \quad (3)$$

4. Calculate $\mathbf{A}\mathbf{Y}_1'$, which is shown in Table 2. Note that, for UNFRIEND, the only nonzero loading

Table 2 $\mathbf{B} = \mathbf{A}\mathbf{Y}'_1$

	b_1	b_2	b_3	res. h^2
EFFORT	0.21	0.56	-0.44	0.51
RESTLESS	0.00	0.43	-0.53	0.47
DEPRESS	0.26	0.73	0.15	0.56
HAPPY	0.13	0.71	0.15	0.52
LONELY	0.16	0.63	0.29	0.48
UNFRIEND	0.84	0.00	0.00	0.00
ENJOYLIF	0.11	0.53	0.36	0.41
FELTSAD	0.20	0.69	0.28	0.55
DISLIKED	0.82	0.00	-0.12	0.01
GETGOING	0.22	0.54	-0.51	0.55

Note: res. h^2 = residual communality = $b_2^2 + b_3^2$.

is on the first component (or first column). This loading is equal to the square root of the sum of squares of the original loadings of UNFRIEND in matrix $\mathbf{A}$ (thus, no 'information' explained by UNFRIEND is lost during the rotation process). For each of the remaining variables in Table 2, we have the following: (i) the squares of the elements in the first column are the portions of the variances of these variables which are accounted for by UNFRIEND; and (ii) the sum of the squares of the elements in the second and third columns is the residual variance (i.e., the variance of the variables not accounted for by UNFRIEND).

5. Find the variable which explains the largest proportion of residual variance (i.e., has the largest residual communality). This is the variable DEPRESS, with a sum of squares of loadings across the last two columns of Table 2 which is equal to $0.73^2 + 0.15^2 = 0.56$.
6. Take the loadings of DEPRESS from Table 2 (0.73, 0.15) and normalize them. This yields the normalized loadings: 0.98, 0.20.
7. Create a 2×2 matrix $\mathbf{Y}_2$, which, in the Gram-Schmidt process, is given by

$$\mathbf{Y}_2 = \begin{bmatrix} b & c \\ c & -b \end{bmatrix}, \quad (4)$$

where $b = 0.98$, $c = 0.20$ (the normalized row of DEPRESS from the last two columns of Table 2). Thus,

$$\mathbf{Y}_2 = \begin{bmatrix} 0.98 & 0.20 \\ 0.20 & -0.98 \end{bmatrix}. \quad (5)$$

Table 3 Final rotated reduced component matrix, $\mathbf{C}$

	c_1	c_2	c_3	h^2
EFFORT	0.21	0.46	0.54	0.55
RESTLESS	0.00	0.31	0.61	0.46
DEPRESS	0.26	0.75	0.00	0.63
HAPPY	0.13	0.73	0.00	0.55
LONELY	0.16	0.68	-0.16	0.51
UNFRIEND	0.84	0.00	0.00	0.70
ENJOYLIF	0.11	0.59	-0.25	0.42
FELTSAD	0.20	0.73	-0.14	0.59
DISLIKED	0.82	-0.02	0.12	0.67
GETGOING	0.22	0.43	0.61	0.60

Note: $h^2 = c_1^2 + c_2^2 + c_3^2$ is the final communality.

8. Postmultiply the last two columns of $\mathbf{A}\mathbf{Y}'_1$ by $\mathbf{Y}'_2$; the result is shown in the last two columns of Table 3. The first column of Table 3 is the first column of $\mathbf{A}\mathbf{Y}'_1$. Together, the three columns are called the rotated reduced component matrix (matrix $\mathbf{C}$ of Table 3).

Note that, for DEPRESS, the loading on the last component (or last column) is zero. The sum of squares of the loadings (the final communality) of DEPRESS in Table 3 is, within rounding error, equal to the square root of the sum of squares of the loadings of DEPRESS in the initial component matrix $\mathbf{A}$ (0.63 vs. 0.62; thus, no 'information' explained by DEPRESS is lost during the rotation process). For the remaining variables in the second column of Table 3, the elements are the portions of the variances of these variables which are accounted for by DEPRESS.

9. The last of the three variables which explains the largest portion of variance in the original 10 variables is GETGOING, since its loading is largest in the last column of Table 3.
10. The sum of squares of all the loadings in Table 3 is approximately equal, within rounding error, to the sum of squares of loadings in $\mathbf{A}$.

Thus the three variables UNFRIEND, DEPRESS, and GETGOING alone retain approximately the same variance that was retained by the first three principal components (which involved all 10 original variables). We have reduced the original battery of 10 questions to three.

The above is presented only as an illustration. It is unlikely that a researcher would need to perform

4 Battery Reduction

a battery reduction on 10 simple items such as those in the example. However, there could be a tremendous gain if the original n was, say, 100 and the number of retained components m was only 10.

Also, the above example focused on finding the m variables that reproduce the variance retained by the principal components. There may be variables with low communalities (thus not related to the other variables). The researcher may want to retain these also. For a discussion of this and presentations of other battery reduction methods, see Cureton & D'Agostino [1, Chapter 12].

References

- [1] Cureton, O. & D'Agostino, R.B. (1983). *Factor Analysis: An Applied Approach*, Lawrence Erlbaum, Hillsdale.
- [2] D'Agostino, R.B., Dukes, K.A., Massaro, J.M. & Zhang, Z. (1992). in *Proceedings of the Fifth Annual Northeast SAS Users Group Conference*, pp. 464–474.
- [3] SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Release 6.04*, 4th Edition, SAS Institute, Cary.

JOSEPH M. MASSARO

Bayes, Thomas

BRIAN S. EVERITT

Volume 1, pp. 129–130

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bayes, Thomas

Born: 1701.

Died: April 17, 1761, London, UK.

Thomas Bayes was the son of a nonconformist minister, Joshua Bayes (1671–1746). The date of Bayes's birth is uncertain, but he was probably born in 1701; he died on 17th April 1761 and was buried in Bunhill Fields in the City of London in the same grave as his father. Little is known about his life except that he received an education for the ministry and was, for most of his life, an ordained nonconformist minister in Tunbridge Wells. He was known as a skilful mathematician, but none of his works on mathematics was published in his lifetime. Despite not having published any scientific work, Bayes was elected a Fellow of the Royal Society in 1742. The certificate proposing him for election, dated 8th April 1742, reads

The Revd Thomas Bayes of Tunbridge Wells, Desiring the honour of being selected into this Society,

we propose and recommend him as a Gentleman of known merit, well skilled in Geometry and all parts of Mathematical and Philosophical learning, and every way qualified to be a valuable member of the same.

The certificate is signed by Stanhope, James Burrow, Martin Folkes, Cromwell Mortimer, and John Eames.

Today, Bayes is remembered for a paper that his friend Richard Price claimed to have found among his possessions after his death. It appeared in the Proceedings of the Royal Society in 1763 and has often been reprinted. It is ironic that the work that assured his fame (at least among statisticians), the posthumously published 'Essay toward solving a problem in the doctrine of chance', was ignored by his contemporaries and seems to have little or no impact on the early development of statistics. The work contains the quintessential features of what is now known as Bayes's Theorem (*see* **Bayesian Belief Networks**), a procedure for the appropriate way to combine evidence, which has had and continues to have a major influence in modern statistics.

BRIAN S. EVERITT

Bayesian Belief Networks

LAWRENCE D. PHILLIPS

Volume 1, pp. 130–134

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bayesian Belief Networks

Bayesian Belief Networks (BBNs) represent knowledge in a directed graph consisting of conditional probabilities (*see Probability: An Introduction*). These probabilities capture both expert judgement and data in a form that makes possible the drawing of valid inferences, which can inform decision making. BBNs describe how knowledge of one event can change the uncertainty associated with related events. In this way, BBNs characterize both causal relationships between events and what can be inferred from diagnostic information.

A brief survey of the laws of probability will aid understanding of BBNs. First to be given is the definition of a conditional probability. The probability of E given F , $p(E|F)$, represents the conditioning of uncertainty about one thing on something else. For example, in scientific inference, $p(H|D)$ could represent your uncertainty about some hypothesis, H , that interests you, given the data, D , you collected in an experiment. In a business application, $p(S|M)$ might represent the probability of a certain level of sales, S , given a successful marketing campaign, M .

Next, the multiplication law of probability shows how probabilities of two events combine to give the probability of the joint event: $p(E \text{ and } F) = p(E) \times p(F|E)$. The probability you will purchase the ticket, event E , and win the lottery, event F , is equal to the probability you will purchase the ticket times the probability of winning the lottery given that you purchased the ticket. In general, if $p(F)$ is judged to be equal to $p(F|E)$, then the two events are considered to be independent; knowledge about E has no effect on uncertainty about F . Then $p(E \text{ and } F) = p(E) \times p(F)$, clearly not the case for this example.

The addition law applies to mutually exclusive events: $p(A \text{ or } B) = p(A) + p(B)$. The probability that team A will win or that team B will win equals the sum of those two individual probabilities, leaving some probability left over for a draw. Next, the probabilities of mutually exclusive and exhaustive events must sum to one, with the probability of the null event (the impossible event) equal to zero.

Combining those laws makes it possible to 'extend the conversation' so that uncertainty about F can include both E and $\sim E$, 'not- E '. Suppose that $\sim E$ represents the probability that your partner purchases a ticket, and that F represents the probability that one of you wins. Then, $p(F) = p(F|E)p(E) + p(F|\sim E)p(\sim E)$.

The mathematics can be shown in the event tree of Figure 1, and in more compact form as a BBN. The tree shows the probabilities on the branches, and because of the conditioning, they are different on the upper and lower branches. The joint probabilities shown at the right of the tree give the product of the two probabilities on the branches. It is the sum of two of those joint probabilities that make up the probability of event F , and of the other two for event $\sim F$, the probability that neither of you wins. The arrow, or arc, connecting the two circles in the belief network indicates that uncertainty about winning is conditional on purchasing the ticket, without explicitly showing the probabilities.

An example shows the application of these laws to BBNs. Assume that the weather is a major contributor to your mood, and consider the event that by noon tomorrow you will be in a grouchy mood. You are uncertain about both your mood and the weather, so a BBN would look like Figure 2.

This compact representation shows the two events, with the arrow representing the conditioning of mood on weather. In BBNs, an arrow often shows the

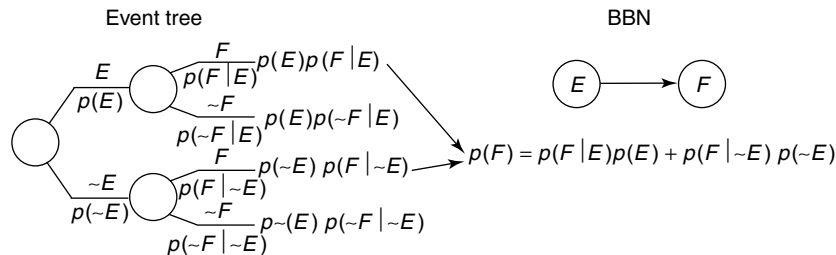


Figure 1 An event tree with its corresponding BBN

2 Bayesian Belief Networks



Figure 2 A BBN for the influence of weather on mood

Weather		Mood	
Clear	70.0	Pleasant	59.0
Rainy	30.0	Grouchy	41.0

Figure 3 The expanded BBN for the effect of weather on mood

direction of causality, though it can also represent simple relevance, in that knowing about one event reduces uncertainty about the other, even though they are not causally related.

Now assume that weather can be either clear or rainy, and your mood either pleasant or grouchy. It's a moderately dry time of year, but you aren't sure about tomorrow, so you look in an almanac to find that it has been clear on tomorrow's date 70% of the time. Knowing how you react to poor weather, and that other causes can lead to grouchiness, you judge that if the weather is clear, your probability of becoming grouchy is 0.2, but if it is rainy, the probability of grouchiness is 0.9. A more complete representation is shown in Figure 3.

The 70 to 30 repeats the input data, but where does the 59 to 41 come from? The probability of

grouchiness is only 20% sure if the weather is clear, and 90% sure if it rains, so those two probabilities are weighted with the 70 to 30 weather probabilities, and those two products are summed, to give 41% as can be seen in Figure 4.

Your lack of certainty about tomorrow's weather bothers you, so you consult the local weather forecast. You also know the research on the reliability of weather forecasts: in your area when they say 'clear', they are right about 85% of the time, and when they say 'rain', they are right about 75% of the time. Modifying your BBN to include these new data gives the representation shown in Figure 5.

Note that the arrow between Weather and Forecast follows the direction of causality; it is the weather that gives rise to the forecast, not the other way around. The Clear-Rain forecast probabilities of 67 to 33 are the result of applying calculations like those that led to 59 to 41 for mood.

You now check the forecast for tomorrow; it is 'clear'. Changing the probability of clear to 100, which represents that 'clear' was definitely *forecast* rather than 'rain', gives the result in Figure 6.

In light of this forecast, there is now an 88.8% chance of clear tomorrow, and a 27.8% chance of your being in a grouchy mood. The latter probability depends on the weather probabilities, as in Figure 3. But where does the 88.8% come from? Calculating the probabilities for the weather given the forecast requires application of *Bayes's theorem*.

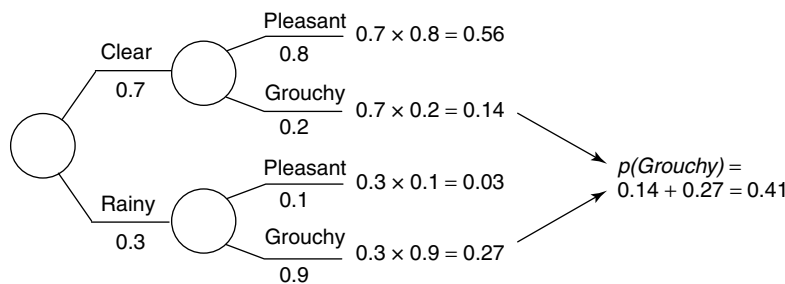


Figure 4 Calculations for the probability of being grouchy for the weather on mood BBN

Forecast		Weather		Mood	
Clear	67.0	Clear	70.0	Pleasant	59.0
Rain	33.0	Rainy	30.0	Grouchy	41.0

Figure 5 The expanded BBN to take account of the weather forecast, with its less-than-perfect reliability

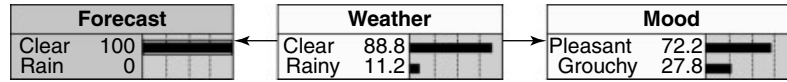


Figure 6 The effect on uncertainty about weather and mood of a ‘clear’ forecast

Let H , a hypothesis, stand for tomorrow’s weather, and D , data, today’s forecast. Bayes’ theorem provides the basis for calculating $p(H|D)$ from the inverse probability, which we know, $p(D|H)$, representing the forecaster’s reliability. Bayes’s theorem is a simple consequence of the above probability laws, with the added recognition that $p(H \text{ and } D)$ must equal $p(D \text{ and } H)$; the order in which H and D are written down makes no difference to their joint probability. Since $p(H \text{ and } D) = p(H|D) \times p(D)$, and $p(D \text{ and } H) = p(D|H) \times p(H)$, then equating the right hand sides of the equations, and rearranging terms, gives Bayes’s theorem:

$$p(H|D) = \frac{p(H) \times p(D|H)}{p(D)} \quad (1)$$

posterior probability = prior probability $\times$ likelihood/probability of the data (see **Bayesian Statistics**).

This result can be shown by ‘flipping’ the original event tree, but Bayes’s theorem is easier to apply in tabular form; see Table 1. Recall that the datum, D , is the forecaster’s prediction today, ‘clear’, and H is the weather that will be realized tomorrow.

Note that D , the forecaster’s ‘clear’, stays the same in the table, while the hypothesis H , next day’s weather, changes, ‘Clear’ in the first row and ‘Rainy’ in the second.

The unreliability of the forecast has increased your original assessment of a 20% chance of grouchiness, if the actual weather is clear, to a 27.8% chance if the forecast is for clear. And by changing the rain

Table 1 Application of Bayes’s theorem after receipt of the forecast ‘clear’

Weather H	Prior $p(H)$	Likelihood $p(D H)$	$P(H) \times p(D H)$	Posterior $P(H D)$
Clear	0.70	0.85	0.595	0.888
Rainy	0.30	0.25	0.075	0.112
Sum = $P(D) = 0.67$				

forecast probability to 100% in the BBN (Figure 7), your chance of grouchy becomes 67.7%, rather less than your original 90%, largely because the forecasts are less reliable for ‘rain’ than they are for ‘clear’.

So, applying the laws of probability, the multiplication and addition laws operating in one direction, and Bayes’s theorem applied in the other direction, allows information to be propagated throughout the network. Suppose, for example, that several days later you recall being in a pleasant mood on the day in question, but can’t remember what the weather was like. Changing the probability of your mood to 100 for ‘pleasant’ in the BBN gives the result shown in Figure 8.

The chance of good weather, while in reality now either zero or 100%, is, for you at this moment, nearly 95%, and the chance of a ‘clear’ forecast the day before, about 82%. In summary, propagating information in the direction of an arrow requires application of the multiplication and addition laws of probability, whereas propagating information against an arrow’s direction invokes Bayes’s theorem.

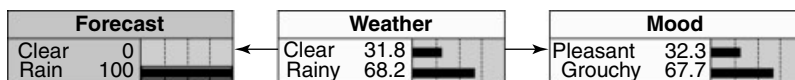


Figure 7 The effect on uncertainty about weather and mood of a forecast of ‘rain’

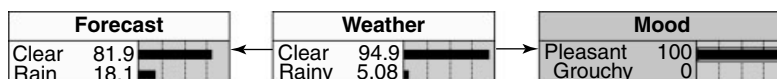


Figure 8 Inferences about the weather forecast and subsequently realized weather if a pleasant mood is all that can be recollected for the day of the original inference

The directed graph indicates conditional dependence between events, with missing links in the graph showing independence. Thus, the lack of an arc between forecast and mood shows that the forecast has no impact on your uncertainty about tomorrow's mood. For this simple problem, lack of arcs is trivial, but for complex BBNs consisting of tens or hundreds of nodes, the presence or absence of arcs provides a compact display that allows a user to grasp quickly the structure of the representation. A corresponding event tree could only be displayed on a computer in small sections or on a very large printed surface, and even then the structure would not be easily grasped even with the probabilities displayed on the corresponding branches.

BBNs break the problem down into many relatively simple probability statements, and from these new insights can emerge. It is this property that was recognised early by psychologists, who first developed the fundamental idea of using human expertise to provide the probabilistic inputs [1, 3, 4]. Their studies initially assumed that data were reliable, though not definitive, in pointing to the correct hypothesis, and their systems assumed a single level of inference, from reliable datum to the hypothesis of interest. Studies comparing actual human inferences to the properties of Bayes's theorem led to the surprising conclusion that in general people do not revise their uncertainty as much as is prescribed by Bayesian calculations, a replicable phenomenon called 'conservatism' by the psychologists who discovered it [2, 14].

But real-world data are often unreliable, ambiguous, redundant, or contradictory, so many investigators developed 'cascaded inference' models to accommodate data unreliability and intermediate levels of uncertainty. Examples abound in medical diagnosis: unreliable data (reports of symptoms from patients) may point to physical conditions (signs only observable from tests) that in turn bear on hypotheses of interest (possible disease states). Comparing actual unaided inferences with these cascaded inference models, as reported in a special issue of *Organisational Behavior and Human Performance* [13], showed occasions when people became less certain than Bayesian performance, but other occasions when they were over confident. Sometimes they assumed unreliable data were reliable, and sometimes they ignored intermediate levels of inference. Although another psychologist, David Schum, picked

up these ideas in the 1960s and studied the parallels between legal reasoning and Bayesian inference [16] the agenda for studying human judgement in the face of uncertainty took a new turn with studies of heuristics and biases (*see* **Heuristics: Fast and Frugal; Subjective Probability and Human Judgement**).

The increasing availability of convenient and substantial computer power saw the growth of BBNs from the mid-1980s to the present day. This growth was fuelled by developments in decision analysis and artificial intelligence [8, 11]. It became feasible to apply the technology to very complex networks [5], aided by computer programs that facilitate structuring and entry of data, with the computational complexity left to the computer (the 'mood' model, above, was constructed using Netica [10]). For complex models, special computational algorithms are used, variously developed by Schachter [15], Lauritzen and Spiegelhalter [7], Pearl [12] and Spiegelhalter and Lauritzen [17]. Textbooks by Jensen [6] and Neapolitan [9] provide guidance on how to construct the models.

BBNs are now in widespread use in applications that require consistent reasoning and inference in situations of uncertainty. They are often invisible to a user, as in Microsoft's help and troubleshooting facilities, whose behind-the-scene BBNs calculate which questions would be most likely to reduce uncertainty about a problem. At other times, as in medical diagnostic systems, the probabilistic inferences are displayed. A web search on BBNs already displays tens of thousands of items; these are bound to increase as this form of rational reasoning becomes recognised for its power to capture the knowledge of experienced experts along with hard data, and make this available in a form that aids decision making.

References

- [1] Edwards, W. (1962). Dynamic decision theory and probabilistic information processing, *Human Factors* **4**, 59–73.
- [2] Edwards, W. (1968). Conservatism in human information processing, in *Formal Representations of Human Judgment*, B. Kleinmuntz, ed., Wiley, New York, pp. 17–52.
- [3] Edwards, W. (1998). Hailfinder: tools for and experiences with Bayesian normative modeling, *American Psychologist* **53**, 416–428.
- [4] Edwards, W., Phillips, L.D., Hays, W.L. & Goodman, B. (1968). Probabilistic information processing systems:

- design and evaluation, *IEEE Transactions on Systems Science and Cybernetics*, **SSR-4**, 248–265.
- [5] Heckerman, D., Mamdani, A. & Wellman, M.P. (1995). Special issue: real-world applications of Bayesian networks, *Communications of the ACM* **38**, 24–57.
 - [6] Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*, Springer-Verlag.
 - [7] Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society, Series B* **50**, 157–224.
 - [8] Matzkevich, I. & Abramson, B. (1995). Decision analytic networks in artificial intelligence, *Management Science* **41**(1), 1–22.
 - [9] Neapolitan, R.E. (2003). *Learning Bayesian Networks*, Prentice Hall.
 - [10] Netica Application APL, DLL and Users Guide (1995–2004). *Norsys Software Corporation*, Vancouver, Download from www.norsys.com.
 - [11] Oliver, R.M. & Smith, J.Q., eds (1990). *Influence Diagrams, Belief Nets and Decision Analysis*, John Wiley & Sons, New York.
 - [12] Pearl, J. (1988). *Probabilistic Reasoning in Expert Systems*, Morgan Kaufmann, San Mateo.
 - [13] Peterson, C.R. ed. (1973). Special issue: cascaded inference, *Organizational Behavior and Human Performance* **10**, 315–432.
 - [14] Phillips, L.D., Hays, W.L. & Edwards, W. (1966). Conservatism in complex probabilistic inference, *IEEE Transactions on Human Factors in Electronics*, **HFE-7**, 7–18.
 - [15] Schachter, R.D. (1986). Evaluating influence diagrams, *Operations Research* **34**(6), 871–882.
 - [16] Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning*, John Wiley & Sons, New York.
 - [17] Spiegelhalter, D. & Lauritzen, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures, *Networks* **20**, 579–605.
- (See also **Markov Chain Monte Carlo and Bayesian Statistics**)

LAWRENCE D. PHILLIPS

Bayesian Item Response Theory Estimation

HARIHARAN SWAMINATHAN

Volume 1, pp. 134–139

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bayesian Item Response Theory Estimation

The problem of fundamental importance in **item response theory** is the accurate estimation of the parameters that characterize the items and the ability or proficiency levels of examinees. Likelihood-based approaches such as joint, conditional, and marginal maximum likelihood procedures provide reasonable estimates of the item parameters (*see* **Maximum Likelihood Item Response Theory Estimation; Maximum Likelihood Estimation, and Item Response Theory (IRT) Models for Dichotomous Data** for more details). However, more often than not these procedures run into Heywood-type problems and yield estimates in the two- and three-parameter models that are inadmissible in the sense that the estimates of the ability and item parameters fall outside acceptable ranges. Bayesian approaches, by taking into account prior and collateral information, often overcome these problems encountered by likelihood-based approaches (*see* **Bayesian Statistics**).

The Bayesian Framework

At the heart of the Bayesian approach is the well-known Bayes Theorem (*see* **Bayesian Belief Networks**), which provides the well-known relationship among conditional probabilities (*see* **Probability: An Introduction**),

$$\pi(A|B) = \frac{\pi(B|A)\pi(A)}{\pi(B)}, \quad (1)$$

where $\pi(y)$ denotes probability when y is discrete and the probability density function of y when y is continuous. If we denote the vector of unknown parameters γ by A and the observations or Data by B , then,

$$\pi(\gamma|Data) = \frac{\pi(Data|\gamma)\pi(\gamma)}{\pi(Data)}. \quad (2)$$

The expression, $\pi(\gamma|Data)$, is the *posterior density* of γ ; $\pi(Data|\gamma)$ is the joint probability or the probability density of the observations. Once the observations are realized, it ceases to have a probabilistic interpretation and is known as the *likelihood*

function, usually written as $L(Data|\gamma)$; $\pi(\gamma)$ is the *prior density* of the parameters, and is an expression of the information or prior belief a researcher may have about the parameters; $\pi(Data)$ is a function of the observations and hence is a constant, determined so that the posterior density has unit mass, that is, $\pi(Data) = 1 / \int L(Data|\gamma)\pi(\gamma)d\gamma$. Bayes Theorem thus leads to the statement

$$\pi(\gamma|Data) = L(Data|\gamma)\pi(\gamma), \quad (3)$$

that is,

$$Posterior = Likelihood \times Prior. \quad (4)$$

Implicit in the description above is the fact that the parameters are treated as random variables. Thus, in the Bayesian approach, it is meaningful to make probabilistic statements about parameters, for example, determine the probability that a parameter will fall within an interval [3]. This is the point of departure between the Bayesian and the classical or the frequentist approach; in the frequentist approach, the parameters are considered fixed, and hence probabilistic statements about the parameters are meaningless (*see* **Bayesian Statistics; Probability: An Introduction**).

In the Bayesian framework, the posterior density contains all the information about the parameters. Theoretically, a comprehensive description of the parameters can be obtained, for example, the parameters can be described in terms of moments and percentile points. However, in the multiparameter situation, obtaining the moments and percentile points is tedious if not impossible. Consequently, point estimates of the parameters, such as the joint mode or the mean, are usually obtained.

In the context of item response theory, the parameters of interest are the item and ability parameters. In the dichotomous item response models, an item may be characterized by one parameter, the difficulty, b , or two parameters, b, a , the difficulty and discrimination parameters respectively, or three parameters, b, a, c , the difficulty, discrimination, and the pseudo chance-level parameters respectively [5, 6]. In the polytomous case (*see* **Item Response Theory (IRT) Models for Rating Scale Data**), items may be characterized by a set of the threshold/category parameters (the partial credit model) or by threshold/category and discrimination parameters (generalized partial credit model/graded response model) (*see*

2 Bayesian Item Response Theory Estimation

Item Response Theory (IRT) Models for Polytomous Response Data). The examinees are usually characterized by a single ability parameter, θ . The joint posterior density of the item and ability parameters for any one examinee is thus

$$\pi(\xi, \theta|u) = L(u|\xi, \theta)\pi(\xi, \theta), \quad (5)$$

where ξ is the vector of item parameters, θ is the ability parameter for the examinee, and $u = [u_1 u_2 \dots u_n]$ is the vector of responses to n items. The posterior density is determined up to a constant once the likelihood function, $L(u|\xi, \theta)$, is determined and the prior, $\pi(\xi, \theta)$, is specified.

The Likelihood Function

The assumption of local independence in Item Response Theory (IRT) implies that

$$\pi(u|\xi, \theta) = \pi(u_1|\xi, \theta)\pi(u_2|\xi, \theta) \dots \pi(u_n|\xi, \theta), \quad (6)$$

where $\pi(u_j|\xi, \theta)$ is specified by the item response model that is deemed appropriate. In the general case where an item is scored polytomously, with response categories $r_{1j}, r_{2j}, \dots, r_{sj}$ for item j , if we denote the probability of responding in category r_k as $P(u_j = r_{kj}|\xi, \theta) \equiv P_j^{r_{kj}}$ with $r_{kj} = 1$ or 0 , $\sum_k r_{kj} = 1$, and $\sum_k P_j^{r_{kj}} = 1$, then the probability of a response to the item can be expressed as

$$\pi(u_j|\xi, \theta) = P_j^{r_{1j}} P_j^{r_{2j}} \dots P_j^{r_{sj}} = \prod_{k=1}^s P_j^{r_{kj}}. \quad (7)$$

The joint probability of the response vector u is the product of these probabilities, and once the responses are observed, becomes the likelihood

$$L(u|\xi, \theta) = \prod_{j=1}^n \prod_{k=1}^s P_j^{r_{kj}}. \quad (8)$$

With N examinees, the likelihood function is given by

$$L(U|\xi, \theta) = \prod_{i=1}^N \prod_{j=1}^n \prod_{k=1}^s P_j^{r_{kj}}, \quad (9)$$

where U is the response vector of the N examinees on n items, and θ is the vector of ability parameters for the N examinees.

In the dichotomous case with response categories r_1 and r_2 , $r_2 = 1 - r_1$, and $\pi(u_j|\xi, \theta) = P_j^{r_{1j}} Q_j^{1-r_{1j}}$. Thus,

$$\begin{aligned} L(U|\xi, \theta) &= \prod_{i=1}^N \prod_{j=1}^n L(u_j|\xi, \theta_i) \\ &= \prod_{i=1}^N \prod_{j=1}^n P_j^{r_{1j}} Q_j^{1-r_{1j}}. \end{aligned} \quad (10)$$

Prior Specification, Posterior Densities, and Estimation

While the evaluation of the likelihood function is straightforward, the specification of the prior is somewhat complex. In IRT, the prior density, $\pi(\xi, \theta)$, is a statement about the prior belief or information the researcher has about the item and ability parameters. It is assumed *a priori* that the item and ability parameters are independent, that is, $\pi(\xi, \theta) = \pi(\xi)\pi(\theta)$.

Specification of priors for the ability and item parameters may be carried out in a single stage, or a hierarchical procedure may be employed. In the single stage procedure, a distributional form is assumed for ξ and the parameters of the distribution are specified. For example, it may be assumed that the item parameters have a multivariate normal distribution, that is, $\xi|\mu, \Sigma \sim N(\mu, \Sigma)$, and the parameters (μ, Σ) are specified. In the hierarchical procedure, distributional forms are assumed for the parameters (μ, Σ) and the *hyper-parameters* that determine the distribution of (μ, Σ) are specified. In contrast to the single stage approach, the hierarchical approach allows for a degree of uncertainty in specifying priors by expressing prior beliefs in terms of a family of prior distributions. Swaminathan and Gifford [14–16] proposed a hierarchical Bayes procedure for the *joint estimation* of item and ability parameters following the framework provided by Lindley and Smith [8]; Mislevy [9], using the same framework, provided a hierarchical procedure for the *marginal estimation* of item parameters. In the following discussion, only the three-parameter dichotomous item response model is considered since the one- and the two-parameter models are obtained as special cases. The procedures described are easily extended to the polytomous case [11, 12].

In the dichotomous case, when the three-parameter model is assumed, the vector of item parameters

consists of $3n$ parameters: n difficulty parameters, b_j , n discrimination parameters, a_j , and n pseudo chance-level parameters, c_j . While in theory, it is possible to assume a multivariate distribution for the item parameters, specification of the parameters poses some difficulty. To simplify the specification of priors, Swaminathan and Gifford [16] assumed that the sets of item parameters b, a, c are independent. They further assumed that the difficulty parameters b_j are *exchangeable* and that in the first stage, $b_j \sim N(\mu_b, \sigma_b^2)$. In the second stage, they assumed a noninformative prior for μ_b and an inverse chi-square prior with parameters ν_b, λ_b for σ^2 . For the discrimination parameters, they assumed a chi distribution with parameters ν_{aj} and ω_{aj} . Finally, for the c -parameter, they assumed a Beta distribution with parameters s_j and t_j . (see **Catalogue of Probability Density Functions**). The ability parameters are assumed to be exchangeable, and independently and identically distributed normally with mean μ_θ and variance σ_θ^2 . By setting μ_θ and σ_θ^2 as zero and one respectively, the conditions required for identifying the model may be imposed. With these assumptions, the joint posterior density of item and ability parameters after integrating the nuisance parameters, $\nu_b, \sigma_b^2, \mu_\theta, \sigma_\theta^2$, is

$$\begin{aligned} \pi(a, b, c, \theta | \nu_a, \omega_a, s, t, U) &= \int L(U | a, b, c, \theta) \\ &\times \prod_{j=1}^n \pi(a_j | \nu_{aj}, \omega_{aj}) \pi(b_j | \mu_b, \sigma_b^2) \pi(c_j | s_j, t_j) \\ &\times \prod_{i=1}^N \pi(\theta_i | \mu, \sigma^2) d\nu_b d\sigma_b^2 d\mu_\theta d\sigma_\theta^2 \end{aligned} \quad (11)$$

Swaminathan and Gifford [16] provided procedures for specifying the parameters for the prior distributions of the discrimination and the pseudo chance level parameters. Once the parameters of the priors are specified, the posterior density is completely specified up to a constant. These authors then obtained the joint posterior modes of the posterior density as the joint Bayes estimates of the item and ability parameters. Through an empirical study, Gifford and Swaminathan [4] demonstrated that the joint Bayesian procedure offered considerable improvement over the joint maximum likelihood procedure.

In contrast to the approach of Swaminathan and Gifford [16] who assumed that the examinees were

fixed, Mislevy [9], following the approach taken by Bock and Lieberman [1], assumed that the examinees were sampled at random from a population. With this assumption, the marginal joint posterior density of the item parameters is obtained as

$$\pi(\xi | U, \tau) = \int L(U | \xi, \theta) \pi(\theta) \pi(\xi | \tau) d\theta. \quad (12)$$

With the assumption that $\theta \sim N(0, 1)$, the integration is carried out using Gaussian quadrature. The advantage of this marginal Bayesian procedure over the joint Bayesian procedure is that the marginal modes are closer to the marginal means than are the joint modes. It also avoids the problem of improper estimates of structural, that is, item, parameters in the presence of an infinite number of nuisance or incidental ability parameters.

Mislevy [9], rather than specifying priors on the item parameters directly, specified priors on transformed discrimination and pseudo chance-level parameters, that is, on $\alpha_j = \log(a_j)$ and $\gamma_j = \log[c_j/(1 - c_j)]$. With $\beta_j = b_j$, the vector of parameters, $\xi_j = [\alpha_j \beta_j \gamma_j]$, was assumed to have a multivariate normal distribution with mean vector μ_j and variance-covariance matrix Σ_j (see **Catalogue of Probability Density Functions**). At the second stage, it is assumed that μ_j is distributed multivariate normally and that Σ_j has the inverted Wishart distribution, a multivariate form of the inverse chi-square distribution. Although in principle it is possible to specify the parameters of these hyper prior distributions, they present problems in practice since most applied researchers and measurement specialists lack sufficient experience with these distributions. Simplified versions of these prior distributions are obtained by assuming the item parameters are independent. In this case, it may be assumed that α_j is normally distributed, or equivalently that a_j has a lognormal distribution. The parameters of this distribution are more tractable and readily specified. With respect to the pseudo chance-level parameter, computer programs such as BILOG [10] and PARSCALE [12] use the beta prior for the pseudo chance-level parameter, as recommended by Swaminathan and Gifford [16]. A detailed study comparing the joint and the marginal estimation procedures in the case of the two-parameter item response model is provided by Kim et al. [7].

Rigdon and Tsutakawa [13], Tsutakawa [17], [18], and Tsutakawa and Lin [19] have provided alternative marginalized Bayes modal estimation procedures. The procedures suggested by Tsutakawa [18] and Tsutakawa and Lin [19] for specifying priors is basically different from that suggested by Swaminathan and Gifford and Mislevy; Tsutakawa and Lin [19] suggested an ordered bivariate beta distribution for the item response function at two ability levels, while Tsutakawa [18] suggested the ordered Dirichlet prior on the entire item response function. These approaches are promising, but no extensive research has been done to date comparing this approach with other Bayesian approaches.

More recently, the joint estimation procedure outlined above has received considerable attention in terms of **Markov Chain Monte Carlo (MCMC)** procedures. In this approach, observations are sampled from the posterior density, and with these the characteristics of the posterior density, such as the mean, variance, and so on, are approximated. This powerful technique has been widely applied in Bayesian estimation and inference and is receiving considerable attention for parameter estimation in item response models (*see Markov Chain Monte Carlo Item Response Theory Estimation*).

Estimation of Ability Parameters with Known Item Parameters

As mentioned previously, one of the primary purposes of testing is to determine the ability or proficiency level of examinees. The estimation procedure for jointly estimating item and ability parameters may be employed in this case. However, in situations such as computer-adaptive tests, joint estimation may not be possible. The alternative is to employ a two-stage procedure, where in the first stage, item parameters are estimated using the marginal Bayesian or maximum likelihood procedures, and in the second stage, assuming that the item parameters are known, the ability parameters are estimated.

The estimation of ability parameters when item parameters are known is far less complex than the procedure for estimating item parameters or jointly estimating item and ability parameters. Since the examinees are independent, it is possible to estimate

each examinee's ability separately. In this case, if the prior density of θ is taken as normal, then the posterior density of θ is

$$\pi(\theta|u, \xi, \mu, \sigma^2) = L(u|\theta, \xi)\pi(\theta|\mu, \sigma^2) \quad (13)$$

where μ and σ^2 are the mean and variance of the prior distribution of θ . The mode of the posterior density, known as the *maximum a posteriori* (MAP) estimate, may be taken as the point estimate of ability. Alternatively, the mean of the posterior density, the *expected a posteriori* (EAP) estimate [2], defined as

$$\bar{\theta} = \int_{-\infty}^{\infty} \theta \pi(\theta|\xi, \mu, \sigma^2) d\theta, \quad (14)$$

may be taken as the point estimate of θ . The integral given above is readily evaluated using numerical procedures. The variance of the estimate can also be obtained as

$$\text{Var}(\bar{\theta}) = \int_{-\infty}^{\infty} [\theta - \bar{\theta}]^2 \pi(\theta|\xi, \mu, \sigma^2) d\theta. \quad (15)$$

A problem that is noted with the Bayesian estimate of ability is that unless reasonably good prior information is available, the estimates tend to be biased.

In the case of conventional testing where many examinees respond to the same items, a hierarchical Bayesian procedure may prove to be useful. Swaminathan and Gifford [14] applied a two-stage procedure similar to that described earlier to obtain the joint posterior density of the abilities of N examinees. They demonstrated that the hierarchical Bayes procedure, by incorporating collateral information available from the group of examinees, produced more accurate estimates of the ability parameters than maximum likelihood estimates or a single-stage Bayes procedure.

References

- [1] Bock, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika* **35**, 179–197.
- [2] Bock, R.D. & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment, *Applied Psychological Measurement* **6**, 431–444.

- [3] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton.
- [4] Gifford, J.A. & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters in item response models, *Applied Psychological Measurement* **14**, 33–43.
- [5] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, Boston.
- [6] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage, Newbury Park.
- [7] Kim, S.H., Cohen, A.S., Baker, F.B., Subkoviak, M.J. & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory, *Psychometrika* **99**, 405–421.
- [8] Lindley, D.V. & Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 1–41.
- [9] Mislevy, R.J. (1986). Bayes modal estimation in item response models, *Psychometrika* **51**, 177–195.
- [10] Mislevy, R.J. & Bock, R.D. (1990). *BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models*, Scientific Software, Chicago.
- [11] Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm, *Applied Psychological Measurement* **16**, 159–176.
- [12] Muraki, E. & Bock, R.D. (1996). *PARSCALE: IRT Based Test Scoring and Item Analysis for Graded Open-Ended Exercises and Performance Tests*, Scientific Software, Chicago.
- [13] Rigdon, S.E. & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models, *Psychometrika* **48**, 567–574.
- [14] Swaminathan, H. & Gifford, J.A. (1982). Bayesian estimation in the Rasch model, *Journal of Educational Statistics* **7**, 175–191.
- [15] Swaminathan, H. & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model, *Psychometrika* **50**, 175–191.
- [16] Swaminathan, H. & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model, *Psychometrika* **51**, 581–601.
- [17] Tsutakawa, R.K. (1984). Estimation of two-parameter logistic item response curves, *Journal of Educational Statistics* **9**, 263–276.
- [18] Tsutakawa, R.K. (1992). Prior distributions for item response curves, *British Journal of Mathematical and Statistical Psychology* **45**, 51–74.
- [19] Tsutakawa, R.K. & Linn, R. (1986). Bayesian estimation of item response curves, *Psychometrika* **51**, 251–267.

HARIHARAN SWAMINATHAN

Bayesian Methods for Categorical Data

EDUARDO GUTIÉRREZ-PEÑA

Volume 1, pp. 139–146

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bayesian Methods for Categorical Data

Introduction

When analyzing data, it is often desirable to take into account available prior information about the quantity of interest. The Bayesian approach to statistical inference is based on a subjective interpretation of probability (*see* **Bayesian Statistics; Subjective Probability and Human Judgement**). Given a prior distribution describing our prior beliefs on the value of an unknown quantity (typically a model parameter), Bayes' theorem (*see* **Bayesian Belief Networks**) allows us to update those beliefs in the light of observed data. The resulting posterior distribution then summarizes all the available information about the quantity of interest, conditional on the posited model. Bayesian methods for the analysis of categorical data use the same classes of models as the classical approach. However, Bayesian analyses are often more informative – the posterior distribution of the parameter containing more information than mere point estimates or test statistics – and may provide more natural solutions in certain situations such as those involving sparse data or unidentifiable parameters.

The analysis of categorical data is, to some extent, still dominated by the classical approach under which inferences are mostly based on asymptotic theory. However, as in many other areas of statistical practice, the Bayesian literature has been growing steadily in recent years. This is mainly due to the continuous development of efficient computational tools, which make it possible to deal with more complex problems. Among other things, this has prompted the development of fully Bayesian analysis of more realistic (if problematic) situations such as those involving missing data, censoring, misclassification, measurement errors, and so on.

Here, we review some of the most common Bayesian methods for categorical data. We focus on the analysis of **contingency tables**, but other useful models are also briefly discussed. For ease of exposition, we describe most ideas in terms of two-way contingency tables. After presenting some preliminary material in the next section, in the

section titled 'Some History', we provide a brief historical account of early Bayesian approaches to the analysis of contingency tables. In the section titled 'Bayesian Inference for Multinomial Data', we review a Bayesian conjugate analysis of the multinomial distribution, which is the basis of some simple analyses of contingency tables. Then, in the section titled 'Log-linear and Generalized Linear Models', we describe a more general and widely used approach based on the class of **log-linear models**. We discuss estimation, hypothesis testing, and model selection. Finally, in the section titled 'Specialized Models', we mention other, specialized models and provide suggestions for further reading.

Preliminaries

Consider a two-way contingency table with r rows and c columns, with cell probabilities π_{ij} and observed counts n_{ij} ($i = 1, \dots, r; j = 1, \dots, c$). Let $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$ and $N = \sum_i \sum_j n_{ij}$. Three sampling schemes occur in applications:

- *Scheme 1* (multinomial sampling), where only the total N is fixed.
- *Scheme 2* (product-multinomial sampling or stratified sampling), where either the row (n_{i+}) or column (n_{+j}) totals are fixed.
- *Scheme 3*, where both the row or column totals are fixed. For 2×2 tables, this situation is related to Fisher's 'exact test' (*see* **Exact Methods for Categorical Data**) (*see* [8]).

Analogous sampling schemes occur in multi-way contingency tables when some of the various marginal totals are fixed by the experimenter. Scheme 3 is not very common and we shall not be concerned with it in what follows. The methods required for the analysis of data obtained under Scheme 2 are the same as those corresponding Scheme 1, provided independent priors are chosen for the parameters of all the multinomial distributions. This is a common assumption and thus we will be focusing on Scheme 1 without any real loss of generality.

Let $m = rc$ denote the total number of cells in a two-way, $r \times c$ contingency table. We will sometimes find it convenient to arrange both the cell counts and the cell probabilities in an $m \times 1$ vector.

2 Bayesian Methods for Categorical Data

Thus, we will denote by $\tilde{\pi}_l$ and $\tilde{n}_l$ respectively, the probability and observed count for cell l ($l = 1, \dots, m$) and $\boldsymbol{\pi}$ will denote both $(\tilde{\pi}_1, \dots, \tilde{\pi}_m)^T$ and $(\pi_{11}, \dots, \pi_{rc})^T$, with the entries of the latter arranged in lexicographical order. Similarly, $\mathbf{n}$ will denote both $(\tilde{n}_1, \dots, \tilde{n}_m)^T$ and $(n_{11}, \dots, n_{rc})^T$.

Under multinomial sampling, the vector of counts, $\mathbf{n}$, is regarded as an observation from a $(m-1)$ -dimensional multinomial distribution with index $N = \sum_l \tilde{n}_l$ and unknown parameter vector $\boldsymbol{\pi}$:

$$f(\mathbf{n}|\boldsymbol{\pi}, N) = \frac{N!}{\prod_l \tilde{n}_l!} \prod_l \tilde{\pi}_l^{\tilde{n}_l}, \quad (1)$$

where $\tilde{\pi}_l > 0$ and $\sum_l \tilde{\pi}_l = 1$.

Some History

Early accounts of Bayesian analyses for categorical data include [20], in particular, Section 5.11, and [16], [17], [18], and [27].

Suppose that the cell counts $\mathbf{n}$ have a multinomial distribution with density function (1) and that the prior density of $\boldsymbol{\pi}$ is proportional to $\prod_l \tilde{\pi}_l^{-1}$ over the region $\tilde{\pi}_l > 0$, $\sum_l \tilde{\pi}_l = 1$ (this is a limiting case of the Dirichlet distribution and is meant to describe vague prior information; see section titled ‘Bayesian Inference for Multinomial Data’). Write $\mathbf{y} = (\log \tilde{n}_1, \dots, \log \tilde{n}_m)^T$ and $\tilde{\boldsymbol{\phi}} = (\log \tilde{\pi}_1, \dots, \log \tilde{\pi}_m)^T$, and let $\mathbf{C}$ be a $k \times m$ matrix of rank $k < m$ and rows summing to zero. Then Lindley ([27], Theorem 1) showed that, provided the cell counts are not too small, the posterior distribution of the contrasts $\boldsymbol{\phi} = \mathbf{C}\tilde{\boldsymbol{\phi}}$ is given approximately by

$$\boldsymbol{\phi} \sim MVN(\mathbf{C}\mathbf{y}, \mathbf{C}\mathbf{N}^{-1}\mathbf{C}^T), \quad (2)$$

where $\mathbf{N}$ is a diagonal matrix with entries $(\tilde{n}_1, \dots, \tilde{n}_m)$.

This result provides approximate estimates of the ‘log-odds’ $\tilde{\boldsymbol{\phi}}$, or linear functions thereof, but not of the cell probabilities $\boldsymbol{\pi}$. Nevertheless, when testing common hypothesis in two-way contingency tables (such as independence or homogeneity of populations), Lindley found analogies with the classical **analysis of variance** which greatly simplify the analysis. He proposed a Bayesian ‘significance test’ based on highest posterior density credible intervals (see also [18]). Spiegelhalter and Smith [32] discuss an

alternative testing procedure based on Bayes factors (see also [20]).

For three-way contingency tables, the analogy with the analysis of variance is no longer useful, but the analysis can still be carried out at the cost of additional computations.

Good [19] developed a Bayesian approach to testing independence in multiway contingency tables. Unlike Lindley’s, this approach has the advantage that it does not depend on the availability of large samples and so is applicable even when many expected cell frequencies are small. Moreover, this approach allows one to estimate the cell probabilities (see also [16] and [17]). To test for independence, Good proposed the use of Bayes factors where the priors assumed for the nonnull model are mixtures of symmetric Dirichlet distributions (see also [2]).

Bishop et al. [9] consider pseudo-Bayes estimators arising from the use of a two-stage prior distribution, following a suggestion by Good [18]. Such estimators are essentially empirical Bayes estimators (see also [22]). Bishop et al. also provide various asymptotic results concerning the risk of their estimators.

Leonard [23] uses exchangeable normal priors on the components of a set of multivariate logits. He then derives estimators of the cell frequencies from the resulting posterior distributions. In a subsequent paper [24], he also develops estimators of the cell frequencies from several multinomial distributions via two-stage priors (see also [25] and [26]). Albert and Gupta [6] also consider estimation in contingency tables, but use mixtures of Dirichlet distributions as priors. In [7], they discuss certain tailored priors that allow them to incorporate (i) separate prior knowledge about the marginal probabilities and an interaction parameter in 2×2 tables; and (ii) prior beliefs about the similarity of a set of cell probabilities in $r \times 2$ tables with fixed row totals (see also [4]).

Bayesian Inference for Multinomial Data

Conjugate Analysis

The standard conjugate prior for the multinomial parameter $\boldsymbol{\pi}$ in (1) is the Dirichlet distribution, with density function

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_*)}{\prod_l \Gamma(\alpha_l)} \prod_l \tilde{\pi}_l^{\alpha_l-1}, \quad (3)$$

for $\alpha_l > 0$ and $\alpha_* = \sum_l \alpha_l$, where $\Gamma(\cdot)$ is the **gamma function** (see [1]).

This distribution is characterized by a parameter vector $\alpha = (\alpha_1, \dots, \alpha_m)^T$ such that $E(\pi_l) = \alpha_l/\alpha_*$. The value of α_* is interpreted as a ‘hypothetical prior sample size’, and determines the strength of the information contained in the prior: a small α_* implies vague prior information whereas a large α_* suggests strong prior beliefs about π . Owing to the conjugacy property, the corresponding posterior distribution of π is also Dirichlet with parameter $\alpha_n = (\tilde{n}_1 + \alpha_1, \dots, \tilde{n}_m + \alpha_m)^T$. This distribution contains *all* the available information about the cell probabilities π , conditional on the observed counts $\mathbf{n}$.

In the absence of prior information, we would typically use a rather vague prior. One of the most widely used such priors for the multinomial parameter is precisely the Dirichlet distribution with parameter $\alpha = (1/2, \dots, 1/2)$ (see [20]). In practical terms, however, one could argue that the strength of the prior should be measured in relation to the actual observed sample. Keeping in mind the interpretation of α_* as a ‘prior sample size’, the quantity $I = \alpha_*/(N + \alpha_*)$ can be regarded as the proportion of the total information that is contributed by the prior. Thus, a value of α_* yielding $I = 0.01$ would produce a fairly vague prior contributing about 1% of the total information, whereas $I \approx 1$ would imply that the data are completely dominated by the prior. Since $E(\pi_l) = \alpha_l/\alpha_*$, the individual values of the α_i should be chosen according to the prior beliefs concerning $E(\tilde{\pi}_l)$. These may be based on substantive knowledge about the population probabilities or on data from previous studies. In the case of vague prior information ($I \leq 0.05$, say), $\alpha_l = \alpha_*/m$ for all $l = 1, \dots, m$ is a sensible default choice. When $\alpha_* = 1$, this corresponds to the prior proposed by Perks [31] and can be interpreted as a single ‘prior observation’ divided evenly among all the cells in the table.

When analyzing contingency tables, we often wish to provide a table of expected cell probabilities or frequencies that can be used for other purposes such as computing standardized rates. The raw observed counts are usually not satisfactory for this purpose, for example, when the table has many cells and/or when few observations are available. In such cases, Bayesian estimators based on posterior expectations

$$E(\tilde{\pi}_l|\mathbf{n}) = \frac{\tilde{n}_l + \alpha_l}{N + \alpha_*}, \quad (4)$$

are often used as smoothed expected cell probabilities (see also [9], [13], [6] and [23]).

Testing for Independence

When testing hypothesis concerning the cells probabilities or frequencies in a contingency table, the null hypothesis imposes constraints on the space of possible values of π . In other words, under the null hypothesis, the cell probabilities are given by $\tilde{\pi}_l^0 = h_l(\pi)$ for some functions $h_l(\cdot)$, $l = 1, \dots, m$. As a simple example, consider a $r \times c$ contingency table and a null model which states that the two variables are independent. In this case, it is convenient to use the double-index notation to refer to the individual cell probabilities or counts. Then

$$\pi_{ij}^0 = h_{ij}(\pi) \equiv \pi_{i+} \pi_{+j}, \quad (5)$$

where $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ ($i = 1, \dots, r$; $j = 1, \dots, c$).

Since we have the posterior distribution of π , we can, in principle, calculate the posterior probability of any event involving the cell probabilities π . In particular, the posterior distribution of π induces a posterior distribution on the vector $\pi^0 = (\pi_{11}^0, \dots, \pi_{rc}^0)^T$ of cell probabilities constrained by the null hypothesis.

The null model of independence can be tested on the basis of the posterior distribution of

$$\delta = \delta(\pi) \equiv \sum_l \log \left(\frac{\tilde{\pi}_l}{\tilde{\pi}_l^0} \right) \log(\tilde{\pi}_l). \quad (6)$$

This quantity can be regarded as a Bayesian version of the deviance. It is always nonnegative and is zero if and only if the null model and the saturated model are the same, i.e., if and only if $\pi_l^0 = \pi_l$ for all l .

The marginal posterior distribution of δ is not available in closed form, but it can easily be obtained from that of π using Monte Carlo techniques. In this case, we can generate a sample $\{\pi^{(1)}, \dots, \pi^{(M)}\}$ of size M from the posterior (Dirichlet) distribution of π . Next, we compute $\delta^{(k)} = \delta(\pi^{(k)})$ for each $k = 1, \dots, M$. The resulting values $\{\delta^{(1)}, \dots, \delta^{(M)}\}$ then constitute a sample from the marginal posterior distribution of δ . The accuracy of the Monte Carlo approximation increases with the value of M .

4 Bayesian Methods for Categorical Data

Posterior distributions of δ concentrated around zero support the null model, whereas posterior distributions located away from zero lead to rejection of the null model. Following a suggestion by Lindley [27], we can test the null hypothesis of independence by means of a Bayesian ‘significance test’: reject the null hypothesis if the 95% (say) highest posterior density credible interval for δ does not contain the value zero.

Numerical example. The $3 \times 2 \times 4$ cross-classification in Table 1 shows data previously analyzed in [21]. The data concern a small study of alcohol intake, hypertension, and obesity.

A sample of size $M = 10\,000$ was simulated from the posterior distribution of δ for the null model of

Table 1 Alcohol, hypertension, and obesity data

Obesity	High BP	Alcohol intake (drinks/day)			
		0	1–2	3–5	6+
Low	Yes	5	9	8	10
	No	40	36	33	24
Average	Yes	6	9	11	14
	No	33	23	35	30
High	Yes	9	12	19	19
	No	24	25	28	29

independence of the three variables. Figure 1 (upper-left panel) shows the corresponding histogram. In this case, the posterior distribution of δ is located away

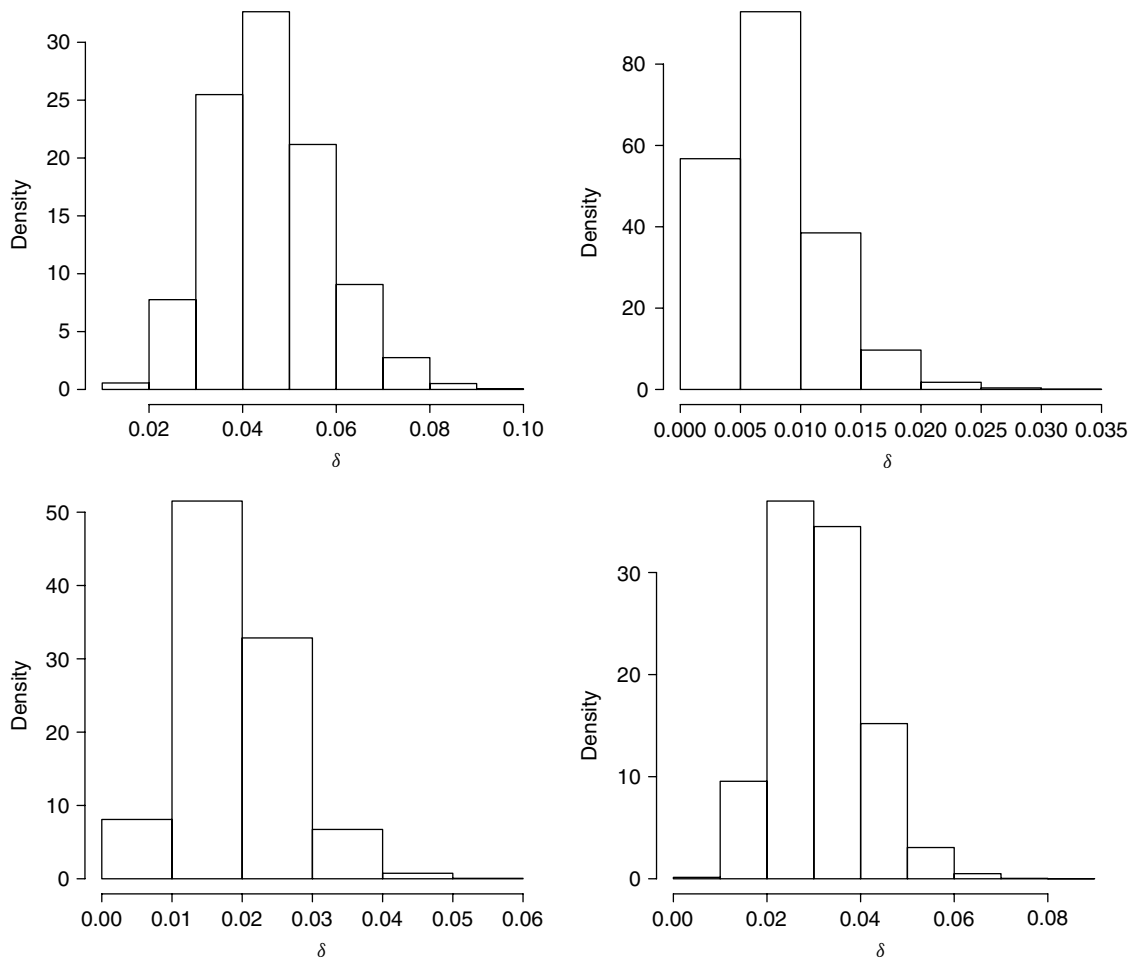


Figure 1 Posterior distribution of δ for various models

from zero, indicating that the model of independence should be rejected.

A similar procedure can be used to perform other analyses of contingency tables, such as tests of homogeneity of populations or even some interaction tests based on simple log-linear model (see section titled ‘Log-linear and Generalized Linear Models’). All we require is (a) the ability to generate large samples from Dirichlet distributions; and (b) standard software capable of producing the estimated expected cell probabilities or cell frequencies under the relevant null model, thus implicitly providing the corresponding functions $h_l(\cdot)$ discussed above.

Other Priors

The Dirichlet distribution is suitable for inputting prior information about the cell probabilities, but it does not allow sufficient structure to be imposed on such probabilities. Alternative classes of prior distribution were mentioned in the section titled ‘Some History’, and, in the next section, we describe yet another alternative which is particularly suitable for log-linear models.

It is often convenient to model multinomial data as observations of independent Poisson variables. This approach leads to valid Bayesian inferences provided that the prior for the Poisson means factors in a particular way (see [27]). This result can be generalized to product-multinomial settings.

Log-linear and Generalized Linear Models

Log-linear models provide a general and useful framework for analyzing multidimensional contingency tables. Consider a two-way contingency table with r rows and c columns. A log-linear model for the cell probabilities has the form

$$\log \pi_{ij} = u_0 + u_{1(i)} + u_{2(j)} + u_{12(ij)},$$

$$i = 1, \dots, r; \quad j = 1, \dots, c; \quad (7)$$

where u_0 is the overall mean, $u_{1(i)}$ and $u_{2(j)}$ represent the main effects of variables 1 and 2 respectively, and $u_{12(ij)}$ represents the interaction between variables 1 and 2 (see, for example, [9]). The number of independent parameters must be equal to the total number of elementary cells in the table, so it is necessary to

impose constraints to reduce the number of independent parameters represented by each u -term. Usually, such constraints take the form $\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0$.

If we consider instead a table of expected counts $\{\mu_{ij}\}$ that sum to the grand total $N = \sum_i \sum_j n_{ij}$, then we have $\mu_{ij} = N\pi_{ij}$ and hence,

$$\log \mu_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)},$$

$$i = 1, \dots, r; \quad j = 1, \dots, c; \quad (8)$$

where $u = u_0 + \log N$.

As mentioned in the section titled ‘Testing for Independence’, simple models of this type can also be analyzed using the procedure outlined there.

Numerical example (ctd). For the data of Table 1, we now consider the following model with no second-order interaction:

$$\log \mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}, \quad (9)$$

where i denotes obesity level, j denotes blood pressure level, and k denotes alcohol intake level.

We simulated a sample of size $M = 10\,000$ from the posterior distribution of δ with the aid of the function `loglin` of the R language and environment for statistical computing (<http://www.R-project.org>). Figure 1 (upper-right panel), shows the corresponding histogram. In this case, the posterior distribution of δ is concentrated around zero, indicating that the model provides a good fit. However, it is possible that a more parsimonious model also fits the data.

Consider, for example, the models

$$\log \mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}, \quad (10)$$

and

$$\log \mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}. \quad (11)$$

The posterior distribution of δ for each of these models is shown in Figure 1 (lower-left and lower-right panels, respectively). For model (10), the 95% highest posterior density credible interval for δ contains the value zero, whereas, for model (11), this is not

the case. Thus, we reject model (11) and retain model (10), which suggests that alcohol intake and obesity are independently associated with hypertension.

The saturated log-linear model allows $\boldsymbol{\varphi} = (\log \mu_{11}, \dots, \log \mu_{rc})^T$ to take any value on R^{rc} . A nonsaturated model constrains $\boldsymbol{\varphi}$ to lie in some vector subspace of R^{rc} , in which case, we can write

$$\boldsymbol{\varphi} = \mathbf{X}\mathbf{u}, \quad (12)$$

where $\mathbf{X}$ is a ‘design’ matrix with columns containing the values of explanatory variables or the values of dummy variables for main effects and interaction terms, and $\mathbf{u}$ is the corresponding vector of unknown regression coefficients or effects.

Knuiman and Speed [21] discuss a general procedure to incorporate prior information directly into the analysis of log-linear models. In order to incorporate constraints on main effects and interaction parameters, they use a structured multivariate normal prior for all parameters taken collectively, rather than specify univariate normal priors for individual parameters, as in [23] and [22]. A useful feature of this general prior is that it allows separate specification of prior information for different interaction terms. They go on to propose an approximate Bayesian analysis, where the mode and curvature of the posterior density at the mode are used as summary statistics.

Dellaportas and Smith [11] show how a specific **Markov chain Monte Carlo** algorithm – known as the Gibbs sampler – may be implemented to produce exact, fully Bayesian analyses for a large class of generalized linear models, of which the log-linear model with a multivariate normal prior is a special case.

Dellaportas and Forster [10] use reversible jump Markov chain Monte Carlo methods to develop strategies for calculating posterior probabilities of hierarchical log-linear models for high-dimensional contingency tables. The best models are those with highest posterior probability. This approach to model selection is closely related to the use of Bayes factors, but it also takes into account the prior probabilities of all of the models under consideration (see also [3]).

Specialized Models

In this section, we present a selective review of some specialized problems for which modern Bayesian techniques are particularly well suited.

Missing Data: Nonresponse

Park and Brown [28] and Forster and Smith [15] develop Bayesian approaches to modeling nonresponse in categorical data problems. Specifically, the framework they consider concerns contingency tables containing both completely and partially cross-classified data, where one of the variables (Y , say) is a response variable subject to nonignorable nonresponse and the other variables (here collectively denoted by X) are regarded as covariates and are always observed. They then introduce an indicator variable R to represent a dichotomous response mechanism ($R = 1$ and $R = 0$ indicating response and nonresponse respectively). A nonresponse model is defined as a log-linear model for the full array of Y , X , and R . A nonignorable nonresponse model is one that contains a YR interaction term.

Park and Brown [28] show that a small shift of the nonrespondents can result in large changes in the maximum likelihood estimates of the expected cell frequencies. Maximum likelihood estimation is problematic here because boundary solutions can occur, in which case the estimates of the model parameters cannot be uniquely determined. Park and Brown [28] propose a Bayesian method that uses data-dependent priors to provide some information about the extent of nonignorability. The net effect of such priors is the introduction of smoothing constants, which avoid boundary solutions.

Nonidentifiability

Censoring. Standard models for censored categorical data (see **Censored Observations**) are usually nonidentifiable. In order to overcome this problem, the censoring mechanism is typically assumed to be ignorable (noninformative) in that the unknown parameter of the distribution describing the censoring mechanism is unrelated to the parameter of interest (see [12] and the references therein). Paulino and Pereira [29] discuss Bayesian conjugate methods for categorical data under general, informative censoring. In particular, they are concerned with Bayesian estimation of the cell frequencies through posterior expectations. Walker [33] considers maximum *a posteriori* estimates, obtained via an EM algorithm, for a more general class of priors.

Misclassification. Paulino et al. [30] present a fully Bayesian analysis of binomial regression data with a

possibly misclassified response. Their approach can be extended to multinomial settings. They use an informative misclassification model whose parameters turn out to be nonidentifiable. As in the case of censoring, from a Bayesian point of view this is not a serious problem since a suitable proper prior will typically make the parameters identifiable. However, care must be taken since posterior inferences on non-identifiable parameters may be strongly influenced by the prior even for large sample sizes.

Latent Class Analysis. A **latent class model** usually involves a set of observed variables called *manifest variables* and a set of unobservable or unobserved variables called **latent variables**. The most commonly used models of this type are the latent conditionally independence models, which state that all the manifest variables are conditionally independent given the latent variables.

Latent class analysis in two-way contingency tables usually suffers from unidentifiability problems. These can be overcome by using Bayesian techniques in which prior distributions are assumed on the latent parameters.

Evans et al. [14] discuss an adaptive importance sampling approach to the computation of posterior expectations, which are then used as point estimates of the model parameters.

Ordered Categories

Albert and Chib [5] develop exact Bayesian methods for modeling categorical response data using the idea of data augmentation combined with Markov chain Monte Carlo techniques. For example, the probit regression model (see **Probits**) for binary outcomes is assumed to have an underlying normal regression structure on latent continuous data. They generalize this idea to multinomial response models, including the case where the multinomial categories are ordered. In this latter case, the models link the cumulative response probabilities with the linear regression structure.

This approach has a number of advantages, especially in the multinomial setup, where it can be difficult to evaluate the likelihood function. For small samples, this Bayesian approach will usually perform better than traditional maximum likelihood methods, which rely on asymptotic results. Moreover, one can

elaborate the probit model by using suitable mixtures of normal distributions to model the latent data.

References

- [1] Abramowitz, M. & Stegun, I.A. (1965). *Handbook of Mathematical Functions*, Dover Publications, New York.
- [2] Albert, J.H. (1990). A Bayesian test for a two-way contingency table using independence priors, *Canadian Journal of Statistics* **18**, 347–363.
- [3] Albert, J.H. (1996). Bayesian selection of log-linear models, *Canadian Journal of Statistics* **24**, 327–347.
- [4] Albert, J.H. (1997). Bayesian testing and estimation of association in a two-way contingency table, *Journal of the American Statistical Association* **92**, 685–693.
- [5] Albert, J.H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**, 669–679.
- [6] Albert, J.H. & Gupta, A.K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables, *The Annals of Statistics* **10**, 1261–1268.
- [7] Albert, J.H. & Gupta, A.K. (1983). Estimation in contingency tables using prior information, *Journal of the Royal Statistical Society B* **45**, 60–69.
- [8] Altham, P.M.E. (1969). Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significant test, *Journal of the Royal Statistical Society B* **31**, 261–269.
- [9] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.
- [10] Dellaportas, P. & Forster, J.J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, *Biometrika* **86**, 615–633.
- [11] Dellaportas, P. & Smith, A.F.M. (1993). Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling, *Applied Statistics* **42**, 443–459.
- [12] Dickey, J.M., Jiang, J.-M. & Kadane, J.B. (1987). Bayesian methods for censored categorical data, *Journal of the American Statistical Association* **82**, 773–781.
- [13] Epstein, L.D. & Fienberg, S.E. (1992). Bayesian estimation in multidimensional contingency tables, in *Bayesian Analysis in Statistics and Economics*, P.K. Goel & N.S. Iyengar, eds, Springer-Verlag, New York, pp. 37–47.
- [14] Evans, M.J., Gilula, Z. & Guttman, I. (1989). Latent class analysis of two-way contingency tables by Bayesian methods, *Biometrika* **76**, 557–563.
- [15] Forster, J.J. & Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response, *Journal of the Royal Statistical Society B* **60**, 57–70.
- [16] Good, I.J. (1956). On the estimation of small frequencies in contingency tables, *Journal of the Royal Statistical Society B* **18**, 113–124.
- [17] Good, I.J. (1965). *The Estimation of Probabilities*, Research Monograph No. 30, MIT Press, Cambridge.

- [18] Good, I.J. (1967). A Bayesian significant test for multinomial distributions (with discussion), *Journal of the Royal Statistical Society B* **29**, 399–431.
- [19] Good, I.J. (1976). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, *The Annals of Statistics* **4**, 1159–1189.
- [20] Jeffreys, H. (1961). *Theory of Probability*, 3rd Edition, Clarendon Press, Oxford.
- [21] Knuiman, M.W. & Speed, T.P. (1988). Incorporating prior information into the analysis of contingency tables, *Biometrics* **44**, 1061–1071.
- [22] Laird, N.M. (1978). Empirical Bayes for two-way contingency tables, *Biometrika* **65**, 581–590.
- [23] Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables, *Journal of the Royal Statistical Society B* **37**, 23–37.
- [24] Leonard, T. (1977). Bayesian simultaneous estimation for several multinomial distributions, *Communications in Statistics – Theory and Methods* **6**, 619–630.
- [25] Leonard, T. (1993). The Bayesian analysis of categorical data – a selective review, in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, P.R. Freeman & A.F.M. Smith, eds, Wiley, New York, pp. 283–310.
- [26] Leonard, T. (2000). *A Course in Categorical Data Analysis*, Chapman & Hall, London.
- [27] Lindley, D.V. (1964). The Bayesian analysis of contingency tables, *The Annals of Mathematical Statistics* **35**, 1622–1643.
- [28] Park, T. & Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse, *Journal of the American Statistical Association* **89**, 44–52.
- [29] Paulino, C.D. & Pereira, C.A. (1995). Bayesian methods for categorical data under informative general censoring, *Biometrika* **82**, 439–446.
- [30] Paulino, C.D., Soares, P. & Neuhaus, J. (2003). Binomial regression with misclassification, *Biometrics* **59**, 670–675.
- [31] Perks, F.J.A. (1947). Some observations on inverse probability including a new indifference rule (with discussion), *Journal of the Institute of Actuaries* **73**, 285–334.
- [32] Spiegelhalter, D.J. & Smith, A.F.M. (1982). Bayes factors for linear and log-linear models with vague prior information, *Journal of the Royal Statistical Society B* **44**, 377–387.
- [33] Walker, S. (1996). A Bayesian maximum a posteriori algorithm for categorical data under informative general censoring, *The Statistician* **45**, 293–298.

EDUARDO GUTIÉRREZ-PEÑA

Bayesian Statistics

LAWRENCE D. PHILLIPS

Volume 1, pp. 146–150

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bayesian Statistics

Bruno de Finetti, one of the founders of Bayesian statistics, wrote in his 1974 *Theory of Probability* [3], ‘PROBABILITIES DO NOT EXIST’. He meant that probabilities do not exist as properties of things; instead, they represent an individual’s degree of belief about some unknown event or uncertain quantity. These personal probabilities must, however, obey the laws of probability, such as the multiplication and addition laws that describe how probabilities assigned to individual events combine to describe the uncertainty associated with compound events (*see Probability: Foundations of*). One consequence of those laws is Bayes’s theorem (*see Bayesian Belief Networks*), which shows how probabilities are revised in the light of new information. It is this crucial theorem that brings individual expressions of uncertainty into contact with real-world data, with the result that with sufficient information, two people holding initially very different views will find their final probabilities in near agreement. For a Bayesian, that agreement defines scientific truth.

The idea is to capture one’s uncertainty about an event or uncertain quantity in the form of ‘prior’ probabilities, then gather data, observe the results and summarize them as a special probability known as a *likelihood*, then apply Bayes’s theorem by multiplying the prior by the likelihood, giving a ‘posterior’ probability (the mathematical details are given in the entry on Bayesian belief networks). Put simply,

$$\begin{aligned} &\text{posterior probability} = \text{prior probability} \\ &\quad \times \text{likelihood.} \end{aligned}$$

An example illustrates the approach. Imagine that you wish to know the proportion, π , of people in a defined population who share some characteristic of interest, such as their ability to smell freesias, the South African flower that your spouse, say, experiences as very fragrant, but for you smells faintly or not at all. On the basis of that limited data from two people, you know that π cannot be either zero or one, but it could be any value in between. One way to represent your uncertainty at this point is as the gentle distribution in Figure 1; no probability at $\pi = 0$ or 1, with probabilities increasing away from those values, peaking at 0.5. Now, you ask

many people to smell a bunch of freesias, quitting when you are tired of asking, and find that you have asked 60 people, of whom 45 reported the freesias as very fragrant. Applying Bayes’s theorem results in the peaked distribution of Figure 1. Now, most of your opinion about the proportion of people who can smell freesias falls between about 0.6 and 0.9; prior uncertainty has been transformed by the data into narrower posterior uncertainty. If more people were to be sampled, the current posterior distribution would become the prior, and after the new data were obtained, the new posterior would be even narrower, though it might shift either left or right.

The beginnings of Bayesian statistics date back to 1763 with the publication, posthumously, of a paper by the **Reverend Thomas Bayes** [1], an English nonconformist who recognized the implications of the laws of probability, though he did not explicitly show the theorem that now bears his name. The modern beginnings can be traced to foundations laid by Ramsey [16], de Finetti [2], Good [5], and others, with the first practical procedures developed by Jeffreys [7, 8]. Although **Savage** [17] failed in his attempt to provide a Bayesian justification for classical statistics, he recognized in the second edition that the two approaches are not reconcilable. His book provided a complete axiomatic treatment of both personal probability and utility to encompass decision making, thus extending foundations laid by von Neumann and Morgenstern [20]. These two books stimulated the publication of Schlaifer’s textbooks [18, 19], Raiffa and Schlaifer’s development of Bayesian decision theory [15] and the two-volume textbook by Lindley [11], who was a student of Jeffreys. Phillips provided the first textbook for social scientists [13], with exercises drawn from research reports in psychology, economics, and sociology. Today, many textbooks provide excellent treatments of the topic, with Winkler’s 1972 textbook a particularly good example, now reissued in a second edition [21].

It is worth distinguishing the early reception of Bayesian ideas by the scientific and business communities. Critics in the sciences argued that it was not appropriate to introduce subjectivism into the objective pursuit of truth, and, anyway, prior opinion was too difficult to assess. The business community, on the other hand, welcomed the ability to combine ‘collateral’ information in the prior along with hard data in the likelihood, thereby allowing both experience and data to inform decisions in a

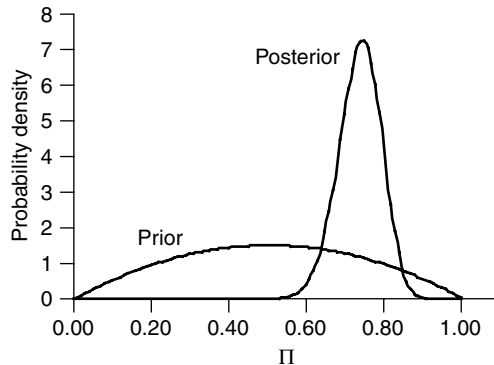


Figure 1 Prior and posterior distributions for the freesias example

formal analysis. In some practical applications, prior probabilities are largely based on hard data, while likelihoods are judged by specialists and experts in the topic at hand. Whatever the source of priors and likelihoods, methods for assessing them are now well developed and routinely applied in many fields [12].

As for the scientists' critique, the thorough examination of the foundations of all statistical inference approaches given by the philosophers Howson and Urbach [6] shows that subjective judgments attend all approaches, Bayesian and classical. For example, the choice of a significance level, the **power** of a test, and Type 1 and Type 2 errors are all judgments in classical methods, though it often appears not to be the case when, for example, social science journals require 0.05 or 0.01 levels of significance for results to be published, thereby relieving the scientist of having to make the judgment.

Bayesian statistics was first introduced to psychologists by Edwards, Lindman, and Savage [4] in their landmark paper that set out two important principles: stable estimation and the likelihood principle. Stable estimation is particularly important for scientific research, for it enables certain properties of prior opinion to justify use of a 'noninformative' prior, that is, a prior that has little control over the posterior distribution, such as a uniform prior. In the freesias example, a uniform prior would be a horizontal line intersecting the y-axis at 1.0. If that were the prior, and the data showed that one person can smell the freesias and the other cannot, then the posterior would be the gentle curve shown as the prior in Figure 1. Stable estimation states that the actual

prior can be ignored if one's prior in the vicinity of the data changes only gently and is not substantially higher elsewhere. That is the case with the freesias example. The proportion of people in the sample who can smell freesias is $45/60 = 0.75$. In the vicinity of 0.75, the actual prior does not change very much (it is mostly about 1.1), and the prior does not show a very much larger amount elsewhere, such as a substantial amount on $\pi = 1.0$, which might be thought appropriate by a flower seller whose clients frequently comment on the strong, sweet smell. Thus, to quote Edwards, Lindman, and Savage, '... far from ignoring prior opinion, stable estimation exploits certain well-defined features of prior opinion and is acceptable only insofar as those features are really present'. For much scientific work, stable estimation justifies use of a uniform prior.

The likelihood principle states that all the information relevant to a statistical inference is contained in the likelihood. For the freesias example, the relevant data are only the number of people who can smell the freesias and the number who cannot. The order in which those data were obtained is not relevant, nor is the rule employed to determine when to stop collecting data. In this case, it was when you became tired of collecting data, a stopping rule that would confound the classical statistician whose significance test requires knowing whether you decided to stop at 60 people, or when 45 'smellers' or 15 'nonsmellers' were obtained.

In most textbooks, much is made of the theory of conjugate distributions, for that greatly simplifies calculations. Again, for the freesias example, the sampling process is judged to be Bernoulli (*see Catalogue of Probability Density Functions*): each 'smeller' is a success, each 'nonsmeller' a failure; the data then consist of s successes and f failures. By the theory of conjugate distributions, if the prior is in the two-parameter Beta family, then with a Bernoulli process generating the data, the posterior is also in the Beta family (*see Catalogue of Probability Density Functions*). The parameters of the posterior Beta are simply the parameters of the prior plus s and f , respectively. For the above example, the prior parameters are 2 and 2, the data are $s = 45$ and $f = 15$, so the posterior parameters are 47 and 17. The entire distribution can be constructed knowing only those two parameters. While Bayesian methods are often more computationally difficult than classical tests, this is no longer a problem with the ready

availability of computers, simulation software, and Bayesian statistical programs (see **Markov Chain Monte Carlo and Bayesian Statistics**).

So how does Bayesian inference compare to classical methods (see **Classical Statistical Inference: Practice versus Presentation**)? The most obvious difference is in the definition of probability. While both approaches agree about the laws of probability, classical methods assume a relative frequency interpretation of probability (see **Probability: An Introduction**). As a consequence, posterior probability distributions play no part in classical methods. The true proportion of people who can smell freesias is a particular, albeit unknown, value, X . There can be no probability about it; either it is X or it is not. Instead, sampling distributions are constructed: if the freesias experiment were repeated over and over, each with, say 60 different people, then the proportion of smellers would vary somewhat, and it is this hypothetical distribution of results that informs the inferences made in classical methods. Sampling distributions enable the construction of confidence intervals, which express the probability that the interval covers the true value of π . The Bayesian also calculates an interval, but as it is based on the posterior distribution, it is called a *credible interval*, and it gives the probability that π lies within the interval. For the freesias example, there is a 99% chance that X lies between 0.59 and 0.86. The **confidence interval** is a probability statement about the interval, while the credible interval is a statement about the uncertain quantity, π , a subtle distinction that often leads the unwary to interpret confidence intervals as if they were credible intervals.

As social scientists know, there are two stages of inference in any empirical investigation, statistical inference concerning the relationship between the data and the statistical hypotheses, and scientific inference, which takes the inference a step beyond the statistical hypotheses to draw conclusions about the scientific hypothesis. A significance level interpreted as if it were a Bayesian inference usually makes little difference to the scientific inferences, which is possibly one reason why social scientists have been slow to take up Bayesian methods.

Hypothesis testing throws up another difference between the approaches. Many significant results in the social science literature establish that a result is not just a chance finding, that the difference on some measure between a treatment and a control group is

‘real’. The Bayesian approach finds the posterior distribution of the difference between the measures, and determines the probability of a positive difference, which is the area of the posterior density function to the right of zero. That probability turns out to be similar to the classical one-tailed significance level, provided that the Bayesian’s prior is noninformative. The Bayesian would report the probability that the difference is positive; if it is greater than 0.95, that would correspond to significance of $p < 0.05$. But the significance level should be interpreted as meaning that there is less than a 5% chance that this result or one more extreme would be obtained if the null hypothesis of no difference were true. Therefore, since this probability is so small, the null hypothesis can be rejected. The Bayesian, on the other hand, asserts that there is better than a 95% chance, based only on the data actually observed, that there is a real difference between treatment and control groups. Thus, the significance level is a probability statement about data, while the Bayesian posterior probability is about the uncertain quantity of interest.

For the freesias example, 60% of the probability density function lies to the left of $\pi = 0.75$, so there is a 60% chance that the proportion π is equal to or less than 0.75. If a single estimate about π were required, the mean of the posterior distribution, $\pi = 0.73$, would be an appropriate figure, slightly different from the sample mean of 0.75 because of the additional prior information.

In comparing Bayesian and classical methods, Pitz [14] showed graphically, cases in which data that led to a classical rejection of a null hypothesis actually provided evidence in favor of the null hypothesis in a Bayesian analysis of the same data, examples of Lindley’s paradox [10]. Lindley proved that as the sample size increases, it is always possible to obtain a significant rejection of a point hypothesis whether it is true or false. This applies for any significance level at all, but only for classical two-tailed tests, which have no interpretation in Bayesian theory.

From the perspective of making decisions, significance levels play no part, which leaves the step between classical statistical inference and decision making bridgeable only by the exercise of unaided judgment (see entries on utility and on strategies of decision making). On the other hand,

Bayesian posterior probabilities or predictive probabilities about uncertain quantities or events are easily accommodated in decision trees, making possible a direct link between inference and decision. While this link may be of no interest to the academic researcher, it is vital in many business applications and for regulatory authorities, where important decisions based on fallible data are made. Indeed, the design of experiments can be very different, as Kadane [9] has demonstrated for the design of clinical trials in pharmaceutical research. This usefulness of Bayesian methods has led to their increasing acceptance over the past few decades, and the early controversies have now largely disappeared.

References

- [1] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society* **53**, 370–418, Reprinted in Barnard, 1958.
- [2] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales De L'Institut Henri Poincaré* **7**, Translated by H.E. Kyburg, (Foresight: Its logical laws, its subjective sources) in Kyburg and Smokler, 1964.
- [3] de Finetti, B. (1974). *Theory of Probability*, Vol. 1, Translated by A. Machi & A. Smith, Wiley, London.
- [4] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research, *Psychological Review* **70**(3), 193–242.
- [5] Good, I.J. (1950). *Probability and the Weighing of Evidence*, Griffin Publishing, London.
- [6] Howson, C. & Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*, 2nd Edition, Open Court, Chicago.
- [7] Jeffreys, H. (1931). *Scientific Inference*, 3rd Edition, 1957, Cambridge University Press, Cambridge.
- [8] Jeffreys, H. (1939). *Theory of Probability*, 3rd Edition, 1961, Oxford, Clarendon.
- [9] Kadane, J., ed. (1996). *Bayesian Methods and Ethics in a Clinical Trial Design*, John Wiley & Sons, New York.
- [10] Lindley, D. (1957). A statistical paradox, *Biometrika* **44**, 187–192.
- [11] Lindley, D. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Vols. 1, 2, Cambridge University Press, Cambridge.
- [12] Morgan, M.G. & Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge.
- [13] Phillips, L.D. (1973). *Bayesian Statistics for Social Scientists*, Thomas Nelson, London; Thomas Crowell, New York, 1974.
- [14] Pitz, G.F. (1978). Hypothesis testing and the comparison of imprecise hypotheses, *Psychological Bulletin* **85**(4), 794–809.
- [15] Raiffa, H. & Schlaifer, R. (1961). *Applied Statistical Decision Theory*, Harvard University Press, Cambridge.
- [16] Ramsey, F.P. (1926). Truth and probability, in R.B. Braithwaite, ed., *The Foundations of Mathematics and Other Logical Essays*, Keegan Paul, (1931), London, pp. 158–198.
- [17] Savage, L.J. (1954). *The Foundations of Statistics*, 2nd Edition, 1972, Dover Publications, Wiley, New York.
- [18] Schlaifer, R. (1959). *Probability and Statistics for Business Decisions*, McGraw-Hill, New York.
- [19] Schlaifer, R. (1961). *Introduction to Statistics for Business Decisions*, McGraw-Hill, New York.
- [20] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.
- [21] Winkler, R.L. (2003). *An Introduction to Bayesian Inference and Decision*, 2nd Edition, Probabilistic Publishing, Gainesville.

LAWRENCE D. PHILLIPS

Bernoulli Family

STEPHEN SENN

Volume 1, pp. 150–153

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bernoulli Family

In his entertaining but irreverent book on the history of mathematics, Eric Temple Bell took the case of the Bernoullis as being clear proof of the heritability of mathematical genius [1]. This remarkable Swiss dynasty produced at least ten mathematicians of note. Because the names, James, John, Nicholas, and Daniel, appear in different generations and also, depending on the account, in Latin, German, French or English forms, identification of individuals is perilous. In this brief account, the English forms of the personal names will be used together with Boyer's [2] system of numbering. Nicholas (1623–1708), a drug merchant but not a mathematician, is the common ancestor and given no numeral but the qualification 'Senior'. Thereafter, numerals I, II, and III are used as the names reappear in successive generations (see Figure 1). One oddity of this system must be noted and that is that the first Nicholas of mathematical note is Nicholas II, his father Nicholas I and grandfather, Nicholas Senior being the only nonmathematicians necessary to include in order to create a connected tree.

Although many of the Bernoullis did work in mathematics of indirect interest to statisticians, three of them, James I, Nicholas II, and Daniel I, did work of direct importance in probability or statistics and they are covered here.

James I

Born: January 6, 1654, in Basle, Switzerland.

Died: August 16, 1705, in Basle, Switzerland.

James (I) Bernoulli was born in Basle in 1654 and studied at the University, graduating in theology in 1676 [9]. In that year, he left to work as a tutor in Geneva and France, returning to Basle in 1681 [6]. From 1677, he started keeping a scientific diary, *Meditationes*, which traces his interests in mathematics. His first publication in mathematics, however, on the subject of the comet of 1680, predicting its return in 1719, was not until in 1681. In 1687, James became professor of mathematics in Basle. With the exception, perhaps, of his younger brother John I (1667–1748), he became the most important early

developer of the infinitesimal calculus in the form proposed by Leibniz, a fact that Leibniz himself recognized in 1694. James's first publication on the subject of probability dates from 1685, but his fame in this field rests primarily on his posthumous work *Ars Conjectandi* (1713), the first 200 pages of which have been described by the distinguished historian of statistics, Anders Hald [6], as a, 'pedagogical masterpiece with a clear formulation of theorems both in abstract form and by means of numerical examples' (p. 225). In the last 30 pages of the book, however, Bernoulli progresses beyond the then current treatment of probability in terms of symmetric chances *a priori* to develop both subjective interpretations of probability and a famous limit theorem in terms of relative frequencies, the first such to be proved in probability theory. In modern terms, we refer to this as the (weak) **law of large numbers** [11].

Nicholas II

Born: October 10, 1687, in Basle, Switzerland.

Died: November 29, 1759, in Basle, Switzerland.

James's nephew, Nicholas (II) Bernoulli is noteworthy as the editor of his uncle's posthumous masterpiece. He also did important work in probability himself, however, although until recently his role in the development of the subject was underestimated [3] and our modern awareness of his importance is largely due to Hald's careful and thorough analysis of his work [6]. Nicholas was born in 1687 in Basle, the year of the publication of Newton's *Philosophiae naturalis principia mathematica* [3]. In 1704, he obtained a master's degree in mathematics and in 1709, aged only 21, a doctorate in jurisprudence for a thesis entitled, *On the Use of the Art of Conjecturing in Law*. This is clearly considerably influenced by his uncle's work and contains many skillful applications of probability to a wide range of insurance and inheritance problems in which the law could be involved. For example, he has a chapter entitled, 'On an absent person presumed dead'. Nicholas was professor at Padua from 1716 to 1719, after which he returned to Basle to hold first a chair in logic (from 1719 according to Hald, but from 1722 according to Csörgő) and then of Roman and canon law. In 1709, he visited Paris and from 1712 to 1713 undertook a grand tour of France, England,

2 Bernoulli Family

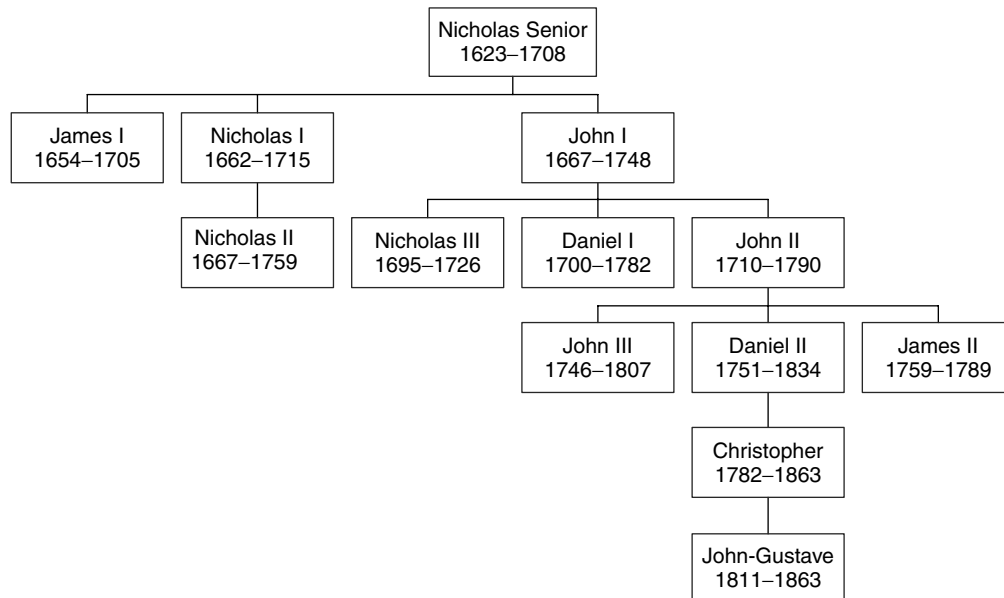


Figure 1 A family tree of the mathematical Bernoullis

and the Netherlands, returning via France. These visits enabled him to establish excellent contacts with many of the leading mathematicians of the day and are the origin of his important correspondence with Montmort through which, together with DeMoivre, he contributed to the solution of a problem posed by William Waldegrave. This involves a circular tournament of n players P_1 to P_n of equal skill. P_1 plays P_2 and the winner plays P_3 , the winner playing P_4 and so on. The game stops once a player has beaten every other player in a row. If necessary, P_1 reenters the game once P_n has played and so on. Montmort and DeMoivre had solutions for $n = 3$ and $n = 4$ but Nicholas was able to provide the general solution.

Nicholas also worked on John Arbuthnot's famous significance test. Arbuthnot had data on christenings by sex in London from 1629 to 1710. Male christenings exceeded female ones in every one of the 82 years, and he used this fact to calculate the probability of this occurring by chance as $(1/2)^{82}$. This is equivalent to, but must not necessarily be interpreted as, a one-sided P value. He then argued that this probability was so small that it could not be interpreted as a chance occurrence and, since it was desirable for the regulation of human affairs that there should be an excess of males at birth, was evidence

of divine providence. The official publication date for Arbuthnot's paper was 1710 and Nicholas Bernoulli discussed it with fellows of the Royal Society during his stay in London in 1712. In a letter to Burnet and 'sGravesande, he uses an improved form of his Uncle James's approximation to the tail area of a binomial distribution (see **Catalogue of Probability Density Functions**) to show that Arbuthnot's data are unsurprising if the probability of a male birth is taken to be $18/35$.

Daniel I

Born: February 8, 1700, in Groningen, The Netherlands.

Died: March 17, 1782, in Basle, Switzerland.

Daniel Bernoulli was in his day, one of the most famous scientists in Europe. His early career in mathematics was characterized by bitter disputes with his father John I, also a brilliant mathematician. Daniel was born in Groningen in 1700, but the family soon returned to Basle. In the same way that John's father, Nicholas Senior, had tried to dissuade his son from studying mathematics, John in turn tried to push Daniel into business [1]. However, when only ten years old, Daniel started to receive lessons in

mathematics from his older brother, Nicholas III. For a while, after a change of heart, he also studied with his father. Eventually, however, he chose medicine as a career instead and graduated in that discipline from Heidelberg in 1721 [5]. A subsequent falling out with his father caused Daniel to be banished from the family home.

Daniel is important for his contribution to at least four fields of interest to statisticians: stochastic processes, tests of significance, likelihood, and utility. As regards the former, his attempts to calculate the advantages of vaccination against smallpox are frequently claimed to be the earliest example of an epidemic model, although as Dietz and Heesterbeek have pointed out in their recent detailed examination [4], the model in question is static not dynamic. However, the example is equally interesting as a contribution to the literature on competing risk.

In an essay of 1734, one of several of Daniel's that won the prize of the Parisian Academy of Sciences, he calculates, amongst other matters, the probability that the coplanarity of the planetary orbits could have arisen by chance. Since the orbits are not perfectly coplanar, this involves his calculating the probability, under a null of perfect random distribution, of a result as extreme or more extreme than that observed. This example, rather than Arbuthnot's, is thus perhaps more properly regarded as a forerunner of the modern significance test [10].

More controversial is whether Daniel can be regarded as having provided the first example of the use of the concept of maximizing likelihood to obtain an estimate (*see Maximum Likelihood Estimation*). A careful discussion of Bernoulli's work of 1769 and 1778 on this subject and his friend and fellow Basler Euler's commentary of 1778 has been provided by Stigler [12].

Finally, Daniel Bernoulli's work on the famous St. Petersburg Paradox should be noted. This problem was communicated to Daniel by his cousin Nicholas II and might equally well have been discussed in the section Nicholas II. The problem was originally proposed by Nicholas to Montmort and concerns a game of chance in which B rolls a die successively and gets a reward from A, that is dependent on the number of throws x to obtain the first six. In the first variant, the reward is x crowns. This has

expectation six crowns. This is then regarded as a fair price to play the game. In the second variant, however, the reward is 2^{x-1} and this does not have a finite expectation, thus implying that one ought to be prepared to pay any sum at all to play the game [7]. Daniel's solution, published in the journal of the St. Petersburg Academy (hence the 'St. Petersburg Paradox') was to replace money value with utility. If this rises less rapidly than the monetary reward, a finite expectation may ensue. Daniel's resolution of his cousin's paradox is not entirely satisfactory and the problem continues to attract attention. For example, a recent paper by Pawitan includes a discussion [8].

References

- [1] Bell, E.T. (1953). *Men of Mathematics*, Vol. 1, Penguin Books, Harmondsworth.
- [2] Boyer, C.B. (1991). *A History of Mathematics*, (Revised by U.C. Merzbach), 2nd Edition, Wiley, New York.
- [3] Csörgő, S. (2001). Nicolaus Bernoulli, in *Statisticians of the Centuries*, C.C. Heyde & E. Seneta, eds, Springer, New York.
- [4] Dietz, K. & Heesterbeek, J.A.P. (2002). Daniel Bernoulli's epidemiological model revisited, *Mathematical Biosciences* **180**, 1.
- [5] Gani, J. (2001). Daniel Bernoulli, in *Statisticians of the Centuries*, C.C. Heyde & E. Seneta, eds, Springer, New York, p. 64.
- [6] Hald, A. (1990). *A History of Probability and Statistics and their Applications before 1750*, Wiley, New York.
- [7] Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*, 1st Edition, John Wiley & Sons, New York.
- [8] Pawitan, Y. (2004). Likelihood perspectives in the consensus and controversies of statistical modelling and inference, in *Method and Models in Statistics*, N.M. Adams, M.J. Crowder, D.J. Hand & D.A. Stephens, eds, Imperial College Press, London, p. 23.
- [9] Schneider, I. (2001). Jakob Bernoulli, in *Statisticians of the Centuries*, C.C. Heyde & E. Seneta, eds, Springer, New York, p. 33.
- [10] Senn, S.J. (2003). *Dicing with Death*, Cambridge University Press, Cambridge.
- [11] Shafer, G. (1997). Bernoulli, The, in *Leading Personalities in Statistical Sciences*, N.L. Johnson & S. Kotz, eds, Wiley, New York, p. 15.
- [12] Stigler, S.M. (1999). *Statistics on the Table*, Harvard University Press, Cambridge.

STEPHEN SENN

Binomial Confidence Interval

CLIFFORD E. LUNNEBORG

Volume 1, pp. 153–154

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Binomial Confidence Interval

Let $(x_1, x_2, \dots, x_n)$ be a random sample of size n from a population consisting of a proportion p of “successes” and a proportion $(1 - p)$ of “failures.” We observe s successes in the sample and now require a $(1 - \alpha)$ 100% **confidence interval** (CI) for p .

Often, the confidence interval is constructed as a complement to a set of null-hypothesis tests. If a hypothesized value, θ_0 , for a parameter θ lies within the bounds to a $(1 - \alpha)$ 100% CI for that parameter, then we ought not reject the hypothesis that $\theta = \theta_0$ at the α significance level. Equivalently, the $(1 - \alpha)$ 100% CI for θ can be defined as the set of values for θ_0 that cannot be rejected at the α significance level. This logic guided the definition in 1934 of what has come to be known as the Clopper–Pearson [2] or exact CI for the binomial parameter p .

Having observed s successes in a random sample of n observations, we would reject the hypothesis $p = p_0$ at the α significance level if, under this hypothesis, either the probability of observing s or fewer successes or the probability of observing s or more successes does not exceed $\alpha/2$. That is, we take the test to be a nondirectional one with the probability of a Type I error divided equally between the two directions.

The probability of s or fewer successes in a random sample of size n , where the probability of a success at each draw is p , is given by a sum of binomial terms,

$$\sum_{y=0, \dots, s} \left\{ \frac{n!}{[y! \times (n - y)!]} p^y (1 - p)^{n-y} \right\}, \quad (1)$$

and the probability of s or more successes by a second sum,

$$\sum_{y=s, \dots, n} \left\{ \frac{n!}{[y! \times (n - y)!]} p^y (1 - p)^{n-y} \right\}. \quad (2)$$

The upper and lower bounds to the $(1 - \alpha)$ 100% CI are given by the values of p that equate each of these sums to $\alpha/2$:

$$\frac{\alpha}{2} = \sum_{y=0, \dots, s} \left\{ \frac{n!}{[y! \times (n - y)!]} p U^y (1 - p_U)^{n-y} \right\}. \quad (3)$$

and

$$\frac{\alpha}{2} = \sum_{y=s, \dots, n} \left\{ \frac{n!}{[y! \times (n - y)!]} p_L^y (1 - p_L)^{n-y} \right\}. \quad (4)$$

Although the required proportions, p_L and p_U , cannot be found directly from the formulae, these bounds have been tabulated for a range of values of n , s , and α , for example, in [3].

Many statistical packages can routinely produce such exact binomial CIs. As an example, the **binom.test** function in the **R** statistical computing language (www.R-project.org) reports, for $n = 12$ and $s = 5$, lower and upper bounds to a 95% CI of 0.1516 and 0.7233.

Owing to the discreteness of the binomial random variable, however, these exact binomial CIs may not have the desired coverage probabilities. That is, CIs defined in this manner may not cover the true value of the success parameter $(1 - \alpha)$ 100% of the time. The use of a carefully selected normal approximation has been shown to improve coverage [1]. One of the approximations recommended, for example, by [1] and [4], is easily implemented. The lower and upper limits to a $(1 - \alpha)$ 100% binomial CI are approximated by

$$p_{\text{adj}} \pm z_{\alpha/2} \left[\frac{p_{\text{adj}}(1 - p_{\text{adj}})}{n} \right],$$

where $p_{\text{adj}} = (s + 2)/(n + 4)$ and $z_{\alpha/2}$ is the $(\alpha/2)$ quantile of the standard normal distribution. For example, 2.5% of the standard normal distribution falls below -1.96 .

For our example, $s = 5$ and $n = 12$, the approximated lower and upper bounds to a 95% CI for p are 0.1568 and 0.7182. In this instance, they differ only slightly from the exact binomial CI limits.

References

- [1] Agresti, A. & Coull, B.A. (1998). Approximate is better than ‘exact’ for interval estimation of binomial proportions, *The American Statistician* **52**, 119–126.
- [2] Clopper, C.J. & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial, *Biometrika* **26**, 404–413.
- [3] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [4] Garthwaite, P.H., Jolliffe, I.T. & Jones, B. *Statistical Inference*, 2nd Edition, Oxford University Press, Oxford.

CLIFFORD E. LUNNEBORG

Binomial Distribution: Estimating and Testing Parameters

VANCE W. BERGER AND YANYAN ZHOU

Volume 1, pp. 155–157

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Binomial Distribution: Estimating and Testing Parameters

Binomial Test

The binomial distribution (*see Catalogue of Probability Density Functions*) is generally used to model the proportion of binary trials (that can turn out one of two ways) in which a ‘success’ results. For example, each member of a given set of N adolescents may be classified as a smoker or as a nonsmoker. If some assumptions are met, then the number of smokers in this set follows the binomial distribution. The assumptions required are as follows:

1. Each adolescent may be classified as only a smoker or a nonsmoker, and the definition of what constitutes a smoker is common to all adolescents.
2. Each adolescent has the same probability, p , of being a smoker.
3. There is independence, in that the decision of one adolescent to smoke has no bearing on the decision of any other adolescent to smoke.

Each of these assumptions may be challenged. Regarding the first, it is certainly more informative to classify potential smokers on the basis of the amount they smoke, including possibly not at all. In fact, dichotomizing this inherently continuous smoking variable can result in a loss of **power** if the smoking variable is the dependent variable [3, 7, 8], or in a reversal of the direction of the effect if it is a predictor [6]. The second assumption would be violated if parents who smoke are more likely to have children who smoke. The third assumption, independence, seems questionable if smoking is the result of peer pressure to do so. Nevertheless, we proceed with the binomial distribution, because, sometimes, binomial data are the best (or only) data available. The observed number of smokers, r , in a sample of n adolescents, is represented as an observation on a random variable X , which, if all the assumptions are true, follows the binomial distribution with parameters n (the number of ‘trials’) and p (the ‘success’

probability):

$$P(X = r) = \frac{n!}{[r!(n-r)!]} p^r (1-p)^{n-r},$$
$$r = 0, 1, \dots, n. \quad (1)$$

While there are cases in which one would want to study and estimate the size of a population (what we call N), we consider cases in which N is known prior to conducting the experiment. In many of these cases, it is of interest to learn about the binomial proportion or success probability, p . One may calculate p quite simply as the ratio of the observed frequency of successes in the population to the size N of the population. If, however, it is not feasible to study an entire population and find the numerator of the aforementioned ratio, then p may not be calculated, and it must instead be estimated. The estimation of p is also fairly straightforward, as a sample of size n may be taken, hopefully a representative sample, and then this sample may serve as the population, so that p is calculated on the sample and then used as an estimate for the population.

Clearly, this estimate will not be a very good one if the sampling schemes systematically overrepresents some segments of the population relative to others [5]. For example, if one wanted to study the proportion of a given population that was on a diet, then the local gym would probably not be the best place to conduct the survey, as there could be a bias toward either inflated estimation (more health-conscious people diet and exercise) or deflated estimation (fewer people who exercise are overweight). This concern is beyond the scope of the present article, as it presupposes that there are recognizable subsets (in the example, members of a gym would be a recognizable subset) of the population with a success probability that differed from that of the population at large. If this were the case, then the heterogeneity would invalidate the second binomial assumption.

While estimation is a useful procedure in some contexts, it is also often useful to conduct a formal one-sample hypothesis test that specifies, as its null hypothesis, that the population success proportion p is equal to some prespecified number p_0 . For example, if a claim is made that a certain treatment can make it more or less likely that a child will be born a boy, then one might take the ‘null’ success probability to be 0.5 (to reflect the null state in which girls and boys are equally likely) and ask if 0.5 is still

2 Binomial Distribution

the success probability once this treatment is used. If so, then the treatment has not had its intended effect, but if not, then perhaps it has (depending on the direction of the shift). This analysis could be set up as a two-sided test:

$$\begin{aligned} H_0: p &= 0.5 \text{ versus} \\ H_A: p &\neq 0.5 \end{aligned} \quad (2)$$

or as a one-sided test in either of two directions:

$$\begin{aligned} H_0: p &= 0.5 \text{ versus} \\ H_A: p &> 0.5 \end{aligned} \quad (3)$$

or

$$\begin{aligned} H_0: p &= 0.5 \text{ versus} \\ H_A: p &< 0.5 \end{aligned} \quad (4)$$

To test any of these hypotheses, one would use the **binomial test**. The binomial test is based on the binomial distribution with the null value for p (in this case, the null value is $p = 0.5$), and whatever n happens to be appropriate. The binomial test is generally conducted by specifying a given significance level, α , although one could also provide a P value, thereby obviating the need to specify a given significance level. If α is specified and we are conducting a one-sided test, say with $H_A: p > 0.5$, then the rejection region will consist of the most extreme observations in the direction of the hypothesized effect. That is, it will take a large number of successes, r , to reject H_0 and conclude that $p > 0.5$.

There is some integer, k , which is termed the *critical value*. Then H_0 is rejected if, and only if, the number of successes, r , is at least as large as k . What the value of k is depends on α , as well as on n and p . The set of values of the random variable $\{X \geq k\}$ makes up a rejection region. The probability (computed under the null hypothesis) that the binomial random variable takes a value in the rejection region cannot exceed α and should be as close to α as possible. As a result, the critical value k is the smallest integer for which $P_0\{X \geq k\} \leq \alpha$. This condition ensures that the test is exact [1, 4].

Because the distribution of the number of successes is discrete, it will generally turn out that for the critical k $P_0\{X \geq k\}$ will be strictly less than α . That is, we will be unable to find a value of k for

which the null probability of a success count in the rejection region will be exactly equal to α . When $P_0\{X \geq k\} < \alpha$, the test is said to be conservative. That is, the probability of a Type I error will be less than an intended or nominal significance level α .

The actual level of significance, α' , is computed as

$$\begin{aligned} P_0\{X \geq k\} &= \sum_{r=k, \dots, n} \left\{ \frac{n!}{[r!(n-r)!]} p_0^r (1-p_0)^{n-r} \right\} \\ &= \alpha' \end{aligned} \quad (5)$$

and should be reported. The value of α' depends only on n , p_0 , and α (these three determine k) and can be computed as soon as these parameters are established.

It is a good idea to also report P values, and the P value can be found with the above formula, except replacing k with the observed number of successes. The discreteness of the binary random variable and consequent conservatism of the hypothesis test can be managed by reporting a P value interval [2].

This discussion was based on testing H_0 against $H_A: p > p_0$, but with an obvious modification it can be used also for testing H_0 against $H_A: p < p_0$. In this case, the rejection region would be on the opposite side of the distribution, as we would reject H_0 for small values of the binomial X . The modification for the two-sided test, $H_A: p \neq p_0$, is not quite as straightforward, as it requires rejecting H_0 for either small or large values of X . Finally, we mention that with increasing frequency one encounters tests designed to establish not superiority but rather **equivalence**. In such a case, H_0 would specify that p is outside a given equivalence interval, and the rejection region would consist of intermediate values of X .

As mentioned, the binomial test is an **exact test** [1, 4], but when $np \geq 5$ and $n(1-p) \geq 5$, it is common to use the normal distribution to approximate the binomial distribution. In this situation, the z -test may be used as an approximation of the binomial test.

References

- [1] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [2] Berger, V.W. (2001). The p-value interval as an inferential tool, *Journal of the Royal Statistical Society D (The Statistician)* **50**(1), 79–85.

-
- [3] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [4] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**(1), 74–82.
- [5] Berger, V.W., Rezvani, A. & Makarewicz, V. (2003). Direct effect on validity of response run-in selection in clinical trials, *Controlled Clinical Trials* **24**(2), 156–166.
- [6] Brenner, H. (1998). A potential pitfall in control of covariates in epidemiologic studies, *Epidemiology* **9**(1), 68–71.
- [7] Moses, L.E., Emerson, J.D. & Hosseini, H. (1984). Analyzing data from ordered categories, *New England Journal of Medicine* **311**, 442–448.
- [8] Rahlfs, V.W. & Zimmermann, H. (1993). Scores: ordinal data with few categories – how should they be analyzed? *Drug Information Journal* **27**, 1227–1240.

VANCE W. BERGER AND YANYAN ZHOU

Binomial Effect Size Display

ROBERT ROSENTHAL

Volume 1, pp. 157–158

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Binomial Effect Size Display

The binomial effect size display (BESD) was introduced in 1982 as an intuitively appealing general purpose display of the magnitude of experimental effect (*see Effect Size Measures*) [3]. Although there had been a growing awareness of the importance of estimating sizes of effects along with estimating the more conventional levels of significance, there was still a problem in interpreting various effect size estimators such as the **Pearson correlation r** . For example, experienced behavioral researchers and experienced statisticians were quite surprised when they were shown that the Pearson r of 0.32 associated with a coefficient of determination (r^2) of only 0.10 was the correlational equivalent of increasing a success rate from 34 to 66% by means of an experimental treatment procedure; for example, these values could mean that a death rate under the control condition is 66% but is only 34% under the experimental condition. There appeared to be a widespread tendency to underestimate the importance of the effects of behavioral (and biomedical) interventions simply because they are often associated with what are thought to be low values of r^2 [2, 3]. The interpretation of the BESD is quite transparent, and it is useful because it is (a) easily understood by researchers, students, and lay persons; (b) applicable in a wide variety of contexts; and (c) easy to compute.

The question addressed by the BESD is: What is the effect on the success rate (survival rate, cure rate, improvement rate, selection rate, etc.) of instituting a certain treatment procedure? It displays the change in success rate (survival rate, cure rate, improvement rate, selection rate, etc.) attributable to a certain treatment procedure. An example shows the appeal of the procedure.

In their **meta-analysis** of psychotherapy outcome studies, Smith and Glass [5] summarized the results of some 400 studies. An eminent critic stated that the results of their analysis sounded the ‘death knell’ for psychotherapy because of the modest size of the effect. This modest effect size was calculated to be equivalent to an r of 0.32 accounting for ‘only’ 10% of the variance.

Table 1 is the BESD corresponding to an r of 0.32 or an r^2 of 0.10. The table shows clearly that

Table 1 The binomial effect size display: an example ‘Accounting for Only 10% of the Variance’

Condition	Treatment outcome		Σ
	Alive	Dead	
Treatment	66	34	100
Control	34	66	100
Σ	100	100	200

Table 2 Binomial effect size displays corresponding to various values of r^2 and r

r^2	r	Success rate increased		Difference in success rates
		From	To	
0.01	0.10	0.45	0.55	0.10
0.04	0.20	0.40	0.60	0.20
0.09	0.30	0.35	0.65	0.30
0.16	0.40	0.30	0.70	0.40
0.25	0.50	0.25	0.75	0.50
0.36	0.60	0.20	0.80	0.60
0.49	0.70	0.15	0.85	0.70
0.64	0.80	0.10	0.90	0.80
0.81	0.90	0.05	0.95	0.90
1.00	1.00	0.00	1.00	1.00

it would be misleading to label as ‘modest’ an effect size equivalent to increasing the success rate from 34 to 66% (e.g., reducing a death rate from 66 to 34%).

Table 2 systematically shows the increase in success rates associated with various values of r^2 and r . Even so ‘small’ an r as 0.20, accounting for only 4% of the variance, is associated with an increase in success rate from 40 to 60%, such as a reduction in death rate from 60 to 40%. The last column of Table 2 shows that the difference in success rates is identical to r . Consequently, the experimental success rate in the BESD is computed as $0.50 + r/2$, whereas the control group success rate is computed as $0.50 - r/2$. When researchers examine the reports of others and no effect size estimates are given, there are many equations available that permit the computation of effect sizes from the sample sizes and the significance tests that have been reported [1, 4, 6].

References

- [1] Cohen, J. (1965). Some statistical issues in psychological research, in *Handbook of Clinical Psychology*, B.B. Wolman, ed., McGraw-Hill, New York, pp. 95–121.

2 Binomial Effect Size Display

- [2] Rosenthal, R. & Rubin, D.B. (1979). A note on percent variance explained as a measure of the importance of effects, *Journal of Applied Social Psychology* **9**, 395–396.
- [3] Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect, *Journal of Educational Psychology* **74**, 166–169.
- [4] Rosenthal, R. & Rubin, D.B. (2003). $r_{\text{equivalent}}$: A simple effect size indicator, *Psychological Methods* **8**, 492–496.
- [5] Smith, M.L. & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies, *American Psychologist* **32**, 752–760.
- [6] Wilkinson, L., Task Force on Statistical Inference. (1999). Statistical methods in psychology journals, *American Psychologist* **54**, 594–604.

ROBERT ROSENTHAL

Binomial Test

MICHAELA M. WAGNER-MENGHIN

Volume 1, pp. 158–163

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Binomial Test

The Binomial Distribution as Theoretical Model For Empirical Distributions

Several statistical tests are based on the assumption that all elements of a sample follow an identical distribution. Designing and analyzing experiments and surveys require us to select a suitable theoretical model (distribution) to represent the empirical distribution. The **binomial distribution** is a probability distribution and a suitable theoretical model for many of the empirical distributions we encounter in social science experiments and surveys. It describes the behavior of a count variable (more specifically, the probability of observing a particular value of the count variable) if the following conditions are met.

1. The number of observations or trials, n , is fixed. The number of trials resulting with the first of the two possible outcomes is x .
2. The observations are independent.
3. Each observation represents one of two outcomes. The probabilities corresponding to the two outcomes are p , for the first of the two possible outcomes and q , for the second of two possible outcomes. They add up to 1.0, so, often only the probability for p is given, and the probability for q is $1 - p$.
4. The probability of the outcomes is the same for each trial.

The process defined by the four conditions is also called *Bernoulli process* or *sequence of Bernoulli trials* (after Jacques Bernoulli, a Swiss mathematician, 1654–1705) (see **Bernoulli Family**), and some statistical text books will refer to the binomial distribution as the *Bernoulli distribution*.

Count variables we observe in social sciences include the number of female students affiliated to a specific department, the number of people volunteering for social work, the number of patients recovering after obtaining a new treatment, the number of subjects solving a puzzle in a problem-solving task. Note that there are many variables, which already represent one of two outcomes: There are either male or female students, people are volunteering or not, puzzles can be solved or not. Other variables may have to be transformed in order to represent only two possible outcomes.

In general, a probability distribution is specified by a probability function that describes the relationship between random events and random variables. With countable finite values for the random variable, the binomial distribution is classified as a discrete probability function. The binomial distribution is a function of n and p .

The resulting empirical distribution when tossing a coin several times is a typical example to illustrate the parameters n and p that specify a binomial distribution: When tossing several times (n = number of coin tosses), we observe the first of the two outcomes, the result head, for a certain number of tosses (x = number of trials with result head). X can randomly take one of the values of k ($k = 0, 1, 2, \dots, n$). Therefore, k is called *random variable*.

To use the binomial function as a theoretical model for describing this empirical function, we additionally need to specify the relationship between the random event of tossing the coin and the random variable. This assumption about the relationship is expressed as a probability (p) for one of the two possible outcomes and can be derived either theoretically or empirically. When tossing a coin, one usually has the idea (theory) of using a fair coin and one would expect the result ‘head’ in about 50% of all tosses. The theoretically derived probability for the value ‘head’ is therefore $p = 0.5$.

We also can see that the required conditions for using the binomial distribution as a theoretic model apply in this example: The number of observations of coin tosses is fixed, each coin toss can be performed independently from the other tosses, there are only two possible outcomes – either head or tail – and, unless we manipulate the coin between two tosses, the probability of the outcome is the same for each trial.

Formally expressed, we say: A variable x with probability function

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1)$$

for $k = 0, 1, 2, \dots, n$; $0 < p < 1$ is called a *binomial variable* whose distribution has parameters n and p .

We can now calculate the probability of observing the outcome ‘head’ x times in $n = 10$ coin tosses by entering the respective values of the binomial distribution ($n = 10$, $p = 0.5$, $x = 0, 1, 2, \dots, n$) in

2 Binomial Test

the probability function (1), as done here as an example for the value $x = 7$.

$$\begin{aligned}
 P(x = 7) &= \binom{10}{7} \times 0.5^7 \times (0.5)^3 \\
 &= \frac{10!}{7! \times (10 - 7)!} \times (0.5)^{10} \\
 &= \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1) \times (3 \times 2 \times 1)} \\
 &\quad \times (0.5)^{10} = 0.117. \quad (2)
 \end{aligned}$$

The probability of observing seven times ‘head’ in 10 trials is 0.117, which is a rather small probability, and might leave us with some doubt whether our coin is really a fair coin or the experimenter did the tosses really independently from each other. Figure 1(a) gives us the probability distribution for the binomial variable x . The text book of Cohen [1, p. 612] gives a simple explanation of what makes the binomial distribution a distribution: ‘The reason you have a distribution at all is that whenever there are n trials, some of them will fall into one category and some will fall into the other, and this division into categories can change for each new set of n trials.’

The resulting probability distribution for the coin toss experiment is symmetric; the value $p = 0.5$ is indicating this already. The distribution in Figure 1(b) illustrates what happens when the value for p increases, for example, when we have to assume that

the coin is not fair, but biased showing ‘head’ in 80% of all tosses.

There are some other details worth knowing about the binomial distribution, which are usually described in length in statistical text books [1, 4]. The binomial mean, or the expected count of success in n trials, is $E(X) = np$. The standard deviation is $\text{Sqrt}(npq)$, where $q = 1 - p$. The standard deviation is a measure of spread and it increases with n and decreases as p approaches 0 or 1. For any given n , the standard deviation is maximized when $p = 0.5$.

With increasing n , the binomial distribution can be approximated by the normal distribution (with or without continuity correction), when p indicates a symmetric rather than an asymmetric distribution. There is no exact rule when the sample is large enough to justify the normal approximation of the binomial distribution, however, a rule of thumb published in most statistical text books is that when p is not near 0.5, npq should be at least 9. However, according to a study by Osterkorn [3], this approximation is already possible when np is at least 10.

The Binomial Test

However, the possibility of deriving the probability of observing a special value x of a count variable by means of the binomial distribution might be interesting from a descriptive point of view. The

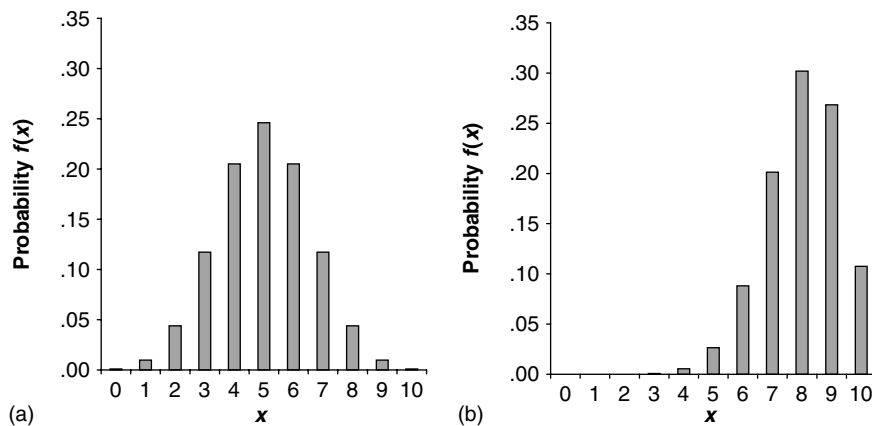


Figure 1 Probability function for the binomial distribution with $n = 10$ and $p = 0.5$ (a) coin toss example and $n = 10$ and $p = 0.8$ (b)

Table 1 Data of the coin toss experiment

Outcome		Observed ratio	Expected ratio
Outcome 1: 'head'	$k = 7$	$p1 = 0.7$	$p = 0.5$
Outcome 2: 'tail'	$n - k = 3$	$p2 = 0.3$	$1 - p = 0.5$
	$n = 10$		

value of the binomial distribution for planning and analyzing social science experiments comes with the possibility of using statistical tests based on the binomial distribution. If the binomial distribution is a suitable theoretical model for describing the expected empirical distribution, the binomial test can be applied. The binomial test is also known by the name *sign test*.

In the coin toss experiment with 10 trials introduced above, we observed 7 'heads' and 3 'tails'. On the basis of our assumption of a fair coin, we expected 5 'heads' and 5 'tails' (see Table 1).

One might argue that the observed deviation is small and due to chance. Using the binomial test, we can examine our argument by using the principle of testing statistical hypotheses: Predefining a probability for rejecting the null hypothesis (α) and comparing the observed probabilities for a statistical hypothesis to decide about rejecting the statistical hypothesis.

The Statistical Hypothesis of the Binomial Test

The binomial test gives us a probability for the assumption that the observed frequencies are equal to the expected frequencies. This probability can then be compared to the predefined α -level to decide about keeping or rejecting this assumption.

In our example, we propose the hypothesis that the observed probability for the outcome head (0.7) is no different than the theoretic expected probability for the outcome head (0.5).

How to Perform the Test

To perform a binomial test, we have to find the probability (P) that corresponds to a tail containing x or more extreme values. Formula 3 shows how this is done in general for the one-sided test, giving the probability that the observed ratio for outcome 1 is

smaller than or equal to the expected ratio:

$$P = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}. \quad (3)$$

As in our example, the observed ratio for outcome 1 'head' is already larger than the expected ratio, testing this side of the hypothesis makes no sense. Inserting $k = 7$ and $p = 0.5$ from Table 1 in (3) will not yield a useful result. We need to transform the problem to apply the test: $k' = N - k = 3$, $p' = 1 - p = 0.5$ and we are now testing the hypothesis that the observed probability for the outcome 'not head' (0.3) is the same as the theoretical expected probability for the outcome 'not head' (0.5). Formally expressed, we write: $H_0: p = 0.5$; $H_A: p < 0.5$; $\alpha = 0.05$. Using only a calculator, we can perform an exact binomial test for this problem by summarizing the following probabilities:

$$\begin{aligned} P(x = 0) &= \binom{10}{0} \times 0.5^7 \times (0.5)^3 = 0.001 \\ P(x = 1) &= \binom{10}{1} \times 0.5^7 \times (0.5)^3 = 0.0098 \\ P(x = 2) &= \binom{10}{2} \times 0.5^7 \times (0.5)^3 = 0.0439 \\ P(x = 3) &= \binom{10}{3} \times 0.5^7 \times (0.5)^3 = 0.1172 \\ P &= 0.1719. \end{aligned} \quad (4)$$

The probability to observe 3 or fewer times the outcome 'not head' in 10 trials is $P = 0.1719$. This value is larger than the significance level of 0.05, and therefore we keep the hypothesis of equal observed and expected ratio. According to this result, we have a fair coin.

Performing the exact binomial test for a larger sample is rather difficult and time consuming (see **Exact Methods for Categorical Data**). Statistic text books therefore recommend the asymptotic binomial test, on the basis of the fact that the binomial distribution can be approximated by the normal distribution when the sample is large, thus allowing a more convenient z-test [1, 4].

4 Binomial Test

Table 2 Data of the ‘representative sample’ example. Statistical hypothesis: The observed ratio of males and females is equal to the expected ratio. $H_0: p = 0.5$; $H_A: p \neq 0.5$; $\alpha = 0.05$

Sex	Observed ratio	Expected ratio
Group 1: ‘male’	$k = 16$	$p1 = 0.17$
Group 2: ‘female’	$n - k = 76$	$p2 = 0.83$
	$n = 92$	$1 - p = 0.5$

Note: Asymptotic significance, two-tailed: $P = 0.000$
 $P = 0.000 < \alpha = 0.05$, reject H_0 .

Some Social Science Examples for Using the Binomial Test Using SPSS

Example 1: A natural grouping variable consisting of two mutually exclusive groups 92 subjects volunteered in a research project. 16 of them are males, 76 are females. The question arises whether this distribution of males and females is representative or whether the proportion of males in this sample is too small (Table 2).

Using SPSS, the syntax below, we obtain a P value for an asymptotic significance (two-tailed), which is based on the z-approximation of the binomial distribution: $P = 0.000$.

NPART TEST

/BINOMIAL (0.50) = sex.

The decision to calculate the two-tailed significance is done automatically by the software, whenever the expected ratio is 0.5. Still, we can interpret a one-sided hypothesis. As the distribution is symmetric, when $p = 0.5$, all we have to do is to divide the obtained P value by two. In our example, the P value is very small, indicating that the statistical hypothesis of equal proportions is extremely unlikely ($P = 0.000 < \alpha = 0.05$, reject H_0). We therefore reject the H_0 and assume that the current sample is not representative for males and females. Males are underrepresented.

Example 2: Establishing two groups by definition

In Example 1, we used a natural grouping variable consisting of two mutually exclusive groups. In Example 2, we establish the two groups on the basis of empirical information.

For the assessment of aspiration level, 92 subjects volunteered to work for 50 sec on a speeded

Table 3 Data of the ‘aspiration level’ example. Statistical hypothesis: The observed ratio of optimistic performance prediction is equal to the expected level of performance prediction. $H_0: p = 0.5$; $H_A: p \neq 0.5$; $\alpha = 0.05$

Group	Observed ratio	Expected ratio
Group 1: ‘optimistic’	$k = 65$	$p1 = 0.71$
Group 2: ‘pessimistic’	$n - k = 27$	$p2 = 0.29$
	$n = 92$	$1 - p = 0.5$

Note: Asymptotic significance, two-tailed: $P = 0.000$
 $P = 0.000 < \alpha = 0.05$, reject H_0 .

symbol-coding task (Table 3). Afterward, they were informed about their performance (number of correct), and they provided a performance prediction for the next trial. We now use the binomial test to test whether the ratio of subjects with optimistic performance prediction (predicted increase of performance = group 1) and pessimistic performance prediction (predicted decrease of performance = group 2) is equal.

The significant result indicates that we can reject the null hypothesis of equal ratio between optimistic and pessimistic performance prediction. The majority of the subjects expect their performance to increase in the next trial.

Example 3: Testing a hypothesis with more than one independent binomial test

As we know, our sample of $n = 92$ is not representative regarding males and females (see Example 1). Thus, we might be interested in testing whether the tendency to optimistic performance prediction is the same in both groups. Using the binomial test, we perform two statistical tests to test the same hypothesis (Table 4). To avoid accumulation of statistical error, we use Bonferroni-adjustment (see **Multiple Comparison Procedures**) $\alpha' = \alpha/m$ (with m = number of statistical tests performed to test one statistical hypothesis) and perform the binomial tests by using the protected α level (see [2] for more information on Bonferroni-adjustment).

The significant result for the females indicates that we can reject the overall null hypothesis of equal ratio between optimistic and pessimistic performance prediction for males and for females. Although not further discussed here, on the basis of the current

Table 4 Data of the ‘aspiration level’ example – split by gender. Statistical hypothesis: The observed ratio of optimistic performance prediction is equal to the expected level of performance prediction. Test 1: male; Test 2: female $H_0: p = 0.5$ for males and for females; $H_A: p \neq 0.5$ for males and/or females; $\alpha = 0.05$

		Observed ratio	Expected ratio
Male			
Group 1: ‘optimistic’	$k = 12$	$p1 = 0.75$	$p = 0.5$
Group 2: ‘pessimistic’	$n - k = 4$	$p2 = 0.25$	$1 - p = 0.5$
	$n = 16$		
Exact significance, two-tailed: $P = 0.077$			

		Observed ratio	Expected ratio
Female			
Group 1: ‘optimistic’	$k = 53$	$p1 = 0.70$	$p = 0.5$
Group 2: ‘pessimistic’	$n - k = 23$	$p2 = 0.30$	$1 - p = 0.5$
	$n = 76$		
Asymptotic significance, two-tailed: $P = 0.001$			

Note: Adjusted $\alpha: \alpha' = \alpha/m$; $\alpha = 0.05$, $m = 2$, $\alpha' = 0.025$; $P(\text{male}) = 0.077 > \alpha' = 0.025$, retain H_0 for males; $P(\text{female}) = 0.001 < \alpha' = 0.025$, reject H_0 for females.

data, we may assume equal proportion of optimistic and pessimistic performance predictions for males, but not for females. The current female sample tends to make more optimistic performance predictions than the male sample.

Example 4: Obtaining p from another sample

The previous examples obtained p for the binomial test from theoretical assumptions (e.g., equal distribution of male and females; equal distribution of optimistic and pessimistic performance). But there are other sources of obtaining p that might be more interesting in social sciences. One source for p might be a result found with an independent sample or published in literature.

Continuing our aspiration level studies, we are interested in situational influences and are now collecting data in real-life achievement situations

Table 5 Data of the ‘comparing volunteers and applicants’ example. Statistical hypothesis: Applicants’ (observed) ratio of optimistic performance prediction is equal to the expected level of performance prediction. $H_0: p = 0.7$; $H_A: p > 0.7$; $\alpha = 0.05$

Group		Observed ratio	Expected ratio
Group 1: ‘optimistic’	$k = 21$	$p1 = 0.87$	$p = 0.7$
Group 2: ‘pessimistic’	$n - k = 3$	$p2 = 0.13$	$1 - p = 0.3$
	$n = 24$		

Note: Exact significance, one-tailed: $P = 0.042$
 $P = 0.042 < \alpha = 0.05$, reject H_0 .

rather than in the research-lab situation. 24 females applying for a training as air traffic control personnel did the coding task as part of their application procedure. 87% of them gave an optimistic prediction.

Now we remember the result of Example 3. In the research lab, 70% of the females had made an optimistic performance prediction, and we ask whether these ratios of optimistic performance prediction are the same (Table 5).

On the basis of the binomial test result, we reject the null hypothesis of equal ratios of optimistic performance prediction between the volunteers and the applicants. Female applicants are more likely to give optimistic than pessimistic performance predictions.

References

- [1] Cohen, B.H. (2001). *Explaining Psychological Statistics*, 2nd Edition, Wiley, New York.
- [2] Haslam, S.A. & McGarty, C. (2003). *Research Methods and Statistics in Psychology*, Sage Publications, London.
- [3] Osterkorn, K. (1975). Wann kann die Binomial und Poissonverteilung hinreichend genau durch die Normalverteilung ersetzt werden? [When to approximate the Binomial and the Poisson distribution with the Normal distribution?], *Biometrische Zeitschrift* **17**, 33–34.
- [4] Tamhane, A.C. & Dunlop, D.D. (2000). *Statistics and Data Analysis: From Elementary to Intermediate*, Prentice Hall, Upper Saddle River.

MICHAELA M. WAGNER-MENGHIN

Biplot

JOHN GOWER

Volume 1, pp. 163–164

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Biplot

Biplots are of two basic kinds (a) those concerned with a multivariate data matrix of a sample of n cases with observations on p variables (see **Multivariate Analysis: Overview**) and (b) those concerned with a single variable classified into a two-way table with r rows and c columns. In both cases, the aim is to give a visual representation in a few dimensions that approximates the values of the data matrix or table, as the case may be. The *bi* in biplots denotes the two modes: in case (a) samples and variables and in case (b) rows and columns. Thus, biplots are not necessarily two dimensional, though they often are. Biplots are useful for detecting patterns possibly suggesting more formal analyses and for displaying results found by more formal methods of analysis. In the following, cases (a) and (b) are treated briefly; for further details see Gower and Hand [3].

The simplest and commonest form of multivariate biplot uses **principal component analysis** (PCA) to represent the cases by n points and the variables by p vectors with origin 0. Then, the length of the projection of the point representing the i th case onto the j th vector predicts the observed (i, j) th value [2]. Gower and Hand [3] replace the vectors by bipolar axes equipped with numerical scales: then the values associated with the projections are immediate. This biplot is essentially a generalized **scatterplot** with more axes, necessarily nonorthogonal, than dimensions. The biplot axes are approximations of the full Cartesian representations with p orthogonal axes, where intercase distances are given by Pythagoras' distance formula (Figure 1). Similar approaches may be used with other distances and other methods of **multidimensional scaling**, when the scale markers on the axes may become nonlinear (see e.g., a logarithmic scale) and the axes themselves may be nonlinear.

Categorical variables, not forming a continuous scale, need special consideration. Each is represented by a set of *category-level-points* (CLPs) one for each category-level; CLPs for ordered categories are collinear. In contrast to a numerical scale, CLPs are labeled by the names of their category levels. In exact representations, the point representing a case is nearer the labels that give the values of its variables than to any other labels. In approximations, we need to predict what are the nearest

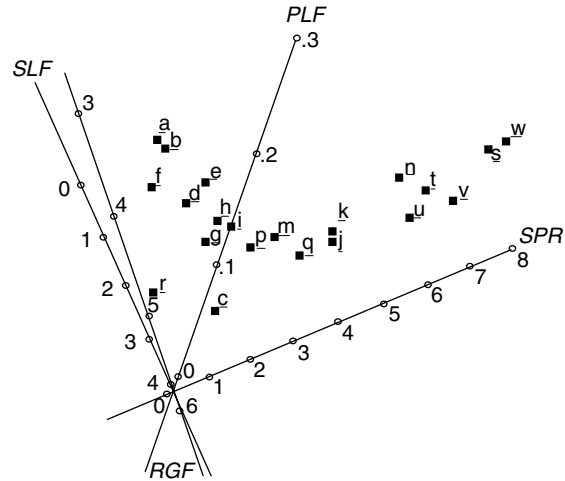


Figure 1 A principal component biplot with 21 cases (aircraft), labeled alphabetically, and four quantitative variables (RGF, SPR, PLF, SLF) referring to performance represented as four scaled biplot axes. This is a version of Figure 2.8 of Gower and Hand [3] modified to emphasize the close relationship between biplot axes and conventional coordinate axes. Predicted values are read off by projecting onto the axes, in the usual way

CLPs and this is done by creating a set of convex regions known as *prediction regions*; a point representing a case is then predicted as having the labeled categories of the prediction regions within which it lies. This setup applies to multiple correspondence analysis (MCA) with its dependence on chi-squared distance but it also applies to any form of distance defined for categorical variables. Figure 2 shows a biplot for ordered categorical variables. Numerical and categorical variables may be included in the same biplot. As well as prediction, sometimes one wishes to add a new sample to a multidimensional scaling display. This interpolation requires a different set of biplot axes than those required for prediction.

A two-way table can always be treated as a data matrix and the previously discussed forms of biplot used. However, the cases and variables of a multivariate data matrix are logically different noninterchangeable entities, whereas the rows and columns of a two-way table may be interchanged. Rows and columns are given parity when each is represented by sets of points without scaled axes. For categorical variables, the table is a **contingency table** and we have **correspondence analysis** biplots. For a

2 Biplot

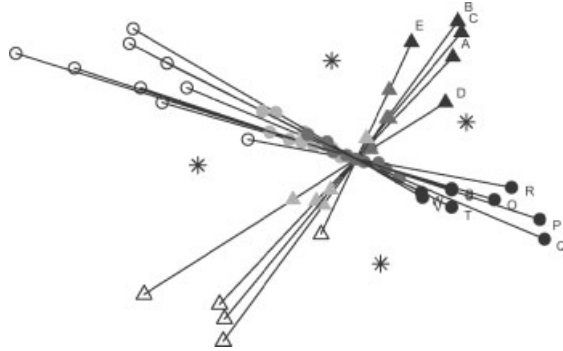


Figure 2 This is a nonlinear principal components biplot, showing 14 ordered categorical variables, A–E giving attitudes to regional identity and O–W giving attitudes to national identity. Each variable has four levels, ‘very’, ‘somewhat’, ‘a little’ and ‘none’ shown as black, dark gray, light gray and unshaded markers, triangular for regional identity and circular for national identity. Just four, of nearly 1000, cases are illustrated. Predictions associated with a case are given by the labels of the nearest markers on each axis. The data refer to Great Britain in the International Social Survey Program for 1995 and are fully discussed by Blasius and Gower [1]

tabulated quantitative variable, the biplot represents the multiplicative interaction term of a biadditive model. In both cases, interpretation is by evaluating the inner product $OR_i \times OC_j \times \cos(R_i OC_j)$, where R_i and C_j are the points representing the i th row and j th column, respectively.

References

- [1] Blasius, J. & Gower, J.C. (2005). Multivariate predictions with nonlinear principal components analysis: application, *Quality & Quantity* **39**, (in press).
- [2] Gabriel, K.R. (1971). The biplot graphical display of matrices with application to principal components analysis, *Biometrika* **58**, 453–467.
- [3] Gower, J.C. & Hand, D.J. (1996). *Biplots*, Chapman & Hall, London.

JOHN GOWER

Block Random Assignment

VANCE W. BERGER

Volume 1, pp. 165–167

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Block Random Assignment

The validity of the comparison between any two treatment groups is predicated on the comparability of these two treatment groups in all relevant respects other than the treatments under study (*see Clinical Trials and Intervention Studies*). Without this condition, any observed differences in the outcomes across the treatment groups can be attributed to either the treatments or the underlying differences having nothing to do with the treatments. To definitively attribute observed differences to the treatments themselves (the intended conclusion), all competing potential attributions must be falsified. For example, if alcohol consumption is observed to be positively associated with certain types of cancer, but alcohol consumption is also observed to be positively associated with tobacco consumption, which in turn is positively associated with these same types of cancer, then tobacco use is a confounder (*see Confounding Variable*).

Without additional information, it is impossible to distinguish between alcohol being truly carcinogenic and alcohol being nothing more than a benign correlate of tobacco use. These are the two extremes explanations, and in the middle one would consider explanations involving attributable fractions. Clearly, confounding can lead to complicated analyses and interpretations, and hence steps are often taken to control, minimize, or eliminate confounding. One such step is **randomization**, which can be conducted for any of several reasons, but one of the best is the creation of comparable comparison groups. Often, discussions of confounding end as soon as randomization is mentioned. That is, there appears to be a widely held belief that randomization by itself can eliminate all confounding, and ensure that any baseline differences between comparison groups formed by randomization are necessarily random and, by implication, unimportant. These views are not only wrong, but they are also dangerous.

In reality, there are many types of randomization, and each is susceptible to various types of bias. Prominent among these various biases are chronological bias resulting from time trends and selection bias resulting from prediction of upcoming allocations. To understand these biases, consider one of

the simplest types of randomization, specifically the random allocation rule [7] by which randomization is unrestricted other than having to adhere to specified allocation proportions. For example, if there are 200 subjects to be randomized, 100 to each group, then there are $200!/(100!)^2$ possible ways to split the 200 subjects evenly between the two treatment groups. The random allocation rule assigns equal probability, $(100!)^2/200!$, to each of these possibilities. This randomization procedure allows for long runs of consecutive subjects being allocated to the same treatment condition, so at some point during the study there may be grossly unbalanced numbers of subjects allocated to the two treatment groups.

If, in addition to a gross imbalance, there is also a time trend, then this can lead to imbalance with regard to not only numbers allocated at some point in time but also a covariate, and this imbalance may well remain at the end of the study. For example, at some point during the study, the composition of the subjects entering the study may change because of some external factor, such as new legislation or the approval of a new drug. This may make ‘early’ patients older or younger, or more or less likely to be male or female than ‘late’ patients, in which case this imbalance over time is transferred to imbalance across treatment groups. To prevent chronological bias from occurring, randomization is often restricted, so that at various points in time the number of subjects randomized to each treatment group is the same. This is referred to as blocking, or using randomized blocks, or permuted blocks. Each block is a set of subjects enrolled between consecutive forced returns to perfect balance. Note that the random allocation rule is a randomized block design with one large block, so it is more resistant to chronological bias than is the unrestricted randomization that results from assigning each subject on the basis of a fair coin toss, as the latter would not even ensure comparable group sizes at the end.

When randomized blocks are used, generally randomization is stratified by the block, meaning that the randomization in any one block is independent from the randomization in any other block. This is not always the case, however, as a study of etanercept for children with juvenile rheumatoid arthritis [5] used blocks within two strata, and corresponding blocks in the two strata were mirror images of each other [1].

In the usual case, not only are the blocks independent of each other, but also randomization within

2 Block Random Assignment

any block occurs by the random allocation rule. The block size may be fixed or random. If the block size is fixed at 4, for example, then there are six admissible sequences per block, specifically AABB, ABAB, ABBA, BAAB, BABA, and BBAA, so each would be selected (independently) with probability 1/6 for each block. There are two admissible sequences per block of size 2 (AB and BA), and 20 admissible sequences within each block of size 6. Of course, these numbers would change if the allocation proportions were not 1 : 1 to the two groups, or if there were more than two groups. Certainly, randomized blocks can handle these situations. Following (Table 1) is an example of blocked randomization with 16 subjects and a treatment assignment ratio of 1 : 1. Three randomization schedules are presented, corresponding to fixed block sizes of 2, 4, and 6. One could vary the block sizes by sampling from each of the columns in an overall randomization scheme. For example, the first six subjects could constitute one block of size 6, the next two could be a block of size 2, and the last eight could be two blocks of size 4 each. Note that simply varying the block sizes does not constitute random block sizes. As the name would suggest, random block sizes means that the block sizes not only vary but are also selected by a random mechanism [2].

The patterns within each block are clear. For example, when the block size is 2, the second

assignment depends entirely on the first assignment within each block; when the block size is 4, the fourth assignment is always predetermined, and the third may be as well. The first two are not predetermined, but the second will tend to differ from the first (they will agree in only two of the six possible sequences).

Smaller block sizes are ideal for controlling chronological bias, because they never allow the treatment group sizes to differ appreciably. In particular, the largest imbalance that can occur in any stratum when randomized blocks are used sequentially within this stratum (that is, within the stratum one block starts when the previous one ends) is half the largest block size, or half the common block size if there is a common block size. If chronological bias were the only consideration, then any randomized block study of two treatments with 1 : 1 allocation would be expected to use a common block size of 2, so that at no point in time could the treatment group sizes ever differ by more than one. Of course, chronological bias is not the only consideration, and blocks of size two are far from ideal. As mentioned, as the block size increases, the ability to predict upcoming treatment assignments on the basis of knowledge of the previous ones decreases.

This is important, because prediction leads to a type of selection bias that can interfere with **internal**

Table 1 Examples of Block Random Assignment

Subject	Block size = 2		Block size = 4		Block size = 6	
	Block	Treatment	Block	Treatment	Block	Treatment
1	1	A	1	A	1	A
2		B		A		A
3	2	A		B		B
4		B		B		A
5	3	B	2	A		B
6		A		B		B
7	4	B		B	2	B
8		A		A		B
9	5	B	3	A		A
10		A		B		B
11	6	A		A		A
12		B		B	3	A
13	7	A	4	B		B
14		B		B		B
15	8	A		A		B
16		B		A		A

validity, even in randomized trials. Note that a parallel is often drawn between randomized trials (that use random allocation) and random samples, as ideally each treatment group in a randomized trial constitutes a random sample from the entire sample. This analogy requires the sample to be formed first, and then randomized, to be valid, and so it breaks down in the more common situation in which the recruitment is sequential over time. The problem is that if a future allocation can be predicted and the subject to be so assigned has yet to be selected, then the foreknowledge of the upcoming allocation can influence the decision of which subject to select. Better potential responders can be selected when one treatment is to be assigned next and worse potential responders can be selected when another treatment is to be assigned next [3]. This selection bias can render analyses misleading and estimates unreliable.

The connection between blocked designs and selection bias stems from the patterns inherent in the blocks and allowing for prediction of future allocations from past ones. Clearly, the larger the block size, the less prediction is possible, and so if selection bias were the only concern, then the ideal design would be the random allocation rule (that maximizes the block size and minimizes the number of blocks), or preferably even unrestricted randomization. But there is now a trade-off to consider between chronological bias, which is controlled with small blocks, and selection bias, which is controlled with large blocks. Often this trade-off is addressed by varying the block sizes. While this is not a bad idea, the basis for doing so is often the mistaken belief that varying the block sizes eliminates all prediction of future allocations, and, hence, eliminates all selection bias. Yet varying the block sizes can, in some cases, actually lead to more prediction of future allocations than fixed block sizes would [4, 6].

A more recent method to address the trade-off between chronological bias and selection bias is the maximal procedure [4], which is an alternative to randomized blocks. The idea is to allow for any allocation sequence that never allows the groups sizes to differ beyond an acceptable limit. Beyond this, no additional restrictions, in the way of forced returns to perfect balance, are imposed. In many ways, the maximal procedure compared favorably to randomized blocks of fixed or varied size [4].

References

- [1] Berger, V. (1999). FDA product approval information—licensing action: statistical review, <http://www.fda.gov/cber/products/etanimm052799.htm> accessed 3/7/02.
- [2] Berger, V.W. & Bears, J.D. (2003). When can a clinical trial be called “randomized”? *Vaccine* **21**, 468–472.
- [3] Berger, V.W. & Exner, D.V. (1999). Detecting selection bias in randomized clinical trials, *Controlled Clinical Trials* **20**, 319–327.
- [4] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through Less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.
- [5] Lovell, D.J., Giannini, E.H., Reiff, A., Cawkwell, G.D., Silverman, E.D., Nocton, J.J., Stein, L.D., Gedalia, A., Ilowite, N.T., Wallace, C.A., Whitmore, J. & Finck, B.K. (2000). Etanercept in children with polyarticular juvenile rheumatoid arthritis, *The New England Journal of Medicine* **342**, 763–769.
- [6] Rosenberger, W. & Lachin, J.M. (2001). *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons.
- [7] Rosenberger, W.F. & Rukhin, A.L. (2003). Bias properties and nonparametric inference for truncated binomial randomization, *Nonparametric Statistics* **15**(4–5), 455–465.

VANCE W. BERGER

Boosting

ADELE CUTLER

Volume 1, pp. 168–169

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Boosting

Boosting was introduced by Freund and Schapire in 1996 [2], who developed the *Adaboost* algorithm. Adaboost is an ensemble classifier, which works by voting the predictions of many individual classifiers (*see Discriminant Analysis*). More information about voting ensembles is given in the related article on **bagging**. Another popular boosting algorithm is the LogitBoost algorithm of Friedman et al. 2000 [4]. The main benefit of boosting is that it often improves prediction accuracy, however, the resulting classifier is not interpretable.

The idea behind boosting is to form an ensemble by fitting simple classifiers (so-called base learners) to weighted versions of the data. Initially, all observations have the same weight. As the ensemble grows, we adjust the weights on the basis of our

knowledge of the problem. The weights of frequently misclassified observations are increased, while those of seldom-misclassified observations are decreased. Heuristically, we force the classifiers to tailor themselves to the hard-to-classify cases, and hope that the easy-to-classify cases will take care of themselves.

The base learners vary according to the implementation, and may be as simple as the so-called stumps classifier, which consists of a decision tree with only one split; cases with low values of the split variable are classified as one class and those with high values are classified as the other class. The base learner must allow us to weight the observations. Alternatively, we can randomly sample cases with replacement, with probabilities proportional to the weights, which makes boosting appear similar to bagging. The critical difference is that in bagging, the members of the ensemble are independent; in boosting, they are not.

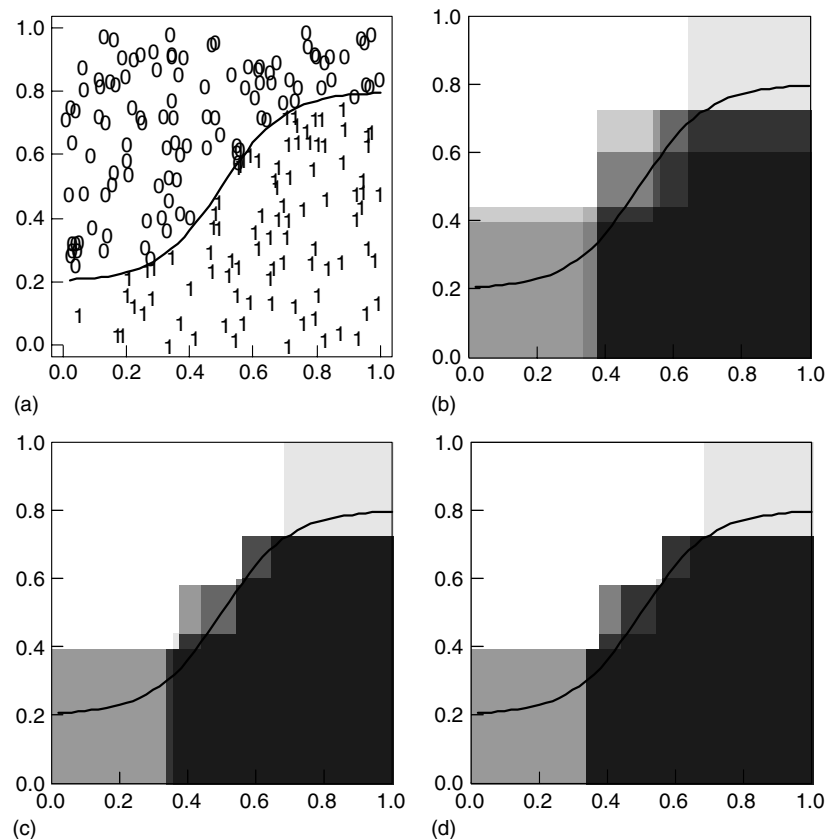


Figure 1 (a) Data and underlying function; (b) 10 boosted stumps; (c) 100 boosted stumps; (d) 400 boosted stumps

Once we have the ensemble, we combine them by weighted voting. These weights are chosen so that highly accurate classifiers get more weight than do less accurate ones. Again, the particular choice of weights depends on the implementation.

To illustrate, we use the R [8] function Logit-Boost [1] to fit a classifier to the data in Figure 1. The classification boundary and the data are given in Figure 1(a). In Figure 1(b), (c), and (d), the shading intensity indicates the weighted vote for class 1. As more stumps are included, the boosted classifier obtains a smoother, more accurate estimate of the classification boundary; however, it is still not as accurate as bagged trees (see Figure 1d in the article on bagging).

Recent theoretical work, for example [6], has shown that Adaboost is not consistent. That is, there are examples for which Adaboost does not converge to the optimal classification rule. In practice, this means that Adaboost will overfit the data in noisy situations, so it should not be used without some form of **cross validation** to prevent overfitting. For example, we might run the algorithm until the prediction error on a cross validation test set starts to increase.

An application of boosting to dementia data is described in the article on bagging.

Statistical references on boosting include [3], [4] and [5] and a machine learning summary is given in [7]. Related methods include bagging and **random forests**.

References

- [1] Dettling, M. & Bühlmann, P. (2003). Boosting for tumor classification with gene expression data, *Bioinformatics* **19**(9), 1061–1069.
- [2] Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm, in *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156.
- [3] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**, 1189–1232.
- [4] Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting, *Annals of Statistics* **28**, 337–407, (with discussion).
- [5] Hastie, T., Tibshirani, R. & Friedman, J.H. (2001). *The Elements of Statistical Learning*, Springer-Verlag.
- [6] Jiang, W. (2001). Some theoretical aspects of boosting in the presence of noisy data, in *Proceedings of the Eighteenth International Conference on Machine Learning*, 234–241.
- [7] Meir, R. & Ratsch, G. (2003). An introduction to boosting and leveraging, in *Lecture Notes in Artificial Intelligence: Advanced lectures on machine learning*, Springer-Verlag.
- [8] R Development Core Team (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, www.R-project.org.

(See also **Neural Networks; Pattern Recognition**)

ADELE CUTLER

Bootstrap Inference

A.J. CANTY AND ANTHONY C. DAVISON

Volume 1, pp. 169–176

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bootstrap Inference

The data analyst buys insight with assumptions – for example that data are independent, or follow a specified distributional form, or that an estimator has a given distribution, typically obtained by a mathematical argument under which the sample size tends to infinity (see **Maximum Likelihood Estimation**). Sometimes these assumptions may be justified by background knowledge or can be checked empirically, but often their relevance to the situation at hand is questionable, and, if they fail, it can be unclear how to proceed. Bootstrap methods are a class of procedures that may enable the data analyst to produce useful results nonetheless.

Consider, for instance, Table 1, which shows how an integer measure `hand` of the degree of left-handedness of $n = 37$ individuals varies with a genetic measure `dnan`. Figure 1 shows a strong positive relation between them, which is tempting to summarize using the sample product–moment correlation coefficient (see **Pearson Product Moment Correlation**), whose value is $\hat{\theta} = 0.509$. If the sample is representative of a population, we may wish to give a **confidence interval** for the population correlation. Exact intervals are hard to calculate, but a standard approximation yields a 95% confidence interval of (0.221, 0.715). This, however, presupposes that the population distribution is bivariate normal (see **Catalogue of Probability Density Functions**), an assumption contradicted by Figure 1. What then is the status of the interval? Can we do better?

Data Resampling

A key aspect of non-Bayesian statistical inference is the use of the sampling variability of an estimator, $\hat{\theta}$, to build inferences for the quantity of interest, θ . This entails using a model, for example, that the observations $y_1, \dots, y_n$ are drawn independently from an underlying population distribution F . In Table 1, each observation y consists of a pair (`dnan`, `hand`), and thus F has support in a subset of the plane. If F were known, the sampling variability of θ could be found either by a theoretical calculation, if this were possible, or empirically, by simulation from F . This latter method entails generating a sample

$y_1^*, \dots, y_n^*$ from F , computing the corresponding value $\hat{\theta}^*$ of the estimator, repeating this sampling procedure R times, and using the resulting replicates $\hat{\theta}_1^*, \dots, \hat{\theta}_R^*$ to estimate the distribution of $\hat{\theta}$. These two possibilities are illustrated in (a) of Figure 2, which superimposes the theoretical density of $\hat{\theta}$ under a fitted bivariate-normal distribution upon a histogram of $R = 10\,000$ values of $\hat{\theta}^*$ simulated from the same bivariate normal distribution. As one would expect, the two are close: if R were increased to infinity, then the histogram would converge to the theoretical density. The power of modern computing means that R is limited only by our impatience for results, but taking R huge is of little use here, because the bivariate-normal distribution fits the data badly. An alternative is to estimate the population F directly from the data using the empirical distribution function $\hat{F}$, which puts probability mass n^{-1} on each of the $y_1, \dots, y_n$. This discrete distribution is in fact the nonparametric **maximum-likelihood estimate** of F : in the present case, it puts masses on the points in Figure 1, proportional to the multiplicities, and with unit total mass.

The bootstrap idea is to use resampling from the fitted distribution $\hat{F}$ to mimic sampling from the distribution F underlying the data. To perform this we take samples $y_1^*, \dots, y_n^*$ of size n with replacement from the original observations $y_1, \dots, y_n$. Thus, for the handedness data, a bootstrap sample is created by taking $n = 37$ pairs (`dnan`, `hand`) from those in Table 1 with replacement and equal probability. Repetitions of the original observations will occur in the bootstrap samples, and though not a problem here, this can have repercussions when the statistic depends sensitively on the exact sample values; in some cases, smoothing or other modifications to the resampling scheme may be required. Panel (b) of Figure 2 shows a histogram of values $\hat{\theta}^*$ computed from 10 000 bootstrap samples, together with the probability density of $\hat{\theta}$ under the bivariate-normal distribution fitted above. The sharp difference between them would not disappear even with an infinite simulation, because the data are nonnormal – the histogram better reflects the sampling variability of the correlation coefficient under the unknown true distribution F .

The bootstrap replicates can be used to estimate sampling properties of $\hat{\theta}$ such as its bias and variance,

2 Bootstrap Inference

Table 1 Data from a study of handedness; hand is an integer measure of handedness and dnan a genetic measure. Data due to Dr Gordon Claridge, University of Oxford

	dnan	hand	dnan	hand	dnan	hand	dnan	hand
1	13	1	11	28	1	21	29	2
2	18	1	12	28	2	22	29	1
3	20	3	13	28	1	23	29	1
4	21	1	14	28	4	24	30	1
5	21	1	15	28	1	25	30	1
6	24	1	16	28	1	26	30	2
7	24	1	17	29	1	27	30	1
8	27	1	18	29	1	28	31	1
9	28	1	19	29	1	29	31	1
10	28	2	20	29	2	30	31	1

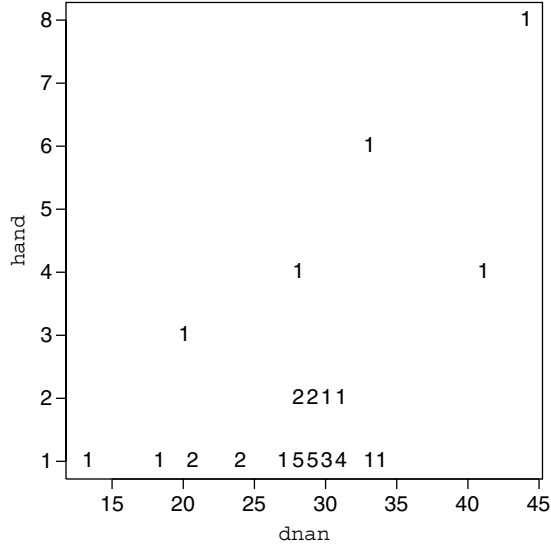


Figure 1 Scatter plot of handedness data. The numbers show the multiplicities of the observations

which may be estimated by

$$b = \bar{\theta}^* - \hat{\theta}, \quad v = \frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}_r^* - \bar{\theta}^*)^2, \quad (1)$$

where $\bar{\theta}^* = R^{-1} \sum_{r=1}^R \hat{\theta}_r^*$ is the average of the simulated $\hat{\theta}^*$ s. For the handedness data, we obtain $b = -0.046$ and $v = 0.043$ using the 10 000 simulations shown in (b) of Figure 2.

The simulation variability of quantities such as b and v , which vary depending on the random resamples, is reduced by increasing R . As a general rule, $R \geq 100$ is needed for bias and variance estimators to

be reasonably stable. For the more challenging tasks of constructing confidence intervals or significance tests discussed below, $R \geq 1000$ is needed for 95% confidence intervals, and more resamples are needed for higher confidence levels.

The power of the bootstrap should be evident from the description above: although the correlation coefficient is simple, it could be replaced by a statistic $\hat{\theta}$ of a complexity limited only by the data analyst's imagination and the computing power available – see, for example, the article on **finite mixture distributions**. The bootstrap is not a universal panacea, however, and many of the procedures described below apply only when $\hat{\theta}$ is sufficiently smooth as a function of the data values. The resampling scheme may need careful modification for reliable results.

The discussion above has shown how resampling may be used to mimic sampling variability, but not how the resamples can be used to provide inferences on the underlying population quantities. We discuss this below.

Confidence Intervals

Many estimators $\hat{\theta}$ are normally distributed, at least in large samples. If so, then an approximate equitailed $100(1 - 2\alpha)\%$ confidence interval for the estimand θ is $\hat{\theta} - b \pm z_\alpha \sqrt{v}$, where b and v are given at (1) and z_α is the α **quantile** of the standard normal, $N(0, 1)$, distribution. For the handedness data this gives a 95% interval of (0.147, 0.963) for the correlation θ . The quality of the normal approximation and hence the reliability of the confidence interval can be assessed by graphical comparison of $\hat{\theta}_1^*, \dots, \hat{\theta}_R^*$ with a normal

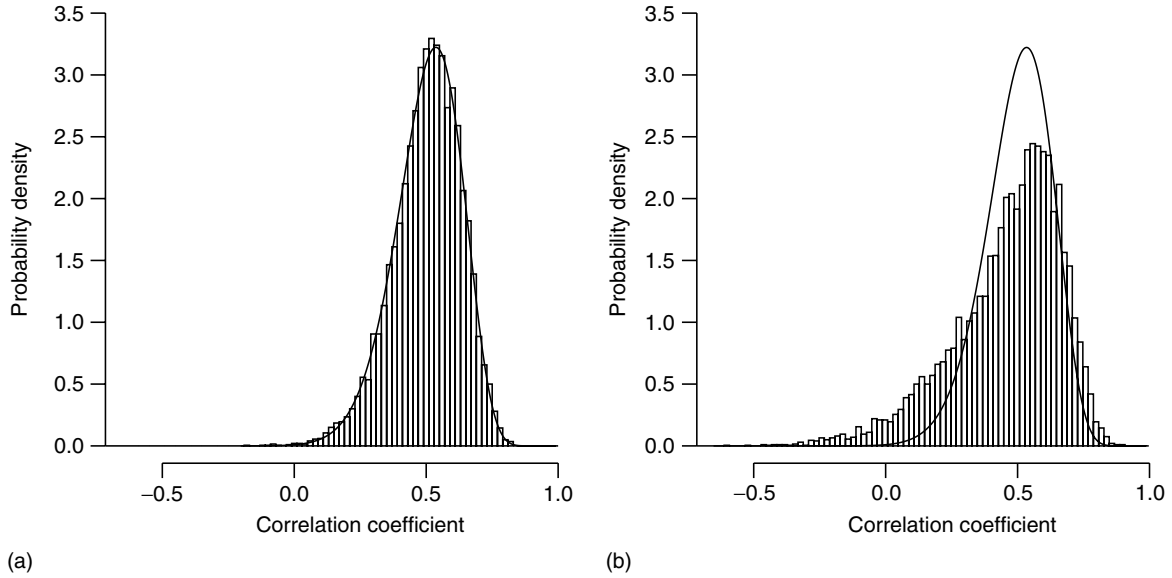


Figure 2 Histograms of simulated correlation coefficients for handedness data. (a): Simulation from fitted bivariate-normal distribution. (b): Simulation from the data by bootstrap resampling. The line in each figure shows the theoretical probability-density function of the correlation coefficient under sampling from the fitted bivariate-normal distribution

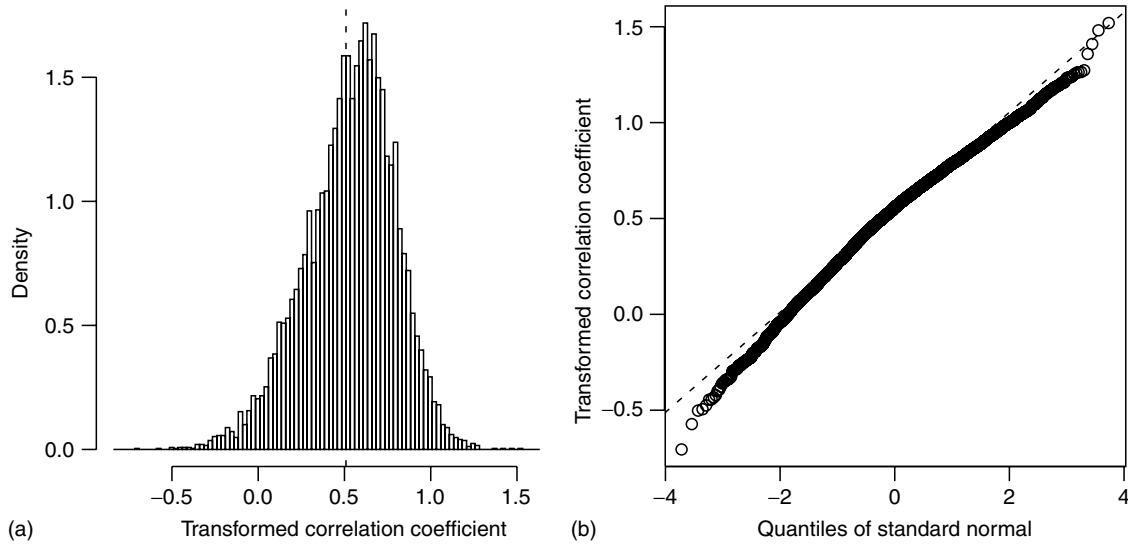


Figure 3 Bootstrap values $\hat{\zeta}^*$ of transformed correlation coefficient ζ . (a): histogram, with vertical dashed line showing original value $\hat{\zeta}$. (b): normal probability plot, with dashed line indicating exact normality

density. The strongly skewed histogram in (b) of Figure 2 suggests that this confidence interval will be quite unreliable for the handedness correlation.

Nonnormality can often be remedied by transformation. In the case of the correlation coefficient,

a well-established possibility is to consider $\zeta = \frac{1}{2} \log\{(1 + \theta)/(1 - \theta)\}$, which takes values in the entire real line rather than just in the interval $(-1, 1)$. Figure 3 shows a histogram and normal probability plot of the values of $\hat{\zeta}^* = \frac{1}{2} \log\{(1 + \hat{\theta}^*)/(1 - \hat{\theta}^*)\}$,

for which the normal distribution is a better, though a not perfect, fit. The 95% confidence interval for ζ computed using values of b and v obtained from the $\hat{\zeta}^*$ s is (0.074, 1.110), and transformation of this back to the original scale gives a 95% confidence interval for θ of (0.074, 0.804), shorter than and shifted to the left relative to the interval obtained by treating the $\hat{\theta}^*$ themselves as normally distributed.

Although simple, normal confidence intervals often require a transformation to be determined by the data analyst, and hence, more readily automated approaches have been extensively studied. The most natural way to use the bootstrap replicates $\hat{\theta}_1^*, \dots, \hat{\theta}_R^*$ of $\hat{\theta}$ to obtain a confidence interval for θ is to use their quantiles directly. Let $\hat{\theta}_{(1)}^* \leq \dots \leq \hat{\theta}_{(R)}^*$ be the ordered bootstrap replicates. Then, one simple approach to constructing an equitailed $100(1 - 2\alpha)\%$ confidence interval is to take the α and $(1 - \alpha)$ quantiles of the $\hat{\theta}^*$, that is, $\hat{\theta}_{(R\alpha)}^*$ and $\hat{\theta}_{(R(1-\alpha))}^*$, where, if necessary, the numbers $R\alpha$ and $R(1 - \alpha)$ are rounded to the nearest integers. Thus, for example, if a 95% confidence interval is required, we set $\alpha = 0.025$ and, with $R = 10\,000$, would take its limits to be $\hat{\theta}_{(250)}^*$ and $\hat{\theta}_{(9750)}^*$. In general, the corresponding interval may be expressed as $(\hat{\theta}_{(R\alpha)}^*, \hat{\theta}_{(R(1-\alpha))}^*)$, which is known as the *bootstrap percentile confidence interval*. This has the useful property of being invariant to the scale on which it is calculated, meaning that the same interval is produced using the $\hat{\theta}^*$ s directly as would be obtained by transforming them, computing an interval for the transformed parameter, and then back-transforming this to the θ scale. Its simplicity and transformation-invariance have led to widespread use of the percentile interval, but it has drawbacks. Typically such intervals are too short, so the probability that they contain the parameter is lower than the nominal value: an interval with nominal level 95% may in fact contain the parameter with probability only .9 or lower. Moreover, such intervals tend to be centered incorrectly: even for equitailed intervals, the probabilities that θ falls below the lower end point and above the upper end point are unequal, and neither is equal to α . For the handedness data, this method gives a 95% confidence interval of $(-0.047, 0.758)$. This seems to be shifted too far to the left relative to the transformed normal interval.

These deficiencies have led to intensive efforts to develop more reliable bootstrap confidence intervals. One variant of the percentile interval, known as the

bias-corrected and accelerated (BC_a) or *adjusted percentile interval*, may be written as $(\hat{\theta}_{(R\alpha')}^*, \hat{\theta}_{(R(1-\alpha''))}^*)$, where α' and α'' are estimated from the $\hat{\theta}_1^*, \dots, \hat{\theta}_R^*$ in such a way that the resulting intervals are closer to equitailed with the required coverage probability $1 - 2\alpha$. Formulae for α' and α'' are given in §5.3 of [2], but often they are built into software libraries for bootstrapping. For the handedness data, we find that $\alpha' = 0.0485$ and $\alpha'' = 0.0085$ resulting in the 95% interval (0.053, 0.792). This method seems to have corrected for the shift to the left we saw in the percentile interval.

Other methods for the calculation of confidence intervals rely on an analogy with the Student t statistic used with normally distributed samples. Suppose that a standard error s for $\hat{\theta}$ is available; then s^2 is an estimate of the variance of $\hat{\theta}$. Then, the basis of more general confidence-interval procedures is the use of bootstrap simulation to estimate the distribution of $z = (\hat{\theta} - \theta)/s$. *Studentized bootstrap confidence intervals* are constructed by using bootstrap simulation to generate R bootstrap replicates $z^* = (\hat{\theta}^* - \hat{\theta})/s^*$, where s^* is the standard error computed using the bootstrap sample that gave $\hat{\theta}^*$. The resulting $z_1^*, \dots, z_R^*$ are then ordered, their α and $(1 - \alpha)$ quantiles $z_{(R\alpha)}^*$ and $z_{(R(1-\alpha))}^*$ obtained, and the resulting $(1 - 2\alpha) \times 100\%$ confidence interval has limits $(\hat{\theta} - sz_{(R(1-\alpha))}^*, \hat{\theta} - sz_{(R\alpha)}^*)$. These intervals often behave well in practice but require a standard error s , which must be calculated for the original sample and each bootstrap sample. If a standard error is unavailable, then the Studentized interval may be simplified to the *bootstrap basic confidence interval* $(2\hat{\theta} - \hat{\theta}_{(R(1-\alpha))}^*, 2\hat{\theta} - \hat{\theta}_{(R\alpha)}^*)$. Either of these intervals can also be used with transformation, but unlike the percentile intervals, they are not transformation-invariant. It is generally advisable to use a transformation that maps the parameter range to the whole real line to avoid getting values that lie outside the allowable range. For the handedness data, the 95% Studentized and basic intervals using the same transformation as before are $(-0.277, 0.868)$ and $(0.131, 0.824)$ respectively. The Studentized interval seems too wide in this case and the basic interval too short. Without transformation, the upper end points of both intervals were greater than 1.

Standard error formulae can be found for many everyday statistics, including the correlation coefficient. If a formula is unavailable, the bootstrap itself can sometimes be used to find a standard error, using

two nested levels of bootstrapping. The bootstrap is applied Q times to each first-level bootstrap sample $y_1^*, \dots, y_n^*$, yielding second-level samples and corresponding replicates $\hat{\theta}_1^{**}, \dots, \hat{\theta}_Q^{**}$. Then, $s^* = \sqrt{v^*}$, where v^* is obtained by applying the variance formula at (1) to $\hat{\theta}_1^{**}, \dots, \hat{\theta}_Q^{**}$. The standard error of the original $\hat{\theta}$ is computed as $s = \sqrt{v}$, with v computed by applying (1) to the first-level replicates $\hat{\theta}_1^*, \dots, \hat{\theta}_R^*$. Although the number R of first-level replicates should be at least 1000, it will often be adequate to take the number Q of second-order replicates of order 100: thus, around 100 000 bootstrap samples are needed in all. With today's fast computing, this can be quite feasible.

Chapter 5 of [2] gives fuller details of these bootstrap confidence-interval procedures, and describes other approaches.

Hypothesis Tests

Often a sample is used to test a null hypothesis about the population from which the sample was drawn – for example, we may want to test whether a correlation is zero, or if some mean response differs between groups of subjects. A standard approach is then to choose a test statistic T in such a way that large values of T give evidence against the null hypothesis, and to compute its value t_{obs} for the observed data. The strength of evidence against the hypothesis is given by the significance probability or P value p_{obs} , the probability of observing a value of T as large as or larger than t_{obs} if the hypothesis is true. That is, $p_{\text{obs}} = P_0(T \geq t_{\text{obs}})$, where P_0 represents a probability computed as if the null hypothesis were true. Small values of p_{obs} are regarded as evidence against the null hypothesis.

A significance probability involves the computation of the distribution of the test statistic under the null hypothesis. If this distribution cannot be obtained theoretically, then the bootstrap may be useful. A key step is to obtain an estimator $\hat{F}_0$ of the population distribution under the null hypothesis, F_0 . The bootstrap samples are then generated by sampling from $\hat{F}_0$, yielding R bootstrap replicates $T_1^*, \dots, T_R^*$ of T . The significance probability is estimated by

$$\hat{p}_{\text{obs}} = \frac{\{\text{number of } T^* \geq t_{\text{obs}}\} + 1}{R + 1},$$

where the +1s appear because t_{obs} is also a replicate under the null hypothesis, and $t_{\text{obs}} \geq t_{\text{obs}}$.

For bootstrap hypothesis testing to work, it is essential that the fitted distribution $\hat{F}_0$ satisfy the null hypothesis. This is rarely true of the empirical distribution function $\hat{F}$, which therefore cannot be used in the usual way. The construction of an appropriate $\hat{F}_0$ can entail restriction of $\hat{F}$ or the testing of a slightly different hypothesis. For example, the hypothesis of no correlation between `hand` and `dnan` could be tested by taking as test statistic $T = \hat{\theta}$, the sample correlation coefficient, but imposing this hypothesis would involve reweighting the points in Figure 1 in such a way that the correlation coefficient computed using the reweighted data would equal zero, followed by simulation of samples $y_1^*, \dots, y_n^*$ from the reweighted distribution. This would be complicated, and it is easier to test the stronger hypothesis of independence, under which any association between `hand` and `dnan` arises purely by chance. If so, then samples may be generated under the null hypothesis by independent bootstrap resampling separately from the univariate empirical distributions for `hand` and for `dnan`; see (a) of Figure 4, which shows that pair (`hand*`, `dnan*`) that were not observed in the original data may arise when sampling under the null hypothesis. Comparison of (b) of Figure 2 and (b) of Figure 4 shows that the distributions of correlation coefficients generated using the usual bootstrap and under the independence hypothesis are quite different.

The handedness data have correlation coefficient $t_{\text{obs}} = 0.509$, and 18 out of 9999 bootstrap samples generated under the hypothesis of independence gave values of T^* exceeding t_{obs} . Thus $\hat{p}_{\text{obs}} = 0.0019$, strong evidence of a positive relationship. A two-sided test could be performed by taking $T = |\hat{\theta}|$, yielding significance probability of about 0.004 for the null hypothesis that there is neither positive nor negative association between `hand` and `dnan`.

A parametric test that assumes that the underlying distribution is bivariate normal gives an appreciably smaller one-sided significance probability of .0007, but the inadequacy of the normality assumption implies that this test is less reliable than the bootstrap test.

Another simulation-based procedure often used to test independence is the permutation method (*see Permutation Based Inference*), under which the values of one of the variables are randomly permuted. The resulting significance probability is typically very close to that from the bootstrap test, because the

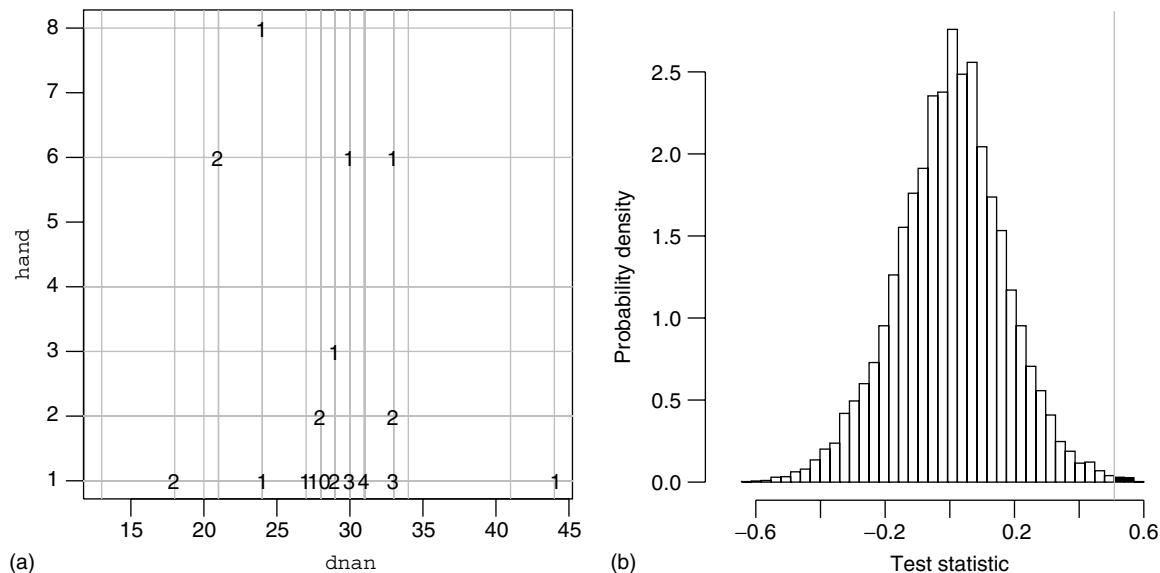


Figure 4 Bootstrap hypothesis test of independence of hand and dnan. (a) intersections of the grey lines indicate possible pairs (hand^* , dnan^*) when resampling values of hand and dnan independently with replacement. The numbers show the multiplicities of the sampled pairs for a particular bootstrap resample. (b) histogram of correlation coefficients generated under null hypothesis of independence. The vertical line shows the value of the correlation coefficient for the dataset, and the shaded part of the histogram corresponds to the significance probability

only difference between the resampling schemes is that permutation samples without replacement and the bootstrap samples with replacement. For the handedness data, the permutation test yields one-sided and two-sided significance probabilities of .002 and .003 respectively.

Owing to the difficulties involved in constructing a resampling distribution that satisfies the null hypothesis, it is common in practice to use confidence intervals to test hypotheses. A two-sided test of the null hypothesis that $\theta = \theta_0$ may be obtained by bootstrap resampling from the usual empirical distribution function $\hat{F}$ to compute a $100(1 - \alpha)\%$ confidence interval. If this does not contain θ_0 , we conclude that the significance probability is less than α . For one-sided tests, we use one-sided confidence intervals. This approach is most reliable when used with an approximately pivotal statistic but this can be hard to verify in practice. For the handedness data, the value $\theta_0 = 0$ lies outside the 95% BCa confidence interval, but within the 99% confidence interval, so we would conclude that the significance probability for a two-sided test is between .01 and .05. This is appreciably larger than found above, but still gives evidence of

a relationship between the two variables. Using the transformed Studentized interval, however, we get a P value of greater than .10, which contradicts this conclusion. In general, inverting a confidence interval in this way seems to be unreliable and is not advised.

Chapter 4 of [2] contains a more complete discussion of bootstrap hypothesis testing.

More Complex Models

The discussion above considers only the simple situation of random sampling, but many applications involve more complex statistical models, such as the linear regression model (*see Multiple Linear Regression*). Standard assumptions for this are that the mean response variable is a linear function of explanatory variables, and that deviations from this linear function have a normal distribution. Here the bootstrap can be applied to overcome potential nonnormality of the deviations, by using the data to estimate their distribution. The deviations are unobserved because the true line is unknown, but they can be estimated using residuals, which can then

be resampled. If $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ are estimates of the linear model coefficients and $e_1^*, \dots, e_n^*$ is a bootstrap sample from the residuals $e_1, \dots, e_n$, then bootstrap responses y^* can be constructed as

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki} + e_i^*, \quad i = 1, \dots, n.$$

The values of the explanatory variables in the bootstrap sample are the same as for the original sample, but the response variables vary. Since the matrix of explanatory variables remains the same, this method is particularly appropriate in designed experiments where the explanatory variables are set by the experimenter. It does however presuppose the validity of the linear model from which the coefficients and residuals are estimated.

An alternative procedure corresponding to our treatment of the data in Table 1 is to resample the vector observations $(y_i, x_{1i}, \dots, x_{ki})$. This may be appropriate when the explanatory variables are not fixed by the experimenter but may be treated as random. One potential drawback is that the design matrix of the resamples may be singular, or nearly so, and if so there will be difficulties in fitting the linear model to the bootstrap sample. These procedures generalize to the **analysis of variance**, the **generalized linear model**, and other regression models.

Extensions of the bootstrap to **survival analysis** and time series (see **Time Series Analysis**) have also been studied in the literature. The major difficulty in complex models lies in finding a resampling scheme that appropriately mimics how the data arose. A detailed discussion can be found in Chapters 3 and 6–8 of [2].

Computer Resources

Although the bootstrap is a general computer intensive method for nonparametric inference, it does not appear in all statistical software packages. Code for bootstrapping can be written in most packages, but this does require programming skills. The most comprehensive suite of code for bootstrapping is the `boot` library written by A. J. Canty for S-Plus, which can be downloaded from <http://statwww.epfl.ch/davison/BMA/library.html>. This code has also been made available as a package for R by B. D. Ripley and can be downloaded free from <http://cran.r-project.org> as part of the binary releases of

the R package. Another free package that can handle a limited range of statistics is David Howell's Resampling Statistics available from <http://www.uvm.edu/dhowell/StatPages/Resampling/Resampling.html>. S-Plus also has many features for the bootstrap and related methods. Some are part of the base software but most require the use of the S + Resample library, which can be downloaded from <http://www.insightful.com/downloads/libraries/>. The commercial package Resampling Stats is available as a stand-alone program or as an add-in for Excel or Matlab from <http://www.resample.com/>. Statistical analysis packages that include some bootstrap functionality are Systat, Stata, and SimStat. There are generally limits to the types of statistics that can be resampled in these packages but they may be useful for many common statistics (see **Software for Statistical Analyses**).

Literature

Thousands of papers and several books about the bootstrap have been published since it was formulated by Efron [3]. Useful books for the practitioner include [7], [1], and [6]. References [2] and [4] describe the ideas underlying the methods, with many further examples, while [5] and [8] contain more theoretical discussion. Any of these contains many further references, though [1] has a particularly full bibliography. The May 2003 issue of *Statistical Science* contains recent surveys of aspects of the research literature.

Acknowledgment

The work was supported by the Swiss National Science Foundation and by the National Sciences and Engineering Research Council of Canada.

References

- [1] Chernick, M.R. (1999). *Bootstrap Methods: A Practitioner's Guide*, Wiley, New York.
- [2] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge.
- [3] Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, 1–26.
- [4] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

8 Bootstrap Inference

- [5] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer, New York.
- [6] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Grove.
- [7] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methodology in Biology*, 2nd Edition, Chapman & Hall, London.
- [8] Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*, Springer, New York.

A.J. CANTY AND ANTHONY C. DAVISON

Box Plots

SANDY LOVIE

Volume 1, pp. 176–178

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Box Plots

Of all the graphical novelties unleashed on a disbelieving world by the late **John Tukey** [3] under the banner of **exploratory data analysis (EDA)**, only **stem and leaf plots** and box plots seem to have stood the test of time. Box plots, or box and whisker plots, follows the spirit of EDA by using only robust components in their construction, that is, ones which minimize the effects of **outliers** or crude data simplifying schemes (see [2] and [4] for more details). Thus, the length of the box is the midspread (**interquartile range**), while the measure of location marked within the box is the **median**. The most common rule for determining the lengths of the whiskers emanating from the top and bottom of the box uses the midspread – specifically, the whiskers are extended to the largest and smallest values (called ‘adjacent’ values) lying within upper and lower limits (fences) defined by 1.5 times the midspread above and below the box. Most box plots also display outliers that are also defined by the midspread in that they are values in the sample or batch that lie above or below these upper and lower limits. Alternative rules extend the whiskers to the largest and smallest values in the sample, regardless of whether these might be outliers, in order to reveal additional distributional properties of the sample, for example, whether it is symmetric or skewed, and if so in which direction? The labeled display in Figure 1 below, using data from Minitab’s *Pulse1* data set, illustrates the basic box plot components and terminology.

Early versions of box and whisker plots did not use the upper and lower **quartiles** to define the box

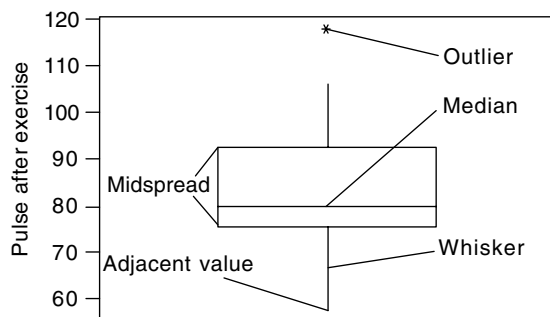


Figure 1 One sample box plot with the major parts labeled

length, but used closely related measures, termed the upper and lower ‘hinges’ by Tukey [3]. These, like the median, represent natural folding points in an ordered sample. Other refinements included fences, which add structure to the areas beyond the upper and lower limits defined earlier, and notches with different angles cut into the sides of the box to represent the size of the sample’s 95% confidence interval about the median.

Tukey also recommended employing the *width* of the box plot to represent the sample size where one wants to compare samples of differing size. Unfortunately, the latter advice can lead to the analyst coming under the malign influence of the so-called *Wundt illusion* when eyeballing plots where the width and length of the box can vary simultaneously and independently. However, while few of these refinements have been incorporated into the major statistics packages, other novelties are on offer. For example, in Minitab, a **Confidence Interval** box can be substituted for the midspread box, and adjacent box plots linked by their medians to facilitate comparisons between them. Thus, box plots can yield a surprising amount of information about differences in the location, spread, and shape of the samples, and about the presence of outliers. Figure 2 is an example of a side by side display for comparing two samples. The data used for the plot is from Minitab’s *Pulse* dataset.

With multiple box plots, one can also compare the distributions and outlier proneness of the samples: this, in turn, provides the basis for a conservative informal graphical inference in that plots whose

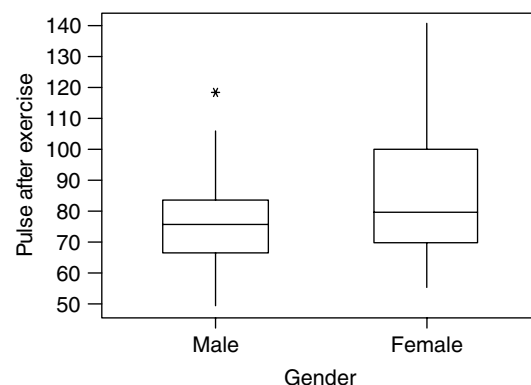


Figure 2 Two sample box plots for comparing pulse after exercise for males and females

2 Box Plots

boxes do not overlap contain medians that are also different. A more precise version of the latter test is available with notched box plots, where the probability of rejecting the null hypothesis of true equal medians is 0.05 when the 95% notches just clear each other, provided that it can be assumed that the samples forming the comparison are roughly normally distributed, with approximately equal spreads (see [1] for more details).

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.
- [2] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1983). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.
- [3] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.
- [4] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

SANDY LOVIE

Bradley–Terry Model

JOHN I. MARDEN

Volume 1, pp. 178–184

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bradley–Terry Model

A paired comparison experiment involves several ‘judges’ comparing several ‘objects’. Each judge performs a number of paired comparisons, that is, considers a pair of the objects, and indicates which of the two is preferred. Bradley and Terry [3] present an example in which the objects are three types of pork roast, depending on the ingredients in the feed of the hogs (just corn, corn plus some peanut, corn plus much peanut). For each comparison, a judge was given two samples of pork roast from different feeds, and the judge indicated the preferred one. Each judge performed several paired comparisons. In a sports context, the objects would be the players (e.g., in tennis) or the teams (e.g., in baseball). There are no judges as such, but a comparison is a match or game, and the determination of which in a pair is preferred is based on who wins the match.

Models for such data would generally be based on the probabilities that Object i is preferred to Object j for each i, j . A fairly simple model would have the chance Object i is preferred to Object j be based on just the relative overall strength or popularity of the two objects. That is, there would be no particular interaction between objects. The notion of no interaction would imply that if Team A is likely to beat Team B, and Team B is likely to beat Team C, then Team A is even more likely to beat Team C. The Bradley–Terry model, sometimes called the Bradley–Terry–Luce model, is one popular method for explicitly exhibiting this lack of interaction. Each object i is assumed to have a positive measure of strength v_i , where the chance that Object i is preferred is

$$p_{ij} = P[\text{Object } i \text{ is preferred to Object } j] = \frac{v_i}{v_i + v_j}. \quad (1)$$

The model is thought to be first proposed by Zermelo [16], but Bradley and Terry [3] brought it into wide popularity among statistical practitioners. See [7] for an extensive bibliography, and the book [5], which covers other models as well.

It is often more convenient to work with odds than with probabilities (see **Odds and Odds Ratios**). If p

is the probability of an event, then the odds are

$$\text{Odds} = \frac{p}{1-p}, \quad \text{hence} \quad p = \frac{\text{Odds}}{1 + \text{Odds}}. \quad (2)$$

The Bradley–Terry probability in (1) then translates to

$$\begin{aligned} \text{Odds}_{ij} &= \text{Odds}[\text{Object } i \text{ is preferred to Object } j] \\ &= \frac{v_i}{v_j}. \end{aligned} \quad (3)$$

An approach to motivating this model follows. Imagine two tennis players, named A and B. The chance Player A has of winning a match against a typical opponent is $p_A = 0.80$, and the chance Player B has of winning against a typical opponent is $p_B = 0.65$. Without any other information, what would be a reasonable guess of the chance that Player A would beat Player B? It would be greater than 50%, because Player A is a better player than B, and would be less than 80%, because Player B is better than the typical player. But can we be more informative?

One approach is to set up a competition in which Player A has a coin with chance of heads being p_A , and Player B has a coin with chance of heads being p_B . The two players flip their coins simultaneously. If one is heads and one is tails, the player with the heads wins. If both are heads or both are tails, they flip again. They continue until a pair of flips determines a winner. The chance Player A wins this competition is the chance A flips heads given that there is a decision, that is, given that one of A and B flips heads and the other flips tails. Thus

$$p_{AB} = P[A \text{ beats } B] = \frac{p_A(1-p_B)}{p_A(1-p_B) + (1-p_A)p_B}. \quad (4)$$

With $p_A = 0.80$ and $p_B = 0.65$, we have $P[A \text{ beats } B] = 0.28/(0.28 + 0.13) = 0.6829$. That is, we guess that the chance A beats B is about 68%, which is at least plausible. The relationship between $P[A \text{ beats } B]$ and p_A, p_B in (4) is somewhat complicated. The formula can be simplified by looking at odds instead of probability. From (4),

$$\begin{aligned} \text{Odds}_{AB} &= \text{Odds}[A \text{ beats } B] \\ &= \frac{p_A(1-p_B)/(p_A(1-p_B) + (1-p_A)p_B)}{1 - p_A(1-p_B)/(p_A(1-p_B) + (1-p_A)p_B)} \end{aligned}$$

2 Bradley–Terry Model

$$\begin{aligned} &= \frac{p_A}{1 - p_A} \frac{1 - p_B}{p_B} \\ &= \frac{Odds_A}{Odds_B}, \end{aligned} \quad (5)$$

where $Odds_A$ is the odds player A beats a typical player. These odds are the Bradley–Terry odds (3), with v_A being identified with $Odds_A$. That is, the Bradley–Terry parameter v_i can be thought of as the odds Object i is preferred to a typical object.

Sections titled Luce’s Choice Axiom and Thurstone’s Scaling discuss additional motivations based on Luce’s Choice Axiom and Thurstonian choice models, respectively. The next section exhibits the use of the Bradley–Terry model in examples. The section titled Ties deals briefly with ties, and the section titled Calculating the Estimates has some remarks concerning calculation of estimates.

Modeling Paired Comparison Experiments

The basic Bradley–Terry model assumes that there are L objects, labeled $1, \dots, L$, Object i having associated parameter $v_i > 0$, and n independent comparisons between pairs of objects are performed. The chance that Object i is preferred to Object j when i and j are compared is given in (1). The data can be summarized by the counts n_{ij} , the number of paired comparisons in which i is preferred to j . The likelihood function (see **Maximum Likelihood Estimation**) is then

$$\begin{aligned} L(v_1, \dots, v_L; \{n_{ij} | i \neq j\}) &= \prod_{i \neq j} p_{ij}^{n_{ij}} \\ &= \prod_{i \neq j} \left(\frac{v_i}{v_i + v_j} \right)^{n_{ij}}. \end{aligned} \quad (6)$$

In the example comparing pork roasts in [3], let Object 1 be the feed with just corn, Object 2 the feed with corn and some peanut, and Object 3 the feed with corn and much peanut. Then the results from Judge I, who made five comparisons of each pair, are (from Table 4 in their paper)

Corn preferred to (Corn + Some peanut)	$n_{12} = 0$
(Corn + Some peanut) preferred to Corn	$n_{21} = 5$
Corn preferred to Corn + Much peanut	$n_{13} = 1$
(Corn + Much peanut) preferred to Corn	$n_{31} = 4$
(Corn + Some peanut) preferred to (Corn + Much peanut)	$n_{23} = 2$
(Corn + Much peanut) preferred to (Corn + Some peanut)	$n_{32} = 3$

(7)

The likelihood (6) in this case is

$$\begin{aligned} &\left(\frac{v_1}{v_1 + v_2} \right)^0 \left(\frac{v_2}{v_1 + v_2} \right)^5 \left(\frac{v_1}{v_1 + v_3} \right)^1 \\ &\times \left(\frac{v_3}{v_1 + v_3} \right)^4 \left(\frac{v_2}{v_2 + v_3} \right)^2 \left(\frac{v_3}{v_2 + v_3} \right)^3 \\ &= \frac{v_1 v_2^7 v_3^7}{(v_1 + v_2)^5 (v_1 + v_3)^5 (v_2 + v_3)^5}. \end{aligned} \quad (8)$$

Note that the numerator in (8) has parameter v_i raised to the total number of times Object i is preferred in any of the paired comparisons, and the denominator has for each pair i, j of objects the sum of their parameters $(v_i + v_j)$ raised to the total number of times those two objects are compared. That is, in general, the likelihood in (6) can be written

$$L(v_1, \dots, v_L; \{n_{ij} | i \neq j\}) = \frac{\prod_{i=1}^n v_i^{n_i}}{\prod_{i < j} (v_i + v_j)^{N_{ij}}}, \quad (9)$$

where

$$n_i = \sum_{j \neq i} n_{ij} \quad \text{and} \quad N_{ij} = n_{ij} + n_{ji}. \quad (10)$$

In fact, $(n_1, \dots, n_L)$ is a sufficient statistic for this model.

The parameters can be estimated using maximum likelihood (see section titled Calculating the Estimates for some computational approaches), but notice that these probabilities do not change if every v_i is multiplied by the same positive constant c , which means that the v_i ’s are not uniquely determined by the p_{ij} ’s. One usually places a constraint on the v_i ’s, for example, that $v_L = 1$, or that $v_1 + \dots + v_L = 1$.

I prefer scaling the v_i 's so that

$$\frac{1}{L} \sum_{i=1}^L \frac{v_i}{1+v_i} = \frac{1}{2}, \quad (11)$$

which means that v_i is the odds that Object i is preferred to a 'typical' ideal object, where the typical object has on an average a 50% chance of being preferred over the others.

Maximizing (8) subject to the constraint in (11) yields the estimates

$$\hat{v}_1 = 0.2166, \hat{v}_2 = \hat{v}_3 = 1.9496. \quad (12)$$

This judge likes some or much peanut equally, and any peanut substantially better than no peanut.

Bradley and Terry also present the results from another judge, Judge II, who again made five comparisons of each pair. Let m_{ij} be the number of times Judge II preferred Object i to Object j , and τ_1, τ_2, τ_3 be the corresponding Bradley–Terry parameters. One can imagine at least two models for the combined data of the two judges: (1) The judges have the same preferences, so that $v_1 = \tau_1, v_2 = \tau_2, v_3 = \tau_3$; (2) The judges may have different preferences, so that v_i does not necessarily equal τ_i . The likelihoods for the two models are

$$L_{\text{Same}} = \prod_{i \neq j} \left(\frac{v_i}{v_i + v_j} \right)^{n_{ij} + m_{ij}} \quad (13)$$

and

$$L_{\text{Different}} = \prod_{i \neq j} \left(\frac{v_i}{v_i + v_j} \right)^{n_{ij}} \left(\frac{\tau_i}{\tau_i + \tau_j} \right)^{m_{ij}}, \quad (14)$$

respectively.

The data for Judge II are

$$\begin{aligned} m_{12} &= 3, m_{21} = 2, m_{13} = 4, \\ m_{31} &= 1, m_{23} = 3, m_{32} = 2. \end{aligned} \quad (15)$$

In the second model (14), the estimates of the v_i 's are those in (12), and the estimate of the τ_i 's are

$$\hat{\tau}_1 = 1.7788, \hat{\tau}_2 = 1.0000, \hat{\tau}_3 = 0.5622. \quad (16)$$

On the basis of the estimates, Judge II appears to have distinctly different preferences than Judge I; Judge II prefers less peanut over more.

The question arises whether the apparent difference between the judges is statistically significant, which we address by testing the hypotheses

$$\begin{aligned} H_0: (v_1, v_2, v_3) &= (\tau_1, \tau_2, \tau_3) \quad \text{versus} \\ H_A: (v_1, v_2, v_3) &\neq (\tau_1, \tau_2, \tau_3). \end{aligned} \quad (17)$$

The likelihood ratio test (*see* **Maximum Likelihood Estimation**) uses the statistic

$$W = 2[\log(L_{\text{Different}}) - \log(L_{\text{Same}})], \quad (18)$$

where in $L_{\text{Different}}$ we replace the parameters with their estimates from (12) and (16), and in L_{Same} with the maximum likelihood estimates from (13), which are

$$(\hat{v}_1, \hat{v}_2, \hat{v}_3) = (\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3) = (0.7622, 1.3120, 1.0000). \quad (19)$$

Under the null hypothesis that the two judges have the same preference structure, W will be approximately χ^2_2 . The degrees of freedom here are the difference between the number of free parameters in the alternative hypothesis, which is 4 because there are six parameters but two constraints, and the null hypothesis, which is 2. In general, if there are L objects and M judges, then testing whether all M judges have the same preferences uses $(L-1) \times (M-1)$ degrees of freedom.

For these data, $W = 8.5002$, which yields an approximate P value of 0.014. Thus, it appears that indeed the two judges have different preferences.

Example: Home and Away Games

Agresti ([1], pages 437–438) applies the Bradley–Terry model to the 1987 performance of the seven baseball teams in the American League Eastern Division. Each pair of teams played 13 games. In the simplest Bradley–Terry model, each team has its v_i . The estimates are given in the first column of Table 1.

Each game is played in the home ballpark of one of the teams, and there is often an advantage to playing at home. A more complex model supposes that each team has two parameters, one for when it is playing at home and one for when it is playing away. Thus there are 14 parameters, v_i^{Home} and v_i^{Away} for each $i = 1, \dots, L = 7$, with one constraint as

4 Bradley–Terry Model

Table 1 Estimated odds for the baseball teams

	Simple	Home	Away	Effect	Neutral	Home	Away
Milwaukee	1.6824	2.1531	1.3367	1.6107	1.6970	1.9739	1.4590
Detroit	1.4554	1.8827	1.1713	1.6074	1.4691	1.7088	1.2630
Toronto	1.2628	1.0419	1.5213	0.6849	1.2667	1.4734	1.0890
New York	1.2050	1.3420	1.0997	1.2203	1.2100	1.4074	1.0403
Boston	1.0477	1.8104	0.6108	2.9639	1.0545	1.2266	0.9066
Cleveland	0.6857	0.8066	0.5714	1.4115	0.6798	0.7907	0.5844
Baltimore	0.3461	0.3145	0.3688	0.8526	0.3360	0.3908	0.2889

usual. In this case, not every pair is observed, that is, Milwaukee cannot be the home team and away team in the same game. That fact does not present a problem in fitting the model, though. The second and third columns in Table 1 give the estimated home and away odds for each team in this model. The ‘Effect’ column is the ratio of the home odds to the away odds, and estimates the effect of being at home versus away for each team. That is, it is the odds the home team would win in the imaginary case that the same team was the home and away team. For most teams, the odds of winning are better at home than away, especially for Boston, although for Toronto and Baltimore the reverse is true.

To test whether there is a home/away effect, one can perform the likelihood ratio test between the two models. The $W = 10.086$ on 7 degrees of freedom (7 because the simple model has $7 - 1 = 6$ free parameters, and the more complicated model has $14 - 1 = 13$). The approximate P value is 0.18, so it appears that the simpler model is not rejected.

Agresti considers a model between the two, where it is assumed that the home/away effect is the same for each team. The parameter v_i can be thought of as the odds of team i winning at a neutral site, and a new parameter θ is introduced that is related to the home/away effect, where

$$v_i^{\text{Home}} = \theta v_i \quad \text{and} \quad v_i^{\text{Away}} = \frac{v_i}{\theta}. \quad (20)$$

Then the odds that team i beats team j when the game is played at team i ’s home is

$$\text{Odds}(i \text{ at home beats } j) = \theta^2 \frac{v_i}{v_j}, \quad (21)$$

and the home/away effect is the same for each team, that is,

$$\text{Effect}(\text{team } i) = \frac{v_i^{\text{Home}}}{v_i^{\text{Away}}} = \frac{\theta v_i}{v_i / \theta} = \theta^2. \quad (22)$$

Table 1 contains the estimated v_i ’s in the ‘Neutral’ column. Notice that these values are very close to the estimates in the simple model. The estimate of θ is 1.1631, so that the effect for each team is $(1.1631)^2 = 1.3529$. The last two columns in the table give the estimated v_i^{Home} ’s and v_i^{Away} ’s. This model is a ‘smoothed’ version of the second model.

To test this model versus the simple model, we have $W = 5.41$. There is 1 degree of freedom here, because the only additional parameter is θ . The P value is then 0.02, which is reasonably statistically significant, suggesting that there is indeed a home/away effect. One can also test the last two models, which yields $W = 4.676$ on $13 - 7 = 6$ degrees of freedom, which is clearly not statistically significant. Thus, there is no evidence that the home/away effect differs among the teams.

Luce’s Choice Axiom

In the seminal book *Individual Choice Behavior* [10], Luce proposed an axiom to govern models for choosing one element among any given subset of elements. The Axiom is one approach to specifying ‘lack of interaction’ between objects when choosing from subsets of them. That is, the relative preference between two objects is independent of which other objects are among the choices. This idea is known as ‘independence from irrelevant alternatives.’

To be precise, let $\mathcal{O} = \{1, \dots, L\}$ be the complete set of objects under consideration and T be any subset of $\mathcal{O}$. Then, for Object $i \in T$, denote

$$P_T(i) = P[i \text{ is most preferred among } T]. \quad (23)$$

The special case of a paired comparison has T consisting of just two elements, for example,

$$P_{\{i,j\}}(i) = P[i \text{ is preferred to } j]. \quad (24)$$

Luce's Axiom uses the notation that for $S \subset T$,

$$\begin{aligned} P_T(S) &= P[\text{The most preferred object among } T \text{ is in } S] \\ &= \sum_{i \in S} P_T(i). \end{aligned} \quad (25)$$

The axiom follows.

Luce's Choice Axiom. For any $T \subset \mathcal{O}$,

$$(i) \quad \text{If } P_{\{a,b\}}(a) \neq 0 \text{ for all } a, b \in T, \text{ then for } i \in S \subset T, \quad P_T(i) = P_S(i)P_T(S); \quad (26)$$

$$(ii) \quad \text{If } P_{\{a,b\}}(a) = 0 \text{ for some } a, b \in T, \text{ then if } i \in T, i \neq a, \quad P_T(i) = P_{T-\{a\}}(i). \quad (27)$$

The essence of the Axiom can be best understood when $P_{\{a,b\}}(a) \neq 0$ for all $a, b \in \mathcal{O}$, that is, when in any paired comparison, there is a positive chance that either object is preferred. Then equation (26) is operational for all $i \in S \subset T$.

For an example, suppose $\mathcal{O} = \{\text{Coke, Pepsi, 7-up, Sprite}\}$. The axiom applied to $T = \mathcal{O}$, $S = \{\text{Coke, Pepsi}\}$ (the colas) and $i = \text{Coke}$ is

$$\begin{aligned} &P(\text{Coke is the favorite among all four}) \\ &= P(\text{Coke is preferred to Pepsi}) \\ &\quad \times P(\text{A cola is chosen as the favorite} \\ &\quad \text{among all four}). \end{aligned} \quad (28)$$

Thus, the choosing of the favorite can be decomposed into a two-stage process, where the choosing of Coke as the favorite starts by choosing colas over noncolas, then Coke as the favorite cola.

There are many implications of the Axiom. One is a precise formulation of the independence from irrelevant alternatives:

$$\frac{P_{\{i,j\}}(i)}{P_{\{i,j\}}(j)} = \frac{P_S(i)}{P_S(j)} \quad (29)$$

for any subset S that contains i and j . That is, the relative preference of i and j remains the same no

matter which, if any, other objects are available. In the soft drink example, this independence implies, in particular, that

$$\begin{aligned} \frac{P_{\{\text{Coke}, 7\text{-up}\}}(\text{Coke})}{P_{\{\text{Coke}, 7\text{-up}\}}(7\text{-up})} &= \frac{P_{\{\text{Coke}, 7\text{-up}, \text{Sprite}\}}(\text{Coke})}{P_{\{\text{Coke}, 7\text{-up}, \text{Sprite}\}}(7\text{-up})} \\ &= \frac{P_{\mathcal{O}}(\text{Coke})}{P_{\mathcal{O}}(7\text{-up})}. \end{aligned} \quad (30)$$

The main implication is given in the next Theorem, which is Theorem 3 in [10].

Luce's Theorem. Assume that the Choice Axiom holds, and that $P_{\{a,b\}}(a) \neq 0$ for all $a, b \in \mathcal{O}$. Then, there exists a positive finite number for each object, say v_i for Object i , such that for any $i \in S \subset \mathcal{O}$,

$$P_S(i) = \frac{v_i}{\sum_{j \in S} v_j}. \quad (31)$$

The interest in this paper is primarily paired comparison, and we can see that for paired comparisons, the Luce Choice Axiom with $S = \{i, j\}$ leads to the Bradley–Terry model (1), as long as the v_i 's are strictly between 0 and ∞ .

Thurstone's Scaling

Scaling models in preference data is addressed fully in [2], but here we will briefly connect the two. Thurstone [14] models general preference experiments by assuming that a given judge has a one-dimensional response to each object. For example, in comparing pork roasts, the response may be based on tenderness, or in the soft drinks, the response may be based on sweetness, or caffeine stimulation. Letting Z_a be the response to Object a , the probability of preferring i among those in subset T (which contains i) is given by

$$\begin{aligned} P_T(i) &= P[i \text{ is most preferred among } T] \\ &= P[Z_i > Z_j, \quad j \in T - \{i\}]. \end{aligned} \quad (32)$$

That is, the preferred object is that which engenders the largest response.

6 Bradley–Terry Model

Thurstone gives several models for the responses $(Z_1, \dots, Z_L)$ based on the Normal distribution. Daniels [4] looks at cases in which the Z 's are independent and from a location-family model with possibly different location parameters; [8] and [13] used gamma distributions. A question is whether any such Thurstonian model would satisfy Luce's Choice Axiom. The answer, given by Luce and Suppes in [11], who attribute the result to Holman and Marley, is the Gumbel distribution. That is, the Z_i 's are independent with density

$$f_{\mu_i}(z_i) = \exp(-(z_i - \mu_i)) \exp(-\exp(-(z_i - \mu_i))), \\ -\infty < z_i < \infty, \quad (33)$$

where the μ_i measures the typical strength of the response for Object i . Then the Thurstonian choice probabilities from (32) coincide with the model (31) with $v_i = \exp(\mu_i)$. Yellott [15] in fact shows that the Gumbel is the only distribution that will satisfy the Axiom when there are three or more objects.

Ties

It is possible that paired comparisons result in no preference, for example, in soccer, it is not unusual for a game to end in a tie, or people may not be able to express a preference between two colas. Particular parameterizations of the paired comparisons when extending the Bradley–Terry parameters to the case of ties are proposed in [12] and [6]. Rao and Kupper [12] add the parameter θ , and set

$$P[\text{Object } i \text{ is preferred to Object } j] = \frac{v_i}{v_i + \theta v_j}. \quad (34)$$

In this case, the chance of a tie when i and j are compared is

$$1 - \frac{v_i}{v_i + \theta v_j} - \frac{v_j}{v_j + \theta v_i} = \frac{\theta(v_i + v_j)}{(v_i + \theta v_j)(v_j + \theta v_i)}. \quad (35)$$

Davidson [6] adds positive c_{ij} 's so that the probability that Object i is preferred to Object j is

$v_i/(v_i + v_j + c_{ij})$. He suggests taking $c_{ij} = c\sqrt{v_i v_j}$ for some $c > 0$, so that

$$P[\text{Object } i \text{ is preferred to Object } j] \\ = \frac{v_i}{v_i + v_j + c\sqrt{v_i v_j}} \quad (36)$$

and

$$P[\text{Object } i \text{ and Object } j \text{ are tied}] \\ = \frac{c\sqrt{v_i v_j}}{v_i + v_j + c\sqrt{v_i v_j}}. \quad (37)$$

Davidson's suggestion may be slightly more pleasing than (34) because the v_i 's have the same meaning as before conditional on there being a preference, that is,

$$P[\text{Object } i \text{ is preferred to Object } j | \\ \text{Object } i \text{ and Object } j \text{ are not tied}] \\ = \frac{v_i}{v_i + v_j}. \quad (38)$$

Calculating the Estimates

In the Bradley–Terry model with likelihood as in (6), the expected number of times Object i is preferred equals

$$E[n_i] = E \left[\sum_{j \neq i} n_{ij} \right] = \sum_{j \neq i} N_{ij} \frac{v_i}{v_i + v_j}. \quad (39)$$

where n_i is the total number of times Object i is preferred, and N_{ij} is the number of times Objects i and j are compared, as in (10). The **maximum likelihood estimates** of the parameters are those that equate the n_i 's with their expected values, that is, they satisfy

$$n_i = \sum_{j \neq i} N_{ij} \frac{\hat{v}_i}{\hat{v}_i + \hat{v}_j}. \quad (40)$$

Rewriting (ignore the superscripts for the moment),

$$\hat{v}_i^{(k+1)} = \frac{n_i}{\sum_{j \neq i} N_{ij} / (\hat{v}_i^{(k)} + \hat{v}_j^{(k)})}, \quad (41)$$

This equation is the basis of an iterative method for finding the estimates. Starting with a guess $(\hat{v}_1^{(0)}, \dots, \hat{v}_L^{(0)})$ of the estimates, a sequence $(\hat{v}_1^{(k)}, \dots, \hat{v}_L^{(k)})$, $k = 0, 1, \dots$, is produced where the $(k + 1)$ st vector is obtained from the k th vector via (41). After finding the new estimates, renormalize them, for example, divide them by the last one, so that the last is then 1. Zermelo [16] first proposed this procedure, and a number of authors have considered it and variations since. Rao and Kupper [12] and Davidson [6] give modifications for the models with ties presented in the section titled Ties. Under certain conditions, this sequence of estimates will converge to the maximum likelihood estimates. In particular, if one object is always preferred, the maximum likelihood estimate does not exist, and the algorithm will fail. See [9] for a thorough and systematic presentation of these methods and their properties.

An alternative approach is given in [1], pages 436–438, that exhibits the Bradley–Terry model, including the one in (20), as a logistic regression model. Thus widely available software can be used to fit the model. The idea is to note that the data can be thought of as $(L/2)$ independent binomial random variables,

$$n_{ij} \sim \text{Binomial}(N_{ij}, p_{ij}), \quad \text{for } i < j. \quad (42)$$

Then under (1) and (3),

$$\log(\text{Odds}_{ij}) = \log(v_i) - \log(v_j) = \beta_i - \beta_j, \quad (43)$$

that is, the $\log(\text{Odds})$ is a linear function of the parameters $\beta_i (= \log(v_i))$'s. The constraint that $v_L = 1$ means $\beta_L = 0$. See [1] for further details.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.
- [2] Böckenholt, U. (2005). *Scaling of preferential choice*, *Encyclopedia of Behavioral Statistics*, Wiley, New York.
- [3] Bradley, R.A. & Terry, M.A. (1952). Rank analysis of incomplete block designs. I, *Biometrika* **39**, 324–345.
- [4] Daniels, H.E. (1950). Rank correlation and population models, *Journal of the Royal Statistical Society B* **12**, 171–181.
- [5] David, H.A. (1988). *The Method of Paired Comparisons*, 2nd Edition, Charles Griffin & Company, London.
- [6] Davidson, R.R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association* **65**, 317–328.
- [7] Davidson, R.R. & Farquhar, P.H. (1976). A bibliography on the method of paired comparisons, *Biometrics* **32**, 241–252.
- [8] Henery, R.J. (1983). Permutation probabilities for gamma random variables, *Journal of Applied Probability* **20**, 822–834.
- [9] Hunter, D.R. (2004). MM algorithms for generalized Bradley–Terry models, *Annals of Statistics* **32**, 386–408.
- [10] Luce, R.D. (1959). *Individual Choice Behavior*, Wiley, New York.
- [11] Luce, R.D. & Suppes, P. (1965). Preference, utility, and subjective probability, *Handbook of Mathematical Psychology Volume III*, Wiley, New York, pp. 249–410.
- [12] Rao, P.V. & Kupper, L.L. (1967). Ties in paired-comparison experiments: a generalization of the Bradley–Terry model (Corr: V63 p1550-51), *Journal of the American Statistical Association* **62**, 194–204.
- [13] Stern, H. (1990). Models for distributions on permutations, *Journal of the American Statistical Association* **85**, 558–564.
- [14] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.
- [15] Yellott, J.I. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution, *Journal of Mathematical Psychology* **15**, 109–144.
- [16] Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitrechnung, *Mathematische Zeitschrift* **29**, 436–460.

(See also **Attitude Scaling**)

JOHN I. MARDEN

Breslow–Day Statistic

MOLIN WANG AND VANCE W. BERGER

Volume 1, pp. 184–186

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Breslow–Day Statistic

The **case-control study** is often conducted to evaluate the association between exposure and disease in epidemiology. In order to control for potential confounders, one can stratify the data into a series of 2×2 tables, with one table for each value of the potential confounder. Table 1 shows the data in the i th of a series of 2×2 tables, for $i = 1, \dots, K$. If the association between exposure and disease is constant over strata, then Mantel–Haenszel estimator (see **Mantel–Haenszel Methods**), then $\hat{\psi}_{MH} = \sum_i R_i / \sum_i S_i$, where $R_i = a_i d_i / N_i$ and $S_i = b_i c_i / N_i$, is usually used to estimate the common odds ratio, ψ .

Breslow and Day [2, p.142] propose the statistic

$$Z^2(\hat{\psi}) = \sum_{i=1}^K \frac{a_i - \hat{e}_i(\hat{\psi})}{\hat{v}_i(\hat{\psi})} \quad (1)$$

for testing the homogeneity of the K odds ratios against the global alternative of heterogeneity. Here, $\hat{\psi}$ is the unconditional **maximum likelihood estimator** of ψ , $\hat{e}_i(\psi)$, the asymptotic expected number of exposed cases, is the appropriate solution to the quadratic equation

$$\frac{\hat{e}_i(\psi)[N_{0i} - t_i + \hat{e}_i(\psi)]}{[N_{1i} - \hat{e}_i(\psi)][t_i - \hat{e}_i(\psi)]} = \psi, \quad (2)$$

and $\hat{v}_i(\psi)$, the asymptotic variance of exposed cases, is given by

$$\hat{v}_i(\psi) = \left[\frac{1}{\hat{e}_i(\psi)} + \frac{1}{N_{0i} - t_i + \hat{e}_i(\psi)} + \frac{1}{N_{1i} - \hat{e}_i(\psi)} + \frac{1}{t_i - \hat{e}_i(\psi)} \right]^{-1}. \quad (3)$$

When the number of strata, K , is small and each table has large frequencies, $Z^2(\hat{\psi})$ is asymptotically (as each N_i gets large) distributed as χ^2 with

$K - 1$ degrees of freedom, under the homogeneity hypothesis.

Breslow and Day [2, p. 142] also suggest that a valid test can be based on $Z^2(\hat{\psi}_{MH})$. However, since $\hat{\psi}_{MH}$ is not efficient, Tarone [4] and Breslow [1] noted that $Z^2(\hat{\psi}_{MH})$ is stochastically larger than a χ^2 random variable (see **Catalogue of Probability Density Functions**) under the homogeneity hypothesis. The correct form for the test proposed by Tarone [4] is

$$Y^2(\hat{\psi}_{MH}) = \sum_{i=1}^K \frac{a_i - \hat{e}_i(\hat{\psi}_{MH})}{\hat{v}_i(\hat{\psi}_{MH})} - \frac{\sum_i a_i - \sum_i \hat{e}_i(\hat{\psi}_{MH})}{\sum_i \hat{v}_i(\hat{\psi}_{MH})}. \quad (4)$$

When the number of strata is small and each table has large frequencies, $Y^2(\hat{\psi}_{MH})$ follows an approximate χ^2 distribution on $K - 1$ degrees of freedom, under the homogeneity hypothesis. As noted by Breslow [1], since $\hat{\psi}_{MH}$ is nearly efficient, the correction term in $Y^2(\hat{\psi}_{MH})$ (the second term on the right-hand side of (4)) is frequently negligible. Because of the computational simplicity of $\hat{\psi}_{MH}$, the test statistic $Y^2(\hat{\psi}_{MH})$ is recommended in practice.

Shown in Table 2 is an example with $K = 2$ from Tarone [4] and Halperin et al. [3]. For this example,

$$\hat{\psi}_{MH} = 10.6317, \hat{e}_1(\hat{\psi}_{MH}) = 179.7951,$$

$$\hat{e}_2(\hat{\psi}_{MH}) = 765.2785,$$

$$\hat{v}_1(\hat{\psi}_{MH}) = 17.4537, \hat{v}_2(\hat{\psi}_{MH}) = 89.8137. \quad (5)$$

It follows that $Y^2(\hat{\psi}_{MH}) = 8.33$. Since $Y^2(\hat{\psi}_{MH})$ is asymptotically χ_1^2 under the null hypothesis of a common odds ratio, there is evidence of heterogeneity (P value = 0.0039). Note that the correction term in $Y^2(\hat{\psi}_{MH})$ is -0.24 .

Table 1 The i th of a series of 2×2 tables

Exposure	Observed frequencies		
	Case	Control	Total
Exposed	a_i	b_i	t_i
Unexposed	c_i	d_i	$N_i - t_i$
Total	N_{1i}	N_{0i}	N_i

Table 2 Example of $2 \times 2 \times 2$ table

Exposure	The 1st table			The 2nd table		
	Case	Control	Total	Case	Control	Total
Exposed	190	810	1000	750	250	1000
Unexposed	10	990	1000	250	750	1000
Total	200	1800	2000	1000	1000	2000

2 Breslow–Day Statistic

References

- [1] Breslow, N.E. (1996). Statistics in epidemiology: the case-control study, *Journal of the American Statistical Association* **91**, 14–28.
- [2] Breslow, N.E. & Day N.E. (1980). *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*, IARC, Lyon.
- [3] Halperin, M., Ware, J.H., Byar, D.P., Mantel, N., Brown, C.C., Koziol, J., Gail, M. & Green, S.B. (1977). Testing for interaction in and $I \times J \times K$ contingency table, *Biometrika* **64**, 271–275.
- [4] Tarone, R.E. (1985). On heterogeneity tests based on efficient scores, *Biometrika* **72**, 91–95.

MOLIN WANG AND VANCE W. BERGER

Brown, William

PAT LOVIE

Volume 1, pp. 186–187

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Brown, William

Born: December 5, 1881, in Sussex, England.

Died: May 17, 1952, in Berkshire, England.

Putting William Brown into any particular pigeonhole is difficult: he was psychologist, psychiatrist, administrator, and, chiefly in the early days of his career, psychometrician.

The son of a schoolmaster, Brown attended a local school until winning a scholarship to Christ's College, Oxford in 1899. He spent the next six years in Oxford taking Mathematical Moderations in 1902, Final Honours in Natural Science (Physiology) in 1904, and finally *Literae Humaniores* with psychology as a special subject the following year. After a spell in Germany as a John Locke Scholar in Mental Philosophy during 1906, he returned to England to continue medical and statistical studies in London. By 1909 Brown had obtained a lectureship in psychology at King's College, London, followed by readership in 1914, the same year that he qualified in medicine. Meanwhile, he had been working in **Karl Pearson's** Biometrical Laboratory at University College London, earning a DSc and the coveted Carpenter medal in 1910 for a pioneering examination of how Pearson's correlational methods could be applied to psychological measurement.

The following year, Brown published *The Essentials of Mental Measurement* [1] based partly on his DSc work. This book, with its criticisms (almost certainly encouraged by Pearson) of Spearman's two factor theory of mental ability, propelled Brown into the center of a long running dispute with **Spearman**, while at the same time landing himself an ally in **Godfrey Thomson**, who would become an even more implacable critic of Spearman's notions. Subsequent editions of Brown's book were coauthored with Thomson, although it is quite clear that Thomson took the major role in the revisions and their renewed attacks on Spearman. While Thomson and Spearman were never reconciled, Brown eventually recanted in 1932 [3], very publicly going over to Spearman's side.

Brown's experiences as an RAMC officer treating shell-shock victims during the First World

War had shifted him ever more toward psychiatry and psychotherapy, and, once back in civilian life, he began to acquire more medical qualifications (DM in 1918, then MRCP and FRCP in 1921 and 1930, respectively). In 1921 he resigned his post in King's College, returning to Oxford to the Wilde Readership in Mental Philosophy. By the late 1920s, Brown was somehow juggling his work in Oxford, a clinical practice that included appointments as a psychotherapist at King's College and Bethlem Royal Hospitals, and writing prolifically on psychiatry and psychotherapy as well as making the occasional foray into psychometrics. Brown also played a significant role in establishing the Institute of Experimental Psychology in Oxford and was its first Director from 1936 to 1945.

After retiring in 1946, Brown continued with his writing and also remained active in other academic areas, for instance, serving as President of the British Psychological Society for 1951 to 1952. According to Godfrey Thomson, only a few weeks before his death in 1952, Brown was intent on making a return to the psychometric work that had launched him into prominence within the psychological establishment more than 40 years earlier.

What can we say of William Brown's legacy to psychological statistics? His criticism of Spearman's radical notions about mental ability certainly stirred up debate within British psychology and brought Godfrey Thomson into the emerging factor analysis arena. But we remember him chiefly for the Spearman–Brown coefficient (or prophesy formula) for determining the effect of test length on reliability. Unlike certain other cases, such as Spearman's rank correlation, exactly who should be credited with this particular measure is quite clear: in back-to-back articles published in 1910, we find the coefficient set out in a general form by Spearman [5] and then, in almost the same way as it is commonly used today, by Brown [2] (see [4]).

References

- [1] Brown, W. (1911). *The Essentials of Mental Measurement*, Cambridge University Press, London.
- [2] Brown, W. (1910). Some experimental results in the correlation of mental abilities, *British Journal of Psychology* 3, 296–322.

2 Brown, William

- [3] Brown, W. (1932). The mathematical and experimental evidence for the existence of a central intellectual factor (g), *British Journal of Psychology* **23**, 171–179.
- [4] Levy, P. (1995). Charles Spearman's contribution to test theory, *British Journal of Mathematical and Statistical Psychology* **48**, 221–235.
- [5] Spearman, C. (1910). Correlation calculated from faulty data, *British Journal of Psychology* **3**, 271–295.

PAT LOVIE

Bubble Plot

BRIAN S. EVERITT

Volume 1, pp. 187–187

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bubble Plot

The simple xy **scatterplot** has been in use since at least the eighteenth century and is the primary data-analytic tool for assessing the relationship between a pair of continuous variables. But the basic scatterplot can only accommodate two variables, and there have been various suggestions as to how it might be enhanced to display the values of further variables. The simplest of these suggestions is perhaps a graphic generally known as a *bubble plot*. Here, two variables are used in the normal way to construct a scatterplot, and the values of a third variable are represented by circles with radii proportional to these values, centered on the appropriate point in the underlying scatterplot. Figure 1 shows an example of a bubble plot for the chest, hips, and waist measurements (in inches) of 20 individuals.

Bubble plots are often a useful supplement to the basic scatterplot, although, when large numbers of observations are plotted, the diagram can quickly

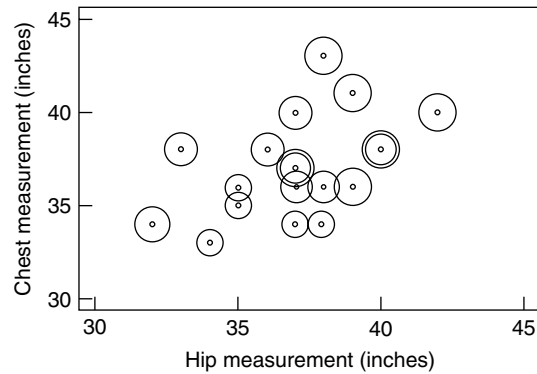


Figure 1 Bubble plot of hip, chest, and waist measurement; the last is represented by the radii of the circles

become difficult to read. In such cases, a **three-dimensional scatterplot** or a **scatterplot matrix** may be a better solution.

BRIAN S. EVERITT

Burt, Cyril Lodowic

BRIAN S. EVERITT

Volume 1, pp. 187–189

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Burt, Cyril Lodowic

Born: March 3, 1883, in Stratford-upon-Avon.

Died: October 10, 1971, in London.

Cyril Burt was the son of a house-physician at Westminster Hospital and could trace his lineage through his father's mother to Newton's mathematical tutor at Cambridge, Sir Isaac Barrow. As a schoolboy, Burt first attended King's School, Warwick, and then in 1895 Christ's Hospital School in London as a boarder. He studied at Jesus College, Oxford, graduating with a degree in classics and philosophy. After completing his degree, Burt traveled to the University of Würzburg in Germany to study psychology under Oswald Külpe, and then returned to Oxford to become the John Locke scholar in mental philosophy.

In 1908, Burt became lecturer in experimental psychology at the University of Liverpool. It was here that his long research career began with a study comparing the intelligence of boys enrolled in an elite preparatory school with the intelligence of boys attending a regular school. Using measures such as mirror drawing that were unlikely to have been learnt during the student's lifetime, Burt showed that the prep school boys scored higher than the boys from the regular school and concluded that they had more innate intelligence. In addition, he noted that the fathers of the prep school students were more successful than the fathers of the other boys, a finding he interpreted as meaning that the prep school boys had benefited from their fathers' superior genetic endowments [1]. Burt, however, did not believe that 100% of intelligence is inherited and acknowledged that environmental influences are also important.

In 1913, Burt became Chief Psychologist for the London County Council and was responsible for the administration and interpretation of mental tests in London's schools. During this time, he developed new tests [2], a special school for the handicapped, and founded child guidance clinics. After twenty years of working for the LCC, Burt took up the Chair of Psychology at University College, London, recently vacated by **Charles Spearman**, where he remained until his retirement in 1950. Early in his career, Burt had worked with Spearman on various

aspects of intelligence and **factor analysis**, clearly invented by Spearman in 1904 [9], although this did not prevent Burt trying to claim the technique as his own later in his life (see both **History of Factor Analyses** entries).

Burt believed that intelligence levels were largely fixed by the age of 11, and so were accurately measurable by standard tests given at that age. On the basis of this belief, he became one of several influential voices that helped introduce the so-called eleven plus examination system in Great Britain, under which all 11-year-olds were given a series of academic and intelligence tests, the results of which determined their schooling for the next 5 to 7 years. And partly for this work, Burt, in 1946, became the first psychologist to be knighted, the Labour Government of the day bestowing the honor for his work on psychological testing and for making educational opportunities more widely available. Whether all those school children whose lives were adversely affected by the result of an examination taken at 11 would agree that the 11+ really made 'educational opportunities more widely available', is perhaps debatable.

In his working lifetime, Burt was one of the most respected and honored psychologists of the twentieth century. Indeed, according to Hans Eysenck in the obituary of Burt he wrote for *The British Journal of Mathematical and Statistical Psychology* [3], Sir Cyril Burt was 'one of England's outstanding psychologists'. But during his long retirement, he published over 200 articles, amongst them several papers that buttressed his hereditarian claim for intelligence by citing very high correlations between IQ scores of identical twins raised apart; according to Burt, these twins were separated in early childhood and raised in different socioeconomic conditions. Burt's study stood out among all others because he had found 53 pairs of such twins, more than twice the total of any previous attempt. Burt's methodology was generously and largely uncritically praised by some other academic psychologists, for example, Hans Eysenck and Arthur Jensen, but after Burt's death, closer scrutiny of his work by Kamin [7, 8] suggested at the best inexcusable carelessness and at worst conscious fraud and fakery. Kamin noticed, for example, that while Burt had increased his sample of twins from fewer than 20 to more than 50 in a series of publications, the average correlation between pairs for IQ remained

unchanged to the third decimal place. This statistically implausible result, allied to Burt's 'missing' coauthors, and a relatively random distribution of Burt's twins to families from various socioeconomic strata, led many to the uncomfortable conclusion that Burt may have fraudulently manufactured the data to support his belief that intelligence is largely inherited. This conclusion is supported in the biography of Burt published in 1979 [5] despite Hearnshaw's great respect for Burt and his initial skepticism about the accusations being made. Later accounts of Burt's work [4, 6] claim that the case for fraud is not proven and that Burt's critics are guilty of selective reporting. Whatever the truth of the matter, the last sentence of Eysenck's obituary of Burt now has a somewhat sad and ironic ring

'...as the first editor of this Journal (BJMSP) ... he set a very high standard of critical appraisal... This critical faculty combined with his outstanding originality, his great insight and his profound mathematical knowledge, makes him a truly great psychologist: his place in the history books of our science is assured.'

References

- [1] Burt, C. (1909). Experimental tests of general intelligence, *British Journal of Psychology* **3**, 94–177.
- [2] Burt, C. (1921). *Mental and Scholastic Tests*, P.S. King and Son, London.
- [3] Eysenck, H.J. (1972). Sir Cyril Burt (1883–1971), *British Journal of Mathematical and Statistical Psychology* **25**, i–iv.
- [4] Fletcher, R. (1991). *Science, Ideology and the Media*, Transaction Publishers, New Brunswick.
- [5] Hearnshaw, L.S. (1979). *Cyril Burt, Psychologist*, Cornell University Press, Ithaca.
- [6] Joynton, R.B. (1989). *The Burt Affair*, Routledge, New York.
- [7] Kamin, L.J. (1974). *The Science and Politics of IQ*, Wiley, New York.
- [8] Kamin, L.J. (1981). *Intelligence: The Battle for the Mind*, Macmillan, London.
- [9] Spearman, C. (1904). General intelligence: objectively measured and determined, *American Journal of Psychology* **15**, 201–299.

BRIAN S. EVERITT

Bush, Robert R

R. DUNCAN LUCE

Volume 1, pp. 189–190

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Bush, Robert R

Born: July 20, 1920, in Albion, Michigan.

Died: January 4, 1972, New York City.

Robert R. Bush first studied electrical engineering at Michigan State University and then physics at Princeton University, receiving a Ph.D. in 1949. A NRC/SSRC two-year post doc for natural scientists to study social science took him to Harvard's Department of Social Relations under the tutelage of statistician Frederick Mosteller. After first publishing a chapter on statistics for social scientists [5], a research project ensued on the statistical modeling of learning [3]. Their basic idea was that an individual selecting among finitely many alternatives on trial i is characterized by a vector of probabilities $\vec{p}_i$. Following a choice, reinforcement r occurs that is represented by a Markov process (see **Markov Chains**) with $\vec{p}_{i+1} = Q_r \circ \vec{p}_i$, where Q_r is a vector operator. They restricted attention to linear operators; others explored nonlinear ones as well. Bush focused on comparing various models and carrying out experiments to test them (see [1], [2] and [4]). By the mid-1960s, his (and psychology's) interest in such modeling waned. Mosteller, Galanter, and Luce speculate about four underlying reasons: (a) the empirical defeat of the gradual learning models by the all-or-none models developed by the mathematical psychology group at Stanford University; (b) the inability to arrive at a sensible account of the strong resistance to extinction in the face of partial reinforcement; (c) not finding a good way to partition the parameters of the model into aspects attributable to the individual, which should be invariant over experimental designs, and those attributable to the boundary conditions of the design; and (d) the paradigm shift to information processing and memory models [6]. (Note that this obituary describes Bush's career and research in considerable detail and lists all of his papers). Nonetheless, his modeling work influenced others, leading to the excellent synthesis by Norman [7], the important conditioning model of Rescorla and Wagner [8], several attempts to account for sequential effects in signal detection in terms of adjustments of the response criterion, and in a variety of other learning situations to this day arising in several fields: computer modeling, neural networks, and sociology.

In addition to his research, Bush was a superb and devoted teacher, with a number of Ph.D. students at major universities, and a very gifted administrator. During the 1950s, he helped organize and lead a series of summer workshops, mainly at Stanford. These led to recruiting young scientists, much research, several books, and the *Journal of Mathematical Psychology*. In 1956, he became an associate professor of applied mathematics at the New York School of Social Work. During that period, he, Galanter, and Luce met many weekends to carry out research, and they spawned the unlikely idea of his becoming chair of psychology at the University of Pennsylvania, one of the oldest such departments. Adventuresome administrators at Penn decided that a radical change was needed. Their gamble was justified by Bush's creating a powerful department which, to this day, has remained excellent. His success was so great that many expected he would move up the administrative ladder, but the necessary ceremonial aspects were anathema to him. Instead, in 1968, he became chair at Columbia. This move was motivated, in part, by his passion for ballet, especially the American Ballet Theater to which he devoted his money and administrative talents as a fund-raiser. The four years in New York, until his death, were a mixture of frustration at administrative indecision and deteriorating health.

References

- [1] Atkinson, R.C. (1964). *Studies in Mathematical Psychology*, Stanford University Press, Stanford.
- [2] Bush, R.R. & Estes, W.K., (1959). *Studies in Mathematical Learning Theory*, Stanford University Press, Stanford.
- [3] Bush, R.R. & Mosteller, F. (1955). *Stochastic Models for Learning*, Wiley, New York.
- [4] Luce, R.D., Bush, R.R. & Galanter, E., (1963, 1965). *Handbook of Mathematical Psychology*, Vols. 1, 2 & 3, Wiley, New York.
- [5] Mosteller, F. & Bush, R.R. (1954). Selected quantitative techniques, in *Handbook of Social Psychology*, G. Lindzey, ed., Addison-Wesley, Cambridge, pp. 289–334.
- [6] Mosteller, F., Galanter, E. & Luce, R.D. (1974). Robert R. Bush, *Journal of Mathematical Psychology* **11**, 163–189.

2 **Bush, Robert R**

- [7] Norman, M.F. (1972). *Markov Processes and Learning Models*, Academic Press, New York.
- [8] Rescorla, R.A. & Wagner, A.T. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement, in *Classical*

Conditioning II, A. Black & W.F. Prokasy, eds, Appleton Century Crofts, New York, pp. 64–99.

R. DUNCAN LUCE

Calculating Covariance

GILAD CHEN

Volume 1, pp. 191–191

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Calculating Covariance

Covariance captures the extent to which two variables vary together in some systematic way, or whether variance in one variable is associated with variance in another variable. A sample covariance between variable X and variable Y (noted as COV_{xy} or s_{xy}) is an estimate of the population parameter σ_{xy} , and is defined mathematically as:

$$s_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{\sum xy}{N - 1} \quad (1)$$

Where $\bar{X}$ is the sample mean on X , $\bar{Y}$ is the sample mean on Y , $\sum xy$ is the sum of the cross products deviations of pairs of X and Y scores from their respective means (also known as the *sum of cross products*), and N is the sample size. Covariance values can be positive, negative, or zero. A positive covariance indicates that higher scores on X are paired with higher scores in Y , whereas a negative covariance indicates that scores on X are paired with lower scores on Y . When a covariance is equal to zero, X and Y are not linearly associated with each other.

The covariance is similar conceptually to the correlation coefficient (r), and in fact the two indices are perfectly and linearly correlated. It can be shown that the highest possible covariance value for variables X and Y is obtained when X and Y are perfectly correlated (i.e., correlated at -1.0 or $+1.0$) [1]. However, the covariance is scale-dependent, in that its value is heavily dependent on the unit of measurement used for X and Y . In contrast, the correlation coefficient is a standardized covariance and can vary between -1.0 and $+1.0$. Everything else being equal, the same correlation between X and Y will be represented by a larger covariance value when (a) the units of measurement for X and Y are larger, and (b) the variances of X and Y are larger. Thus, the covariance can indicate whether X and Y are correlated in a positive or negative way, but it is not very useful as an indicator of the strength of association between X and Y .

Reference

- [1] Howell, D. (1997). *Statistical Methods for Psychology*, 4th Edition, Wadsworth, Belmont.

GILAD CHEN

Campbell, Donald T

SANDY LOVIE

Volume 1, pp. 191–192

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Campbell, Donald T

Born: November 20, 1916, in Michigan, USA.

Died: May 5, 1996, in Pennsylvania, USA.

'It is the fate of rebels to found new orthodoxies', so begins one of Bertrand Russell's *Nightmares of Eminent Persons*. And, so he might have added, it is also the fate of the holders of such orthodoxies, particularly if they stay active and high profile long enough, to be challenged in their turn by a new generation of rebels. The social theorist Donald T. Campbell is a prime example of this cyclical rule in that he initially made his name by aligning himself with the 'social physics' form of positivism and empirical realism at a time when hardly any of the social sciences employed such hard-nosed experimental and quantitative moves, then saw his new approach flourish in social psychology from the 1950s to the early 1970s (helped in no small way by Campbell's own efforts), only to be overtaken by the relativistic and postpositive backlash whose intellectual form had been sketched out in Thomas Kuhn's 1962 poetic masterpiece, *The Structure of Scientific Revolutions*. There is, however, the possibility of an even more intriguing (if somewhat retrospectively applied) twist to the tale in that by the end of his career, Campbell seemed to regard himself as one of the very first *antipositivist* social psychologists in the world (see [3], p. 504)!

Campbell enrolled at the University of California at Berkeley in the autumn of 1937, where he obtained his AB in 1939 (majoring in psychology). In the same year, he became a graduate student at Berkeley on a doctoral topic supervised by Harold Jones of the Institute of Child Welfare. This was, however, interrupted by two events: first, from 1940 to 1941, he took advantage of a travelling Fellowship to work at Harvard under Henry Murray (where he heard the social psychologist Gordon Allport lecture), and second by America's entry into World War II in 1941. Although his war service somewhat delayed his PhD work, the doctorate was awarded to him in 1947 for his research into the generality of social attitudes amongst five ethnic groups. Characteristically, Campbell's own account spends more time worrying about the methodology of the research than its content ([3],

p. 6). His academic rise was swift, from two stretches as assistant professor, first at Ohio State University and then at the University of Chicago (1947 to 1953), finally to an associate professorship at Northwestern University from 1953 until his retirement as emeritus professor in 1979. He achieved several major honors during his life: he was, for instance, awarded the American Psychological Association's Kurt Lewin Prize in 1974 and became the APA's President the following year. The National Academy of Sciences also honoured him with a Distinguished Scientific Contribution Award in 1970.

Although Donald Campbell is primarily judged by psychologists to be a methodologist, particular instances being his highly influential work on **multitrait-multimethod** and **quasi-experiments** [1], he is better understood as a theoretician of the social sciences who, because of his early and deep commitment to empirical realism and positivism, viewed improvements in *method* as the key to reliable truth. However, one could also characterize his work over the long term as attempts to fend off (or at least ameliorate) the more extreme aspects of a range of what he otherwise considered to be attractive philosophies and methodologies of science. Thus, *single variable operationism* was replaced by *multiple variable operationism* in his multitrait-multimethod work, without, however, dropping the essential commitment to *operationism*. Quasi-experimentation (see **Quasi-experimental Designs**) similarly was an explicit attempt to generalize standard Fisherian design and analysis (with its intellectual and quantitative rigor) from the tightly controlled psychological laboratory to the anarchic society outside on the street [2].

References

- [1] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
- [2] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-experimental Designs for Research*, Rand McNally, New York.
- [3] Overman, E.S. & Campbell, D.T. (1988). *Methodology and Epistemology for Social Science: Selected Papers*, University of Chicago Press, Chicago.

SANDY LOVIE

Canonical Correlation Analysis

BRUCE THOMPSON

Volume 1, pp. 192–196

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistical method employed to investigate relationships among two or more variable sets, each consisting of at least two variables [7, 9, 10]. If a variable set consists of fewer than two variables, then canonical analysis is typically called something else, such as a *t* Test or a regression analysis (*see Multiple Linear Regression*). In theory, the canonical logic can be generalized to more than two variable sets [4], but in practice most researchers use CCA only with two variable sets, partly because commonly available software only accommodates this case.

Researchers use CCA and other multivariate methods (*see Multivariate Analysis: Overview*), instead of univariate methods (e.g., regression), for two reasons. First, CCA avoids the inflation of experiment-wise Type I error rates that typically results when multiple univariate analyses are conducted (*see Multiple Testing*). Second, by simultaneously considering all the variables in a single analysis, CCA honors the ecological reality that in nature all the variables can interact with each other. In other words, the same data when analyzed with multiple univariate methods may yield (a) statistically nonsignificant results and (b) zero effect sizes, but when analyzed with CCA may yield (a) statistically significant results and (b) huge effect sizes. And we tend to believe the multivariate results in such cases, because we tend to believe that in reality all the variables do interact with each other, and that only an analysis that honors this possibility generalizes well to reality.

In addition to being important from an applied research point of view, CCA also is important heuristically because it is a very general case of the **general linear model** [5]. The general linear model is the recognition that *all* analyses are correlational and yield r^2 -type effect sizes (e.g., percentage variance explained R^2 , Cohen's η^2), and all yield weights that are applied to the measured variables to produce estimates of the *latent* or *composite scores* (e.g., regression $\hat{Y}$ scores) that are actually the focus of the analysis. In other words, one can conduct a *t* Test using a CCA program, but not vice versa; an **analysis of variance** (ANOVA) using a CCA program,

but not vice versa; a **descriptive discriminant analysis** with a CCA program, but not vice versa; and so forth.

The first step in a CCA, performed automatically by software, is the computation of the bivariate product-moment **correlation matrix** involving all the variables [7]. Then the correlation matrix is partitioned into quadrants, honoring variable membership in the two variable sets. These quadrants are then multiplied times each other using matrix algebra. The resulting quadruple-product matrix is then subjected to a **principal component analysis** to yield the primary CCA results.

Given the computational starting point for CCA, most researchers use CCA when all the variables are intervally scaled. However, some researchers use CCA with dichotomous data or with a mixture of intervally scaled and dichotomous data (*see Measurement: Overview*). The primary difficulty with such analyses is that CCA assumes the data are multivariate normal (*see Catalogue of Probability Density Functions*), and data cannot be perfectly multivariate normal if some or all of the variables are dichotomous.

The fact that CCA invokes a principal component analysis [11] suggests the possibility of using **factor analysis** rather than CCA. Indeed, if in the researcher's judgment the variables do not constitute meaningful sets (e.g., variables measured at two different points in time, academic outcome variables versus personality variables), then factor analysis would be the appropriate way to explore relationships among the variables existing as a single set. But if sets are present, only CCA (and not factor analysis) honors the existence of the variable sets as part of the analysis.

CCA yields *canonical functions* consisting of both standardized weights, similar to regression beta weights (*see Standardized Regression Coefficients*), that can be used to derive *canonical scores*, and *structure coefficients*, r_s , which are bivariate correlations of the measured variables with the canonical composite scores [7, 10]. The number of functions equals the number of variables in the smaller of the two variable sets and each function represents two weighted sums, one for each set of variables. Each function also yields a *canonical correlation coefficient* (R_C) ranging from 0.0 to 1.0, and a squared canonical correlation coefficient (R_C^2). One criterion that is optimized by CCA is that on a given

2 Canonical Correlation Analysis

function the weights, called *standardized canonical function coefficients*, optimize R_C^2 just as regression beta weights optimize R^2 .

The canonical functions are uncorrelated or orthogonal. In fact, the functions are ‘bi-orthogonal’. For example, if there are two functions in a given analysis, the canonical scores on Function I for the criterion variable set are (a) perfectly uncorrelated with the canonical scores on Function II for the criterion variable set and (b) perfectly uncorrelated with the canonical scores on Function II for the predictor variable set. Additionally, the canonical scores on Function I for the predictor variable set are (a) perfectly uncorrelated with the canonical scores on Function II for the predictor variable set and (b) perfectly uncorrelated with the canonical scores on Function II for the criterion variable set. Each function theoretically can yield squared canonical correlations that are 1.0. In this case, because the functions are perfectly uncorrelated, each function perfectly explains relationships of the variables in the two variable sets, but does so in a unique way.

As is the case throughout the general linear model, interpretation of CCA addresses two questions [10]:

1. ‘Do I have anything?’, and, if so
2. ‘Where does it (my effect size) come from?’.

The first question is addressed by consulting some combination of evidence for (a) statistical significance, (b) effect sizes (e.g., R_C^2 or adjusted R_C^2 [8]), and (c) result replicability (see **Cross-validation; Bootstrap Inference**). It is important to remember

that in multivariate statistics one can only test the statistical significance of a function as a single function (i.e., the last function), unless one uses a **structural equation modeling** approach to the analysis [2].

If the researcher decides that the results reflect nothing, the second question is rendered irrelevant, because the sensible researcher will not ask, ‘From where does my nothing originate?’ If the researcher decides that the results reflect more than nothing, then both the standardized function coefficients and the structure coefficients must be consulted, as is the case throughout the general linear model [1, 3]. Only variables that have weight and structure coefficients of zero on all functions contribute nothing to the analysis.

A heuristic example may be useful in illustrating the application. The example is modeled on real results presented by Pitts and Thompson [6]. The heuristic presumes that participants obtain scores on two reading tests: one measuring reading comprehension when readers have background knowledge related to the reading topic (SPSS variable name *read_yes*) and one when they do not (*read_no*). These two reading abilities are predicted by scores on vocabulary (*vocabulr*), spelling (*spelling*), and self-concept (*self_con*) tests. The Table 1 data cannot be analyzed in SPSS by point-and-click, but canonical results can be obtained by executing the syntax commands:

```
MANOVA read_yes read_no WITH
  vocabulr spelling self_con/
  PRINT=SIGNIF(MULTIV EIGEN
  DIMENR) /
```

Table 1 Heuristic data

Person	SPSS Variable Names						
	<i>read_yes</i>	<i>read_no</i>	<i>vocabulr</i>	<i>spelling</i>	<i>self_con</i>	CRIT1	PRED1
Herbert	61 (−1.49)	58 (1.41)	81 (−1.70)	80 (−0.72)	68 (−0.67)	−1.55	−1.59
Jerry	63 (−1.16)	54 (0.48)	88 (−0.59)	92 (1.33)	84 (1.21)	−1.18	−1.07
Justin	65 (−0.83)	51 (−0.21)	87 (−0.75)	84 (−0.03)	73 (−0.08)	−0.82	−0.77
Victor	67 (−0.50)	49 (−0.67)	86 (−0.90)	77 (−1.23)	63 (−1.25)	−0.46	−0.50
Carol	69 (−0.17)	45 (−1.59)	89 (−0.43)	76 (−1.40)	68 (−0.67)	−0.09	−0.31
Deborah	71 (0.17)	48 (−0.90)	95 (0.52)	84 (−0.03)	80 (0.74)	0.21	0.18
Gertrude	73 (0.50)	49 (−0.67)	99 (1.16)	87 (0.48)	85 (1.32)	0.53	0.63
Kelly	75 (0.83)	52 (0.02)	98 (1.00)	86 (0.31)	75 (0.15)	0.82	1.04
Murray	77 (1.16)	55 (0.72)	95 (0.52)	82 (−0.38)	61 (−1.49)	1.12	1.29
Wendy	79 (1.49)	58 (1.41)	99 (1.16)	94 (1.67)	80 (0.74)	1.42	1.10

Note: The z -score equivalents of the five measured variables are presented in parentheses. The scores on the canonical composite variables (e.g., CRIT1 and PRED1) are also in z -score form.

Table 2 Canonical results

SPSS Variable	Function I			Function II		
	Function	r_s	Squared	Function	r_s	Squared
read_yes	1.002	0.999	99.80%	-0.018	0.046	0.21%
read_no	-0.046	0.018	0.03%	1.001	1.000	100.00%
Adequacy			49.92%			50.11%
Redundancy			48.52%			36.98%
R_c^2			97.20%			73.80%
Redundancy			30.78%			8.28%
Adequacy			31.67%			11.22%
vocabulr	1.081	0.922	85.01%	-0.467	-0.077	0.59%
spelling	0.142	0.307	9.42%	1.671	0.575	33.06%
self_con	-0.520	0.076	0.58%	-1.093	-0.003	0.00%

Note: The canonical adequacy coefficient equals the average squared structure coefficient for the variables on a given function [7, 10]. The canonical redundancy coefficient equals the canonical adequacy coefficient times the R_c^2 [7, 9].

DISCRIM=STAN CORR
ALPHA(.999) / .

Table 2 presents the canonical results organized in the format recommended elsewhere [9, 10]. The standardized function coefficients can be used to compute the composite or latent scores on the canonical functions. For example, on the first canonical function, Wendy's criterion composite score on Function I would be her z -scores times the Function I criterion standardized canonical function coefficients ($[1.49 \times 1.002] + [1.41 \times -0.046] = 1.42$). Wendy's predictor composite score on Function I would be her z -scores times the Function I predictor standardized canonical function coefficients ($[1.16 \times 1.081] + [1.67 \times 0.142] + [0.74 \times -0.520] = 1.10$). It is actually the composite scores that are the focus of the CCA, and *not* the measured variables (e.g., read_yes, vocabulr). The measured variables are useful primarily to obtain the estimates of the construct scores. For example, the Pearson r^2 between the Function I composite scores in Table 1 is 0.972. This equals the R_c^2 on Function I, as reported in Table 2.

The heuristic data are useful in emphasizing several points:

1. The standardized canonical function coefficients, like regression beta weights [1], factor pattern coefficients [11], and so forth, are not generally correlation coefficients (e.g., 1.671, -1.093, and 1.081), and therefore cannot be interpreted as measuring the strength of relationship.
2. Because the canonical functions are orthogonal, all functions theoretically could have squared canonical correlation coefficients (here 0.972 and 0.738) of 1.0, and do not sum to 1.0 across functions.
3. Even variables essentially uncorrelated with the measured variables in another variable set (e.g., self_con) can be useful in improving the effect size, as reflected by such variables having near-zero structure coefficients (e.g., 0.076 and -0.003) but large function coefficients (e.g., -0.520 and -1.093).

The latter dynamic of 'suppression' can occur in canonical analysis, just as it can occur throughout the general linear model [1], in analyses such as the regression analysis or descriptive discriminant analysis.

References

- [1] Courville, T. & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough, *Educational and Psychological Measurement* **61**, 229–248.
- [2] Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling* **4**, 65–79.
- [3] Graham, J.M., Guthrie, A.C. & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: a reminder, *Structural Equation Modeling* **10**, 142–153.
- [4] Horst, P. (1961). Generalized canonical correlations and their applications to experimental data, *Journal of Clinical Psychology* **26**, 331–347.

4 Canonical Correlation Analysis

- [5] Knapp, T.R. (1978). Canonical correlation analysis: a general parametric significance testing system, *Psychological Bulletin* **85**, 410–416.
- [6] Pitts, M.C. & Thompson, B. (1984). Cognitive styles as mediating variables in inferential comprehension, *Reading Research Quarterly* **19**, 426–435.
- [7] Thompson, B. (1984). *Canonical Correlation Analysis: Uses and Interpretation*, Sage, Newbury Park.
- [8] Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: a Monte Carlo study, *Educational and Psychological Measurement* **50**, 15–31.
- [9] Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis, *Measurement and Evaluation in Counseling and Development* **24**, 80–95.
- [10] Thompson, B. (2000). Canonical correlation analysis, in *Reading and Understanding More Multivariate Statistics*, L. Grimm & P. Yarnold, eds, American Psychological Association, Washington, pp. 285–316.
- [11] Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, American Psychological Association, Washington.

BRUCE THOMPSON

Carroll–Arabic Taxonomy

JAN DE LEEUW

Volume 1, pp. 196–197

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Carroll–Arabie Taxonomy

Text

A large number of computerized scaling techniques were developed in the wake of the pioneering work of Shepard, Kruskal, and Guttman [4–6, 8, 9] (*see Multidimensional Scaling*). There have been various attempts to bring some order into this bewildering variety of techniques. Books such as [7] or review articles such as [3] are organized with a clear taxonomy in mind, but the most well-known and comprehensive organization of scaling methods is due to Carroll and Arabie [1].

Before we discuss the taxonomy, we have to emphasize two important points. First, the proposed organization of scaling methods is clearly inspired by earlier work of Coombs [2] and Shepard [10]. The exquisite theoretical work of Coombs was written before the computer revolution, and the scaling methods he proposed were antiquated before they were ever seriously used. This had the unfortunate consequence that the theoretical work was also largely ignored. The same thing is more or less true of the work of Shepard, who actually did propose computerized algorithms, but never got them beyond the stage of research software. Again, this implied that his seminal contributions to multidimensional scaling have been undervalued. Both Coombs and Shepard had some followers, but they did not have an army of consumers who used their name and referred to their papers. The second important aspect of the Carroll–Shepard taxonomy is that it was written around 1980. In the subsequent 25 years, hundreds of additional metric and nonmetric scaling methods have been developed, and some of them fall outside the boundaries of the taxonomy. It is also probably true that the messianistic zeal with which the nonmetric methods were presented around 1970 has subsided. They are now much more widely employed, in many different disciplines, but shortcomings have become apparent and the magic has largely dissipated.

The actual taxonomy is given in Table 1. We give a brief explanation of the concepts that are not self-evident. The ‘number of ways’ refers to the dimensionality of the data array and the ‘number of modes’ to the number of sets of objects that must be represented. Thus a symmetric matrix of proximities has two ways but one mode. ‘Scale type’

Table 1 Carroll–Arabie taxonomy of scaling methods

-
- Data.
 - Number of Modes.
 - Number of Ways.
 - Scale Type of Data.
 - Conditionality.
 - Completeness.
 - Model.
 - Spatiality.
 - * Spatial.
 - Distance.
 - Scalar Products.
 - * Nonspatial.
 - Partitions.
 - Subsets.
 - Trees.
 - Number of Sets of Points.
 - Number of Spaces.
 - External Constraints.
-

refers to the usual nominal, ordinal, and numerical distinction. ‘Conditionality’ defines which elements of the data array can be sensibly compared. Thus a matrix with preference rank order in each row is row-conditional. Matrices with similarity rankings between, say, colors, by a number of different subjects, gives three-way, two-mode, ordinal, matrix conditional data. ‘Completeness’ refers to missing data, sometimes in the more theoretical sense in which we say that unfolding data are incomplete, because they only define an off-diagonal submatrix of the complete similarity matrix (*see Multidimensional Unfolding*).

The taxonomy of models is somewhat dated. It is clear the authors set out to classify the existing scaling techniques, more specifically the computerized ones they and their coworkers had developed (which happened to be a pretty complete coverage of the field at the time). We can clearly distinguish the nonmetric scaling methods, the influence of using Minkovski power metrics, the work on cluster analysis (*see Cluster Analysis: Overview*) and additive partitioning, and the work on internal and external analysis of preferences. Some clarifications are perhaps needed. ‘Number of spaces’ refers to either a joint or a separate representation of the two modes of a matrix (or the multiple modes of an array). Such considerations are especially important in off-diagonal methods such as unfolding or correspondence analysis. ‘External’ analysis implies that coordinates in one of the spaces in which we are representing our data are fixed

2 Carroll–Arabie Taxonomy

(usually found by some previous analysis, or defined by theoretical considerations). We only fit the coordinates of the points in other spaces; for instance, we have a two-dimensional space of objects and we fit individual preferences as either points or lines in that space.

In summary, we can say that the Carroll–Arabie taxonomy can be used to describe and classify a large number of scaling methods, especially scaling methods developed at Bell Telephone Laboratories and its immediate vicinity between 1960 and 1980. Since 1980, the field of scaling has moved away to some extent from the geometrical methods and the heavy emphasis on solving very complicated optimization problems. Item response theory and choice modeling have become more prominent, and they are somewhat at the boundaries of the taxonomy. New types of discrete representations have been discovered. The fact that the taxonomy is still very useful and comprehensive attests to the importance of the frameworks developed during 1960–1980, and to some extent also to the unfortunate fact that there no longer is a center in psychometrics and scaling with the power and creativity of Bell Labs in that area.

References

- [1] Carroll, J.D. & Arabie, P. (1980). Multidimensional scaling, *Annual Review of Psychology* **31**, 607–649.
- [2] Coombs, C.H. (1964). *A Theory of Data*, Wiley.
- [3] De Leeuw, J. & Heiser, W.J. (1980). Theory of multidimensional scaling, in *Handbook of Statistics*, Vol. II, P.R. Krishnaiah, ed., North Holland Publishing Company, Amsterdam.
- [4] Guttman, L. (1968). A general nonmetric technique for fitting the smallest coordinate space for a configuration of points, *Psychometrika* **33**, 469–506.
- [5] Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.
- [6] Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: a numerical method, *Psychometrika* **29**, 115–129.
- [7] Roskam, E.E.C.H.I. (1968). *Metric analysis of ordinal data in psychology*, Ph.D. thesis, University of Leiden.
- [8] Shepard, R.N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function (Part I), *Psychometrika* **27**, 125–140.
- [9] Shepard, R.N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function (Part II), *Psychometrika* **27**, 219–246.
- [10] Shepard, R.N. (1972). A taxonomy of some principal types of data and of the multidimensional methods for their analysis, in *Multidimensional Scaling, Volume I, Theory*, R.N. Shepard, A.K. Romney & S.B. Nerlove, eds, Seminar Press, pp. 23–47.

(See also **Proximity Measures; Two-mode Clustering**)

JAN DE LEEUW

Carryover and Sequence Effects

MARY E. PUTT

Volume 1, pp. 197–201

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Carryover and Sequence Effects

‘Carryover’ and ‘sequence’ effects are nuisance parameters (see **Nuisance Variables**) that may arise when repeated measurements are collected on subjects over time. In experimental studies, carryover is a lingering effect of a treatment administered in one period into the subsequent period. Carryover effects are differences in carryover between two treatments, while sequence effects are differences in overall responses between subjects receiving treatments in different orders [6, 11, 14, 16]. Carryover and sequence effects may occur in **observational studies**. For example, the ordering of items in a questionnaire might affect overall response, while the response to an individual item may be influenced by the previous item(s) [15]. Our review focuses on experimental studies, although the principles could extend to observational studies (for a general discussion of repeated measures studies, see [5] or [16]).

To illustrate, first imagine a study where two levels of a single factor (E : Experimental and S : Standard) are of interest and where n subjects, acting as blocks, are each available for two measurement periods. For example, E might be a behavioral intervention to improve short-term memory in patients with Alzheimer’s disease, whereas S refers to standard care. If effects due to sequence, carryover, and period (temporal changes that affect all patients) are not of concern, then each patient could receive E followed by S . With respect to carryover, we specifically assume that the experimental intervention administered in the first period does not influence short-term memory in the second period. In the analysis of this design, carryover from E into S cannot be estimated separately from the treatment effect, that is, carryover is confounded with the treatment effect.

Alternatively, we might randomly assign patients to E followed by S , in sequence 1 (ES), or S followed by E , in sequence 2 (SE). This ‘crossover’ study is a type of split-plot **factorial design** [6, 11]. In split-plot factorials, we have two factors of interest. In the **repeated measures** setting, we randomize n_i ($i = 1, \dots, s$) subjects to receive one of the s levels of the *between-subject* factor; we apply the p levels of the *within-subject* factor to each subject sequentially (Table 1). Thus, each subject

Table 1 Layout for split-plot factorial design in a repeated measures setting

Between-subject factor	Subject	Within-subject factor			
1	1	1	...	p	
⋮	⋮	⋮		⋮	
1	n_1	1	...	p	
⋮	⋮	⋮		⋮	
s	1	1	...	p	
⋮	⋮	⋮		⋮	
s	n_s	1	...	p	

receives only one of the levels of the between-subject factors but all levels of the within-subject factor. In **crossover designs**, sequence effects (ES versus SE) are individual levels of the between-subject factor while the individual treatments (E and S) can be thought of as levels of the within-subject factor (Table 2). In an alternative split-plot design ($EE:SS$), we might randomly assign patients to receive either E or S and then make two weekly measurements. Here, E and S (or EE and SS) are levels of the between-subject factor, while time is the within-subject factor. As we will see, the estimate of the treatment effect in the crossover design is based on within-subject differences. Typically, this design is more efficient and requires far fewer subjects than the $EE:SS$ design where the estimate of the treatment effect is based on a between-subject comparison [13].

As an example, consider a **clinical trial** comparing experimental therapy (nasal corticosteroids, Treatment E) and placebo (Treatment S) on self-reported daytime sleepiness in patients with allergic rhinitis [4]. Table 3 shows the data for nine patients on the $ES:SE$ portion of the crossover study for whom data were available for both periods. The datum for one patient from the ES sequence was missing for the second period. Patients who were randomly assigned to either S or E self-administered their treatments

Table 2 Layout for $ES:SE$ and $EE:SS$ designs

Design	Sequence	Period	
		1	2
$ES:SE$	ES	E	S
$ES:SE$	SE	S	E
$EE:SS$	EE	E	E
$EE:SS$	SS	S	S

2 Carryover and Sequence Effects

Table 3 Mean weekly IDS score by period and mean difference between periods for individual patients on the TP and PT sequences [12]. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 2002 by the American Statistical Association

Sequence 1 (ES)					
Patient	1	2	3	4	5
Period 1	2.00	1.83	0	0.89	3.00
Period 2	1.29	0	0	NA	3.00
Difference	0.36	0.92	0	NA	0

Sequence 2 (SE)					
Patient	1	2	3	4	5
Period 1	4.00	0	0	1.14	0
Period 2	4.00	0.86	0	2.14	2.29
Difference	0	-0.43	0	-0.50	-1.15

over an eight-week period. They then crossed over to the other treatment (without washout) for a second eight-week period of treatment. Each patient rated their improvement in daytime sleepiness (IDS) on a scale of 0 (worst) to 4 (best) on a daily basis. We analyzed average IDS over the final week in each eight-week treatment period and assumed that weekly averaging of the ordinal IDS scores yielded data on a continuous scale.

Statistical Model. Let Y_{ijkl} be the outcome for the j th subject ($j = 1, \dots, n_i$) from the i th sequence ($i = 1, 2$) on the k th treatment ($k = S, E$) in the l th period ($l = 1, 2$). Then

$$Y_{ijkl} = \mu_k + \pi_l + \lambda_{k'_{l-1}} + \delta_i + \varepsilon_{ijkl}, \quad (1)$$

for μ_k , the mean response for the k th treatment; π_l , the mean added effect for the l th period; $\lambda_{k'_{l-1}}$, the mean added effect due to carryover of the k' th treatment administered in the $(l-1)$ th period into the subsequent period ($\lambda_{k'_0} = 0$); and δ_i , the mean added effect for the i th sequence. The ε_{ijkl} is a random error term. Subjects are independent with $E(\varepsilon_{ijkl}) = 0$, $\text{Var}(\varepsilon_{ijkl}) = \sigma^2$, and $\text{Cov}(\varepsilon_{ijkl}, \varepsilon_{ijk'l'}) = \rho\sigma^2$ for $k \neq k'$ and $l \neq l'$, where ρ is the correlation coefficient.

Estimation and Testing. Table 4 gives the expectation of the outcome for each sequence/period combination as well as for Y_{ij}^* , the mean difference of periods for each subject that is,

$$Y_{ij}^* = \frac{1}{2}(Y_{ijk1} - Y_{ijk2}) \quad (2)$$

and $\mu_D = \mu_E - \mu_S$. The variance of each Y_{ij}^* is

$$\text{Var}(Y_{ij}^*) = \frac{1}{2}\sigma^2(1 - \rho) \quad (3)$$

We combine means of the Y_{ij}^* 's for each sequence, to yield estimates of the treatment effect, that is,

$$\hat{\mu}_D = (\bar{Y}_1^* - \bar{Y}_2^*) \quad (4)$$

for $\bar{Y}_i^* = 1/n_i \sum_{j=1}^{n_i} Y_{ij}^*$. In turn, $\hat{\mu}_D$ has expectation

$$E(\hat{\mu}_D) = \mu_D - \frac{1}{2}\lambda_D \quad (5)$$

for $\lambda_D = \lambda_{E1} - \lambda_{S1}$. The expectation of $\hat{\mu}_D$ is unaffected by period or sequence effects. However in the presence of carryover effects ($\lambda_D \neq 0$), $\hat{\mu}_D$ is biased for μ_D . To put these results in the context of the IDS example, suppose that patients receiving *SE* were, on average, sleepier than patients on *ES*, yielding a sequence effect ($\delta_1 < \delta_2$). In addition, suppose that daytime sleepiness was recorded in the morning for Period 1 and in the afternoon for Period 2, causing a period effect ($\pi_1 > \pi_2$). Under our model, neither nuisance parameter biases $\hat{\mu}_D$. The within-subject comparison eliminates sequence effects; the between-subject comparison eliminates period effects. On the other hand, suppose that daytime sleepiness improved on *E* compared to *S*, that measurements were collected in sequential weeks, rather than over an eight-week period, and that some aspect of the effect of *E* lingered into the week subsequent to its administration. Then, in Period 2 of Sequence 1, we would tend to see IDS scores higher than expected for subjects receiving *S* alone. This carryover effect biases $\hat{\mu}_D$; when λ_D and $\hat{\mu}_D$ have the same sign, $\hat{\mu}_D$ underestimates the true treatment effect, μ_D [13].

Table 4 Expectations for the *ES:SE* design by period

Sequence	Period 1	Period 2	Difference (Y_{ij}^*)
1 (ES)	$\mu_E + \pi_1 + \delta_1$	$\mu_S + \pi_2 + \lambda_{E1} + \delta_1$	$\frac{1}{2}\mu_D + \frac{1}{2}(\pi_1 - \pi_2) - \frac{1}{2}\lambda_{E1}$
2 (SE)	$\mu_S + \pi_1 + \delta_2$	$\mu_E + \pi_2 + \lambda_{S1} + \delta_2$	$-\frac{1}{2}\mu_D + \frac{1}{2}(\pi_1 - \pi_2) - \frac{1}{2}\lambda_{S1}$

In most crossover studies, interest lies in estimating μ_D and testing for nonzero treatment effects. If the sample size is large, or the data are normally distributed, a hypothesis test of

$$H_0: \mu_D - \frac{1}{2}\lambda_D = 0 \quad \text{versus} \quad H_1: \mu_D - \frac{1}{2}\lambda_D \neq 0 \quad (6)$$

can be constructed using

$$t^* = \frac{\hat{\mu}_D}{\sqrt{\frac{s^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}. \quad (7)$$

Here s^2 is the pooled sample variance of the Y_{ij}^* 's ($s^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 / (n_1 + n_2 - 2)$), where s_i^2 is the sample variance of the Y_{ij}^* 's for the i th sequence. Under H_0 , t^* has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. A $100(1 - \alpha)\%$ **confidence interval** (CI) for $\mu_D - 1/2\lambda_D$ is simply

$$\hat{\mu}_D \pm t_{n_1+n_2-2, 1-\alpha/2} \sqrt{\frac{s^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

where α is the level of the CI and $t_{n_1+n_2-2, 1-\alpha/2}$ is the $1 - \alpha/2$ th quantile of the $t_{n_1+n_2-2}$ distribution. We show numerical results for the example in Table 5.

For a Type I error rate of 0.05, these results suggest that corticosteroids improve daytime sleepiness, with the interval (0.01, 1.45 units) containing the mean improvement with high confidence.

It is straightforward to see that for the $EE:SS$ design, the treatment effect is estimated from the difference of the mean of the repeated measurements for each subject, and that the remainder of the procedures described for the crossover can be applied. Using the model in Equation (1), the expectation of this estimate is $\mu_D + 1/2(\delta_1 - \delta_2) + 1/2(\lambda_D)$ that is biased by both sequence and carryover effects. If the study is randomized, then δ_1 and δ_2 should be equal. In addition, carryover here represents the effect of repeated administrations of the same treatment, as opposed to the crossover's lingering effect of one treatment into another, an effect that may be more

interpretable in the context of the study. However, in contrast to the crossover study, the mean of the repeated measurements for each subject has variance $\frac{1}{2}\sigma^2(1 + \rho)$, and this yields a larger variance of the estimated treatment effect. As an example, the IDS study included both EE and SS sequences, each with five subjects (data not shown). Here the estimated treatment effect of 0.85 was slightly larger than the estimate using the crossover portion of the study. However, the 95% confidence interval of $(-0.70, 2.40)$ was substantially wider, while the P value for the hypothesis test was 0.24, suggesting substantial loss in precision compared to the $ES:SE$ portion of the study.

Further Considerations. Potential biases generated by carryover or sequence effects can be addressed in either the design or the analysis of the study. As we showed, two different split-plot designs yield different biases, but also different efficiencies. In addition to split-plot factorial designs, designs where carryover and sequence effects should be considered include **randomized block** and randomized block factorial designs [11]. We focused on simple two-treatment, two-period designs. However, additional measurement and/or washout periods may reduce carryover effects and/or improve the efficiency of these designs [1, 3, 13, 16]. We note that baseline measurements can substantially improve the efficiency of the $EE:SS$ design [12]. We can also add additional levels of either the between or the within-subject factor [11]. Each change in design requires modeling assumptions, and particularly assumptions about carryover, that should be evaluated carefully in the context of the study. Moreover, the reader should be aware of arguments against using more elaborate crossover designs. Senn [14, Chapter 10] argues that assumptions used to develop these approaches do not provide a useful approximation to reality, and that the approach may ultimately yield biased results.

For the $ES:SE$ design, we note that λ_D can be estimated from the difference of the sums of the two periods for each individual. Because this estimate is based on a between-subject comparison, it is

Table 5 Expectations for $ES:SE$ daytime sleepiness example by period

Variable	$\hat{\mu}_D$	s_1	s_2	s	t^*	P value ^a	95% CI
Estimate	0.73	0.864	0.941	0.909	2.39	0.048	(0.01, 1.45)

^aTwo-sided test.

4 Carryover and Sequence Effects

less efficient than the estimate of the treatment effect, and the **power** to detect carryover is typically very small [2]. In contrast, λ_D , while typically not of great interest in the $EE:SS$ design, is estimated efficiently from a within-subject comparison. Grizzle's popular two-stage method used a test for the presence of a carryover effect to determine whether to use both or only the first period of data to estimate μ_D in the crossover design [8, 9]. This analysis is fundamentally flawed, with Type I error rates in excess of the nominal Type I error rates in the absence of carryover [7].

Lastly, **analysis of variance** or mixed-effects models (see **Generalized Linear Mixed Models; Linear Multilevel Models**) extend the analyses we have described here, and provide a unified framework for analyzing the repeated measures studies [5, 10, 11, 16].

References

- [1] Balaam, L.N. (1968). A two-period design with t^2 experimental units, *Biometrics* **24**, 61–67.
- [2] Brown, B.W. (1980). The cross-over experiment for clinical trials, *Biometrics* **36**, 69–79.
- [3] Carriere, K.C. & Reinsel, G.C. (1992). Investigation of dual-balanced crossover designs for two treatments, *Biometrics* **48**, 1157–1164.
- [4] Craig, T.J., Teets, S., Lehman, E.G., Chinchilli, V.M. & Zwillich, C. (1998). Nasal congestion secondary to allergic rhinitis as a cause of sleep disturbance and daytime fatigue and the response to topical nasal corticosteroids, *Journal of Allergy and Clinical Immunology* **101**, 633–637.
- [5] Diggle, P.J., Liang, K. & Zeger, S.L. (1994). *Analysis of Longitudinal Data*, Oxford Science Publications, New York.
- [6] Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*, Wiley, New York.
- [7] Freeman, P. (1989). The performance of the two-stage analysis of two-treatment, two-period cross-over trials, *Statistics in Medicine* **8**, 1421–1432.
- [8] Grizzle, J.E. (1965). The two-period change over design and its use in clinical trials, *Biometrics* **21**, 467–480.
- [9] Grizzle, J.E. (1974). Correction to Grizzle (1965), *Biometrics* **30**, 727.
- [10] Jones, B. & Kenward, M. (2003). *Design and Analysis of Crossover Trials*, Chapman & Hall, New York.
- [11] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Pacific Grove, California, Brooks-Cole.
- [12] Putt, M.E. & Chinchilli, V.M. (2000). A robust analysis of crossover designs using multisample generalized L-statistics, *Journal of the American Statistical Association* **95**, 1256–1262.
- [13] Putt, M.E. & Ravina, B. (2002). Randomized placebo-controlled, parallel group versus crossover study designs for the study of dementia in Parkinson's disease, *Controlled Clinical Trials* **23**, 111–126.
- [14] Senn, S. (2002). *Cross-over Trials in Clinical Research*, Wiley, New York.
- [15] Tourangeau, R., Rasinski, K., Bradburn, N. & D'Andrade, R. (1989). Carryover effects in attitude surveys, *Public Opinion Quarterly* **53**, 495–524.
- [16] Vonesh, E.F. & Chinchilli, V.M. (1997). *Linear and Non-linear Models for the Analysis of Repeated Measurements*, Marcel Dekker Inc, New York.

Further Reading

- Frison, L. & Pocock, S.J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for designs, *Statistics in Medicine* **11**, 1685–1704.

MARY E. PUTT

Case Studies

PATRICK ONGHENA

Volume 1, pp. 201–204

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Case Studies

A case study is an intensive and rigorous empirical investigation of a contemporary phenomenon within its real-life context [9, 16, 17, 20]. In behavioral science, the phenomenon usually refers to an individual, but it can also refer to a group of individuals (e.g., a family or a neighborhood), to an organization (e.g., a company or a school), or to a set of procedures (e.g., a public program or a community intervention). The fact that a case study deals with a contemporary phenomenon makes the distinction with a historical study. The contextual nature of case studies implies that there are many more variables of interest than data points, and this represents the major methodological challenge when trying to draw valid conclusions from case studies. As a further complication, case studies are used most often when the researcher has little or no control over the phenomenon [2, 20].

The Case for Case Study Research

Some researchers and methodologists have discounted case studies as a potential source of valid conclusions because they identified the case study with the preexperimental ‘one-shot case study’, using Campbell and Stanley’s terminology (see **Quasi-experimental Designs**) [4]. However, this dismissal is based on a misunderstanding: ‘one-shot case study’ has been a misnomer for a design that only includes one group and one posttest; a design that has little bearing on case studies as such. In a revision of their design classification and terminology, Cook and Campbell [6] emphasized: ‘Certainly the case study as normally practiced should not be demeaned by identification with the one-group posttest-only design’ (p. 96). In his foreword to Yin’s handbook on case study research, Campbell [3] confirms:

‘[This book on case study research] It epitomizes a research method for attempting valid inferences from events outside the laboratory while at the same time retaining the goals of knowledge shared with laboratory science’ (p. ix).

According to Yin [19, 20], one of the most important strategies for drawing valid conclusions from a case study is the reliance on theory, hypotheses, and concepts to guide design and data collection. Reliance

on theory enables the researcher to place the case study in an appropriate research literature, to help define the boundaries of the case and the unit of analysis, and to suggest the relevant variables and data to be collected. To counter the underidentification of the potential theoretical propositions (cf. the above mentioned ‘many more variables of interest than data points’), case study researchers have to bring into action multiple sources of evidence, with data needing to converge in a triangulating fashion. A prototypical example is Campbell’s pattern-matching strategy [2] that has been used successfully to show that the decreasing level of traffic fatalities in Connecticut was not related to the lowering of the speed limit, and which involved linking several pieces of information from the same case to some theoretical proposition. Other strategies to enable valid conclusions from case studies include collecting systematic and objective data, using continuous assessment or observations during an extended period of time, looking at multiple cases to test tentative hypotheses or to arrive at more general statements, and applying formal data-analytic techniques [9, 10].

What is the Case?

A case may be selected or studied for several reasons. It may be that it is a *unique case*, as, for example, with specific injuries or rare disorders. But it may also be that it represents a *critical case* in testing a well-formulated theory or a *revelatory case* when a researcher has the opportunity to observe and analyze a phenomenon previously inaccessible to scientists [20]. Famous case studies were crucial in the development of several psychological phenomena (e.g., Little Hans, Anna O., Little Albert) [9] and series of cases have exerted a tremendous impact on subsequent research and practice (e.g., in sexology [11] and behavior modification [18]). In more recent years, case study research has been adopted successfully and enthusiastically in cognitive neuropsychology, in which the intensive study of individual brain-damaged patients, their impaired performance and double dissociations have provided valid information about, and invaluable insights into, the structure of cognition [5, 14]. For example, Rapp and Caramazza [13] have described an individual who exhibited greater difficulties in speaking nouns than verbs and greater difficulties in writing verbs

than nouns, and this double dissociation of grammatical category by modality within a single individual has been presented as a serious challenge to current neurolinguistic theories.

Cases in All Shapes and Sizes

When we look at the diversity of the case study literature, we will notice that there are various types of case studies, and that there are several possible dimensions to express this multitude. A first distinction might refer to the kind of paradigm the researcher is working in. On the one hand, there is the more quantitative and analytical perspective of, for example, Yin [19, 20]. On the other hand, there is also the more qualitative and ethnographic approach of, for example, Stake [15, 16]. It is important to give both *quantitative* and *qualitative case studies* a place in behavioral science methodology. As Campbell [3] remarked:

It is tragic that major movements in the social sciences are using the term *hermeneutics* to connote *giving up* on the goal of validity and abandoning disputation as to who has got it right. Thus, in addition to the quantitative and quasi-experimental case study approach that Yin teaches, our social science methodological armamentarium also needs a humanistic validity-seeking case study methodology that, while making no use of quantification or tests of significance, would still work on the same questions and share the same goals of knowledge. (italics in original, p. ix–x)

A second distinction might refer to the kind of research problems and questions that are addressed in the case study. In a *descriptive case study*, the focus is on portraying the phenomenon, providing a chronological narrative of events, citing numbers and facts, highlighting specific or unusual events and characteristics, or using ‘thick description’ of lived experiences and situational complexity [7]. An *exploratory case study* may be used as a tryout or act as a pilot to generate hypotheses and propositions that are tested in larger scale surveys or experiments. An *explanatory case study* tackles ‘how’ and ‘why’ questions and can be used in its own right to test causal hypotheses and theories [20].

Yin [20] uses a third distinction referring to the study design, which is based on the number of cases and the number of units of analysis within each case. In *single-case holistic designs*, there is only one case and a single unit of analysis. In *single-case*

embedded designs, there is also only one case, but, in addition, there are multiple subunits of analysis, creating opportunities for more extensive analysis (e.g., a case study of school climate may involve teachers and pupils as subunits of study). *Multiple-case holistic designs* and *multiple-case embedded designs* are the corresponding designs when the same study contains more than one case (e.g., a case study of school climate that uses a multiple-case design implies involving several schools).

Just in Case

As Yin [20] observed: ‘Case study research is remarkably hard, even though case studies have traditionally been considered to be “soft” research. Paradoxically, the “softer” a research strategy, the harder it is to do’ (p. 16). Here are some common pitfalls in case study research (based on the recommendations in [12]):

- *Bad journalism*. Selecting a case out of several available cases because it fits the researcher’s theory or distorting the complete picture by picking out the most sensational features of the case.
- *Anecdotal style*. Reporting an endless series of low-level banal and tedious nonevents that take over from in-depth rigorous analysis.
- *Pomposity*. Deriving or generating profound theories from low-level data, or by wrapping up accounts in high-sounding verbiage.
- *Blandness*. Unquestioningly accepting the respondents’ views, or only including safe uncontroversial issues in the case study, avoiding areas on which people might disagree.

Cases in Point

Instructive applications of case study research and additional references can be found in [1, 8, 15, 19]. Many interesting case studies from clinical psychology and family therapy can be found in ‘*Clinical Case Studies*’, a journal devoted entirely to case studies.

References

- [1] Bassey, M. (1999). *Case Study Research in Educational Settings*, Open University Press, Buckingham.

-
- [2] Campbell, D.T. (1975). Degrees of freedom and the case study, *Comparative Political Studies* **8**, 178–193.
- [3] Campbell, D.T. (2003). Foreword, in *Case Study Research: Design and Methods*, 3rd Edition, R.K. Yin, ed., Sage Publications, London, pp. ix–xi.
- [4] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand MacNally, Chicago.
- [5] Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: the case for single-patient studies, *Brain and Cognition* **5**, 41–66.
- [6] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand MacNally, Chicago.
- [7] Geertz, C. (1973). Thick description: towards an interpretative theory of culture, in *The Interpretation of Cultures*, C. Geertz, ed., Basic Books, New York, pp. 3–30.
- [8] Hitchcock, G. & Hughes, D. (1995). *Research and the Teacher: A Qualitative Introduction to School-Based Research*, 2nd Edition, Routledge, London.
- [9] Kazdin, A.E. (1992). Drawing valid inferences from case studies, in *Methodological Issues and Strategies in Clinical Research*, A.E. Kazdin, ed., American Psychological Association, Washington, pp. 475–490.
- [10] Kratochwill, T.R., Mott, S.E. & Dodson, C.L. (1984). Case study and single-case research in clinical and applied psychology, in *Research Methods in Clinical Psychology*, A.S. Bellack & M. Hersen, eds, Pergamon, New York, pp. 55–99.
- [11] Masters, W.H. & Johnson, V.E. (1970). *Human Sexual Inadequacy*, Little, Brown, Boston.
- [12] Nisbet, J. & Watt, J. (1984). Case study, in *Conducting Small-Scale Investigations in Educational Management*, J. Bell, T. Bush, A. Fox, J. Goodey & S. Goulding, eds, Harper & Row, London, pp. 79–92.
- [13] Rapp, B. & Caramazza, A. (2002). Selective difficulties with spoken nouns and written verbs: a single case study, *Journal of Neurolinguistics* **15**, 373–402.
- [14] Shallice, T. (1979). Case study approach in neuropsychological research, *Journal of Clinical Neuropsychology* **1**, 183–211.
- [15] Stake, R.E. (1995). *The Art of Case Study Research*, Sage Publications, Thousand Oaks.
- [16] Stake, R.E. (2000). Case studies, in *Handbook of Qualitative Research*, 2nd Edition, N.K. Denzin & Y.S. Lincoln, eds, Sage Publications, Thousand Oaks, pp. 435–454.
- [17] Sturman, A. (1997). Case study methods, in *Educational Research, Methodology and Measurement: An International Handbook*, 2nd Edition, J.P. Keeves, ed., Pergamon, Oxford, pp. 61–66.
- [18] Wolpe, J. (1958). *Psychotherapy by Reciprocal Inhibition*, Stanford University Press, Stanford.
- [19] Yin, R.K. (2003a). *Applications of Case Study Design*, 2nd Edition, Sage Publications, London.
- [20] Yin, R.K. (2003b). *Case Study Research: Design and Methods*, 3rd Edition, Sage Publications, London.

(See also **Single-Case Designs**)

PATRICK ONGHENA

Case–Cohort Studies

BRYAN LANGHOLZ

Volume 1, pp. 204–206

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Case–Cohort Studies

The case-cohort design is a method of sampling from an assembled epidemiologic cohort study or clinical trial in which a random sample of the cohort, called the *subcohort*, is used as a comparison group for all cases that occur in the cohort [10]. This design is generally used when such a cohort can be followed for disease outcomes but it is too expensive to collect and process covariate information on all study subjects. Though it may be used in other settings, it is especially advantageous for studies in which covariate information collected at entry to the study is ‘banked’ for the entire cohort but is expensive to retrieve or process and multiple disease stages or outcomes are of interest. In such circumstances, the work of covariate processing for subcohort members can proceed at the beginning of the study. As time passes and cases of disease occur, information for these cases can be processed in batches. Since the subcohort data is prepared early on and is not dependent on the occurrence of cases, statistical analyses can proceed at regular intervals after the processing of the cases. Further, staffing needs are quite predictable. The design was motivated by the case-base sampling method for simple binary outcome data [6, 8]. Parameters of interest in case-cohort studies are usual rate ratios in a Cox proportional hazards model (see **Survival Analysis**) [4].

Design

The basic components of a case-cohort study are the *subcohort*, a sample of subjects in the cohort, and *non-subcohort cases*, subjects that have had an event and are not included in the subcohort. The subcohort provides information on the person-time experience of a random sample of subjects from the cohort or random samples from within strata of a confounding factor. In the latter situation, differing sampling fractions could be used to better align the person-time distribution of the subcohort with that of the cases. Methods for sampling the subcohort include sampling a fixed number without replacement [10] and sampling based on independent Bernoulli ‘coin flips’ [14]. The latter may be advantageous when subjects are entered into the study prospectively;

the subcohort may then be formed concurrently rather than waiting until accrual into the cohort has ended [12, 14]. Simple case-cohort studies are the same as case-base studies for simple binary outcome data. But, in general, portions of a subject’s time on study might be sampled. For example, the subcohort might be ‘refreshed’ by sampling from those remaining on study after a period of time [10, 15]. These subjects would contribute person-time only from that time forward. While the subcohort may be selected on the basis of covariates [3, 10], a key feature of the case-cohort design is that the subcohort is chosen without regard to failure status; methods that rely on failure status in the sampling of the comparison group are **case-control studies**.

Examples

Study of lung cancer mortality in aluminum production workers in Quebec, Canada

Armstrong et al. describe the results of a case-cohort study selected from among 16 297 men who had worked at least one year in manual jobs at a large aluminum production plant between 1950 and 1988 [1]. This study greatly expands on an earlier cohort mortality study of the plant that found a suggestion of increased rates of lung cancer in jobs with high exposures to coal tar pitch [5]. Through a variety of methods, 338 lung cancer deaths were identified. To avoid the expense associated with tracing subjects and abstraction of work records for the entire cohort, a case-cohort study was undertaken. To improve study efficiency, a subcohort of 1138 subjects was randomly sampled from within year of birth strata with sampling fractions varying to yield a similar distribution to that of cases. This was accommodated in the analysis by stratification by these year of birth categories. The random sampling of subcohort members resulted in the inclusion of 205 cases in the subcohort. Work and smoking histories were abstracted for the subcohort and the additional 133 non-subcohort cases. Cumulative exposure to coal tar pitch volatiles were estimated by linking worker job histories to the measurements of chemical levels made in the plant using a ‘job-exposure matrix’. The analyses confirmed the lung cancer–coal pitch association observed in the earlier study and effectively ruled out confounding by smoking.

Genetic influences in childhood asthma development in the Children's Health Study in Los Angeles, California

The Children's Health Study at the University of Southern California has followed a cohort of school-aged children and recorded measures of respiratory health since 1993 [9]. Buccal cell samples have been collected and stored on a large proportion of cohort subjects and it was desired to retrospectively investigate genetic and environmental influences on incident asthma rates in this age group. Because the genotyping lab work is expensive, this study is to be done using a sample of the cohort. One complication of this study is that asthma was not a primary endpoint of the original study and incident asthma occurrence is available only as reported by the subject or the parents. It is believed that about 30% of the self-reported asthma cases will not be confirmed as 'physician diagnosed', the criterion for study cases. Because control selection is not tied to case determination, a case-cohort design allows selection of the comparison group prior to case confirmation. A subcohort will be randomly selected and all of these subjects interviewed and genotyped. Both subcohort and non-subcohort self-reported asthma subjects will be confirmed as to case status. Non-subcohort self-reported asthma subjects who are not confirmed are dropped from the study, while those in the subcohort are simply assigned their confirmed status.

Computer Software

Rate ratios from the Cox model can be computed using any Cox regression software. However, the variance is not properly estimated. A number of methods have been implemented in software packages. The 'Prentice' estimator [10] is a rather complicated expression and only one software package has implemented it (Epicure, Hirosoft International Corp., Seattle, WA, www.hirosoft.com). Simpler alternatives are the 'asymptotic' [11] and 'robust' variance estimators [7, 2]. Either may be computed by simple manipulation of δ and β diagnostic statistics, which are an output option in many software packages [13]. The asymptotic estimator requires the sampling fractions while the robust estimates these from the data.

References

- [1] Armstrong, B., Tremblay, C., Baris, D. & Gilles, T. (1994). Lung cancer mortality and polynuclear aromatic hydrocarbons: a case-cohort study of aluminum production workers in Arvida, Quebec, Canada, *American Journal of Epidemiology* **139**, 250–262.
- [2] Barlow, W.E. (1994). Robust variance estimation for the case-cohort design, *Biometrics* **50**, 1064–1072.
- [3] Borgan, O., Langholz, B., Samuelsen, S.O., Goldstein, L. & Pogoda, J. (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* **6**, 39–58.
- [4] Cox, D.R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society B* **34**, 187–220.
- [5] Gibbs, G.W. (1985). Mortality of aluminum reduction plant workers, 1950 through 1977, *Journal of Occupational Medicine* **27**, 761–770.
- [6] Kupper, L.L., McMichael, A.J. & Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk, *Journal of the American Statistical Association* **70**, 524–528.
- [7] Lin, D.Y. & Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association* **88**, 1341–1349.
- [8] Miettinen, O.S. (1982). Design options in epidemiology research: an update, *Scandinavian Journal of Work, Environment, and Health* **8**(Suppl. 1), 1295–1311.
- [9] Peters, J.M., Avol, E., Navidi, W., London, S.J., Gauderman, W.J., Lurmann, F., Linn, W.S., Margolis, H., Rappaport, E., Gong, H. & Thomas, D.C. (1999). A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity, *American Journal of Respiratory and Critical Care Medicine* **159**(3), 760–767.
- [10] Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials, *Biometrika* **73**, 1–11.
- [11] Self, S.G. & Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies, *Annals of Statistics* **16**, 64–81.
- [12] Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S.L., Clifford, C., Goldman, S., Urban, N., Sheppard, L. & Greenwald, P. (1988). Statistical design of the women's health trial, *Controlled Clinical Trials* **9**, 119–136.
- [13] Therneau, T.M. & Li, H. (1999). Computing the cox model for case cohort designs, *Lifetime Data Analysis* **5**, 99–112.
- [14] Wacholder, S. (1991). Practical considerations in choosing between the case-cohort and nested case-control designs, *Epidemiology* **2**, 155–158.
- [15] Wacholder, S., Gail, M.H. & Pee, D. (1991). Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort, *Biometrics* **47**, 63–76.

BRYAN LANGHOLZ

Case–Control Studies

VANCE W. BERGER AND JIALU ZHANG

Volume 1, pp. 206–207

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Case–Control Studies

Case–Control Study

A case–control study is an observational study, meaning that there is no **randomization** used in the design. The studied subjects make their own choices regarding exposures or treatments, and the investigators observe both the group to which each individual belongs and the corresponding outcome. In contrast, a randomized trial is an experimental study, in which investigators control the group to which a patient belongs by randomly assigning each patient to a treatment group. There are several types of observational studies, such as the case–control study, the cohort study (*see Case–Cohort Studies*), and the cross-sectional study (*see Cross-sectional Design*) [1]. In case–control studies, investigators observe subjects with and without disease, and then look back to assess the antecedent risk factors. In cohort studies, investigators follow subjects with and without a risk factor or exposure, and follow them to determine if they develop the disease.

Cohort studies may not be suitable for studying associations between exposures and diseases that take a long time to develop, such as the association between smoking and lung cancer. Clearly, such a study, if conducted as a cohort study, would require a long time to follow up and a large sample size. If, however, the investigator decides to use a case–control study instead, then the required time between exposure and disease would have already elapsed, and so the time required would be considerably less. If D represents disease (for example, myocardial infarction), E represents exposure (e.g., oral contraception), $\bar{D}$ represents no myocardial infarction, and $\bar{E}$ represents no oral contraception, then for cohort studies the **relative risk (RR)** should be calculated as follows:

$$RR = \frac{P(D|E)}{P(D|\bar{E})}. \quad (1)$$

In a case–control study, the marginal totals for disease (i.e., the total number of subjects with disease, or cases, and the total number of subjects without disease, or controls) is fixed by design (the sampling scheme determines these totals). What is random is the exposure marginal totals (i.e., the total number of

exposed subjects and the total number of unexposed subjects), and the cell counts. It is more common to estimate the relative risk when the exposure margin is fixed by design and the distribution of cases and controls, both overall and within each exposure group, is random. That this is not the case introduces conceptual difficulties with directly estimating the relative risk in case–control studies. One solution is to estimate $P(D|E)$ indirectly, by first estimating $P(E|D)$ directly, and then applying Bayes’ theorem: (*see Bayesian Belief Networks*)

$$P(D|E) = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\bar{D})P(\bar{D})}. \quad (2)$$

The only unknown quantity in this expression is $P(D)$, which is the prevalence of the disease in the whole population. This quantity, $P(D)$, can be estimated from prior knowledge, or perhaps a range of values can be considered to reflect uncertainty. The **odds ratio (OR)** is computed as

$$OR = \frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(\bar{D}|\bar{E})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}, \quad (3)$$

where the notation in the rightmost expression represents the cell counts. From the above expression, we can see that odds ratio can always be computed regardless of whether the study is retrospective or prospective. Also, the odds ratio has the following relationship with the relative risk.

$$OR = RR \frac{1 - P(D|\bar{E})}{1 - P(D|E)}. \quad (4)$$

When the disease is rare, the probability of having the disease should be close to zero (by definition of ‘rare’), regardless of the exposure. Therefore, both $P(D|\bar{E})$ and $P(D|E)$ should be close to zero. In this case, the odds ratio and the relative risk should be very similar numerically. This means that in a case–control study, even if the prevalence of the disease $P(D)$ is unknown, we can still obtain the relative risk by approximating it with the odds ratio if we know that the disease is rare.

Reference

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, Hoboken, New York.

VANCE W. BERGER AND JIALU ZHANG

Catalogue of Parametric Tests

DAVID CLARK-CARTER

Volume 1, pp. 207–227

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Catalogue of Parametric Tests

The classification ‘nonparametric statistical test’ is often used to refer to statistical tests involving parameters when the data is regarded as being nonquantitative (weaker than interval scale data). This catalogue therefore excludes tests based on ranks (*see Rank Based Inference*) but includes chi-square tests, which can be seen as tests involving differences in proportions even though they are often taken to be nonparametric.

One way to look at the definition of a parametric test is to note that a parameter is a measure taken from a population. Thus, the mean and standard deviation of data for a whole population are both parameters. Parametric tests make certain assumptions about the nature of parameters, including the appropriate probability distribution, which can be used to decide whether the result of a statistical test is significant. In addition, they can be used to find a **confidence interval** for a given parameter.

The types of tests given here have been categorized into those that compare means, those that compare variances, those that relate to correlation and regression, and those that deal with frequencies and proportions.

Comparing Means

z-tests

z-tests compare a statistic (or single score) from a sample with the expected value of that statistic in the population under the null hypothesis. The expected value of the statistic under H_0 (see entry on expected value) is subtracted from the observed value of the statistic and the result is divided by its standard deviation. When the sample statistic is based on more than a single score, then the standard deviation for that statistic is its standard error. Thus, this type of test can only be used when the expected value of the statistic and its standard deviation are known. The probability of a *z*-value is found from the standardized normal distribution, which has a mean of 0 and a standard deviation of 1 (*see Catalogue of Probability Density Functions*).

One-group *z*-test for a Single Score. In this version of the test, the equation is

$$z = \frac{x - \mu}{\sigma}, \quad (1)$$

where

x is the single score

μ is the mean for the scores in the population

σ is the standard deviation in the population

Example

Single score = 10

$$\mu = 5$$

$$\sigma = 2$$

$$z = \frac{10 - 5}{2} = 2.5.$$

Critical value for z at $\alpha = 0.05$ with a two-tailed test is 1.96.

Decision: reject H_0 .

One-group *z*-test for a Single Mean. This version of the test is a modification of the previous one because the distribution of means is the standard error of the mean: $\sigma/\sqrt{n}$, where n is the sample size.

The equation for this *z*-test is

$$z = \frac{m - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}, \quad (2)$$

where

m is the mean in the sample

μ is the mean in the population

σ is the standard deviation in the population

n is the sample size

Example

$$m = 5.5$$

$$\mu = 5$$

$$\sigma = 2$$

$$n = 20$$

$$z = \frac{5.5 - 5}{\left(\frac{2}{\sqrt{20}}\right)} = 1.12.$$

The critical value for z with $\alpha = 0.05$ and a one-tailed test is 1.64.

Decision: fail to reject H_0 .

2 Catalogue of Parametric Tests

t Tests

t Tests form a family of tests that derive their probability from Student's *t* distribution (see **Catalogue of Probability Density Functions**). The shape of a particular *t* distribution is a function of the degrees of freedom. *t* Tests differ from *z*-tests in that they estimate the population standard deviation from the standard deviation(s) of the sample(s). Different versions of the *t* Test have different ways in which the degrees of freedom are calculated.

One-group *t* Test. This test is used to compare a mean from a sample with that of a population or hypothesized value from the population, when the standard deviation for the population is not known. The null hypothesis is that the sample is from the (hypothesized) population (that is, $\mu = \mu_h$, where μ is the mean of the population from which the sample came and μ_h is the mean of the population with which it is being compared).

The equation for this version of the *t* Test is

$$t = \frac{m - \mu_h}{\left(\frac{s}{\sqrt{n}}\right)}, \quad (3)$$

where

- m is the mean of the sample
- μ_h is the mean or assumed mean for the population
- s is the standard deviation of the sample
- n is the sample size.

Degrees of freedom:

In this version of the *t* Test, $df = n - 1$.

Example

$$\begin{aligned} m &= 9, \mu_h = 7 \\ s &= 3.2, n = 10 \\ df &= 9 \\ t_{(9)} &= 1.98. \end{aligned}$$

The critical *t* with $df = 9$ at $\alpha = 0.05$ for a two-tailed test is 2.26.

Decision: fail to reject H_0 .

Between-subjects *t* Test. This test is used to compare the means of two different samples. The null hypothesis is $\mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$, where μ_1 and μ_2 are the means of the two populations from which the samples come).

There are two versions of the equation for this test – one when the variances of the two populations are homogeneous, where the variances are pooled, and one when the variances are heterogeneous and the variances are entered separately into the equation.

Homogeneous variance

$$t = \frac{m_1 - m_2}{\sqrt{pv \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad (4)$$

where m_1 and m_2 are the sample means of groups 1 and 2 respectively n_1 and n_2 are the sample sizes of the two groups and pv is the pooled variance

$$pv = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}, \quad (5)$$

where s_1^2 & s_2^2 are the variances of groups 1 and 2. When the two group sizes are the same this simplifies to

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}, \quad (6)$$

where n is the sample size of each group.

Heterogeneous variances

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (7)$$

Degrees of freedom:

Homogeneous variance

$$df = n_1 + n_2 - 2. \quad (8)$$

Heterogeneous variance

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}. \quad (9)$$

Example of groups with homogeneous variance

$$\begin{aligned} m_1 &= 5.3, m_2 = 4.1 \\ n_1 &= n_2 = 20 \\ s_1^2 &= 1.3, s_2^2 = 1.5 \end{aligned}$$

$$\text{df} = 38$$

$$t_{(38)} = 3.21.$$

The critical t for a two-tailed probability with $\text{df} = 38$, at $\alpha = 0.05$ is 2.02.

Decision: reject H_0 .

Within-subjects t Test. This version of the t Test is used to compare two means that have been gained from the same sample, for example, anxiety levels before a treatment and after a treatment, or from two matched samples. For each person, a difference score is found between the two scores, which that person has provided. The null hypothesis is that the mean of the difference scores is 0 ($\mu_d = 0$, where μ_d is the mean of the difference scores in the population).

The equation for this test is

$$t = \frac{m_d}{\left(\frac{s_d}{\sqrt{n}}\right)}, \quad (10)$$

where

m_d is the mean of the difference scores

s_d is the standard deviation of the difference scores

n is the number of difference scores.

Degrees of freedom:

In this version of the t Test, the df are $n - 1$.

Example

$$m_1 = 153.2, m_2 = 145.1, m_d = 8.1$$

$$s_d = 14.6, n = 20$$

$$\text{df} = 19$$

$$t_{(19)} = 2.48.$$

The critical t for $\text{df} = 19$ at $\alpha = 0.05$ for a one-tailed probability is 1.729.

Decision: reject H_0 .

ANOVA

Analysis of variance (ANOVA) allows the comparison of the means of more than two different conditions to be compared at the same time in a single omnibus test. As an example, researchers might wish to look at the relative effectiveness of two treatments for a particular phobia. They could compare a measure of anxiety from people who have received one treatment with those who received the other treatment and with those who received no treatment (the

control group). The null hypothesis, which is tested, is that the means from different ‘treatments’ are the same in the population. Thus, if three treatments are being compared, the null hypothesis would be $\mu_1 = \mu_2 = \mu_3$, where μ denotes the mean for the measure in the population. ANOVA partitions the overall variance in a set of data into different parts.

The statistic that is created from an ANOVA is the F -ratio. It is the ratio of an estimate of the variance between the conditions and an estimate of the variance, which is not explicable in terms of differences between the conditions, that which is due to individual differences (sometimes referred to as *error*).

$$F = \frac{\text{variance between conditions}}{\text{variance due to individual differences}}. \quad (11)$$

If the null hypothesis is true, then these two variance estimates will both be due to individual differences and F will be close to 1. If the values from the different treatments do differ, then F will tend to be larger than 1. The probability of an F -value is found from the F distribution (see **Catalogue of Probability Density Functions**). The value of F , which is statistically significant, depends on the degrees of freedom. In this test, there are two different degrees of freedom that determine the shape of the F distribution: the df for the variance between conditions and the df for the error.

The variance estimates are usually termed the *mean squares*. These are formed by dividing a sum of squared deviations from a mean (usually referred to as the *sum of squares*) by the appropriate degrees of freedom.

The F -ratio is formed in different ways, depending on aspects of the design such as whether it is a between-subjects or a within-subjects design. In addition, the F -value will be calculated on a different basis if the independent variables are fixed or random (see **Fixed and Random Effects**) and whether, in the case of between-subjects designs with unequal sample sizes (unbalanced designs), the weighted or unweighted means are used. The examples given here are for fixed independent variables (other than participants) and weighted means. For variations on the calculations, see [8]. The methods of calculation shown will be ones designed to explicate what the equation is doing rather than the computationally simplest version.

4 Catalogue of Parametric Tests

One-way ANOVA

One-way Between-subjects ANOVA. This version of the test partitions the total variance into two components: between the conditions (between-groups) and within groups (the error).

Sums of squares

The sum of squares between the groups (SS_{bg}) is formed from

$$SS_{bg} = \sum [n_i \times (m_i - m)^2], \quad (12)$$

where

n_i is the sample size in group i

m_i is the mean of group i

m is the overall mean.

The sum of squares within the groups (SS_{wg}) is formed from

$$SS_{wg} = \sum \sum (x_{ij} - m_i)^2, \quad (13)$$

where

x_{ij} is the j th data point in group i

m_i is the mean in group i .

Degrees of freedom

The df for between-groups is one fewer than the number of groups:

$$df_{bg} = k - 1, \quad (14)$$

where k is the number of groups.

The degrees of freedom for within groups is the total sample size minus the number of groups:

$$df_{wg} = N - k, \quad (15)$$

where

N is the total sample size

k is the number of groups

Mean squares

The mean squares (MS) are formed by dividing the sum of squares by the appropriate degrees of freedom:

$$MS_{bg} = \frac{SS_{bg}}{df_{bg}}. \quad (16)$$

$$MS_{wg} = \frac{SS_{wg}}{df_{wg}}. \quad (17)$$

Table 1 The scores and group means of three groups in a between-subjects design

	Group		
	A	B	C
	2	5	7
	3	7	5
	3	5	6
	3	5	4
	1	5	4
	1	4	7
Mean (m_i)	2.167	5.167	5.500

F-ratio

The F -ratio is formed by

$$F = \frac{MS_{bg}}{MS_{wg}}. \quad (18)$$

Example

Three groups each with six participants are compared (Table 1).

Overall mean (m) = 4.278

$$SS_{bg} = \sum [6 \times (m_i - 4.278)^2] = 40.444$$

$$SS_{wg} = 19.167$$

$$df_{bg} = 3 - 1 = 2$$

$$df_{wg} = 18 - 3 = 15$$

$$MS_{bg} = \frac{40.444}{2} = 20.222$$

$$MS_{wg} = \frac{19.167}{15} = 1.278$$

$$F_{(2,15)} = \frac{20.222}{1.278} = 15.826.$$

The critical F -value for $\alpha = 0.05$, with df of 2 and 15 is 3.68.

Decision: Reject H_0 .

One-way Within-subjects ANOVA. Because each participant provides a value for every condition in this design, it is possible to partition the overall variance initially into two parts: that which is between-subjects and that which is within-subjects. The latter can then be further divided, to form the elements necessary for the F -ratio, into between-conditions and that which is subjects within conditions (the residual), with the final one being the error term for the F -ratio. This is a more efficient test because

the between-subjects variance is taken out of the calculation, and so the error variance estimate is smaller than in the equivalent between-subjects ANOVA.

Sums of squares

The total sum of squares (SS_{Total}) is calculated from

$$SS_{\text{Total}} = \sum \sum (x_{ip} - m)^2, \quad (19)$$

where x_{ip} is the score of participant p in condition i . The between-subjects sum of squares (SS_S) is calculated from

$$SS_S = k \times \sum (m_p - m)^2, \quad (20)$$

where

k is the number of conditions
 m_p is the mean for participant p across the conditions
 m is the overall mean.

The within-subjects sum of squares (SS_{ws}) is found from

$$SS_{\text{ws}} = \sum \sum (x_{ip} - m_p)^2, \quad (21)$$

where

x_{ip} is the value provided by participant p in condition i
 m_p is the mean of participant p across all the conditions.

In words, for each participant, find the deviation between that person's score on each condition from that person's mean score. Square the deviations and sum them for that person. Find the sum of the sums.

The between-conditions sum of squares (SS_{bc}) is calculated the same way as the between-groups sum of squares in the between-subjects design, except that because the sample size in each condition will be the same, the multiplication by the sample size can take place after the summation:

$$SS_{\text{bc}} = n \times \sum (m_i - m)^2. \quad (22)$$

The residual sum of squares (SS_{res}) can be found by subtracting SS_{bc} from SS_{ws}

$$SS_{\text{res}} = SS_{\text{ws}} - SS_{\text{bc}}. \quad (23)$$

Degrees of freedom

The degrees of freedom for the total (df_{Total}) are found from

$$df_{\text{Total}} = (n \times k) - 1, \quad (24)$$

where

n is the sample size

k is the number of conditions.

The df for between subjects (df_S) is found from

$$df_S = n - 1, \quad (25)$$

where n is the sample size.

The df for between the conditions (df_{bc}) is found from

$$df_{\text{bc}} = k - 1, \quad (26)$$

where k is the number of conditions.

The df for the residual (df_{res}) is found from

$$df_{\text{res}} = df_{\text{Total}} - (df_{\text{bc}} + df_S). \quad (27)$$

Mean squares

The mean squares for between-conditions (MS_{bc}) is found from

$$MS_{\text{bc}} = \frac{SS_{\text{bc}}}{df_{\text{bc}}}. \quad (28)$$

The mean squares for the residual (MS_{res}) is found from

$$MS_{\text{res}} = \frac{SS_{\text{res}}}{df_{\text{res}}}. \quad (29)$$

F -ratio

The F -ratio is formed from

$$F = \frac{MS_{\text{bc}}}{MS_{\text{res}}}. \quad (30)$$

Example

Five participants provide scores on four different conditions (Table 2).

The overall mean is 10.2.

Sums of squares

$$SS_{\text{Total}} = 47.2$$

$$SS_S = 3.2$$

$$SS_{\text{ws}} = 44$$

$$SS_{\text{bc}} = 5.2$$

$$SS_{\text{res}} = 44 - 5.2 = 38.8$$

6 Catalogue of Parametric Tests

Table 2 The scores of the five participants in four conditions of a within-subjects design with means for conditions and participants

Participant	Condition				Mean
	1	2	3	4	
1	11	9	12	10	10.5
2	10	10	9	9	9.5
3	10	13	11	8	10.5
4	8	10	13	9	10
5	13	9	9	11	10.5
Mean	10.4	10.2	10.8	9.4	

Degrees of freedom

$$df_{\text{Total}} = (5 \times 4) - 1 = 19$$

$$df_{\text{S}} = 5 - 1 = 4$$

$$df_{\text{ws}} = 19 - 4 = 15$$

$$df_{\text{bc}} = 4 - 1 = 3$$

$$df_{\text{res}} = 19 - (3 + 4) = 12$$

Mean squares

$$MS_{\text{bc}} = \frac{5.2}{3} = 1.733, MS_{\text{res}} = \frac{38.8}{12} = 3.233$$

F-ratio

$$F_{(3,12)} = 0.536$$

The critical value of *F* for $\alpha = 0.05$ with df of 3 and 12 is 3.49.

Decision: fail to reject H_0 .

Multiway ANOVA

When there is more than one independent variable (IV), the way in which these variables work together can be investigated to see whether some act as moderators for others. An example of a design with two independent variables would be if researchers wanted to test whether the effects of different types of music (jazz, classical, or pop) on blood pressure might vary, depending on the age of the listeners. The moderating effects of age on the effects of music might be indicated by an interaction between age and music type. In other words, the pattern of the link between blood pressure and music type differed between the two age groups. Therefore, an ANOVA with two independent variables will have three *F*-ratios, each of which is testing a different

null hypothesis. The first will ignore the presence of the second independent variable and test the main effect of the first independent variable, such that if there were two conditions in the first IV, then the null hypothesis would be $\mu_1 = \mu_2$, where μ_1 and μ_2 are the means in the two populations for the first IV. The second *F*-ratio would test the second null hypothesis, which would refer to the main effect of the second IV with the existence of the first being ignored. Thus, if there were two conditions in the second IV, then the second H_0 would be $\mu_a = \mu_b$ where μ_a and μ_b are the means in the population for the second IV. The third *F*-ratio would address the third H_0 , which would relate to the interaction between the two IVs. When each IV has two levels, the null hypothesis would be $\mu_{a1} - \mu_{a2} = \mu_{b1} - \mu_{b2}$, where the μ_{a1} denotes the mean for the combination of the first condition of the first IV and the first condition of the second IV.

Examples are only given here of ANOVAs with two IVs. For more complex designs, there will be higher-order interactions as well. When there are *k* IVs, there will be 2, 3, 4, ..., *k* way interactions, which can be tested. For details of such designs, see [4].

Multiway Between-subjects ANOVA. This version of ANOVA partitions the overall variance into four parts: the main effect of the first IV, the main effect of the second IV, the interaction between the two IVs, and the error term, which is used in all three *F*-ratios.

Sums of squares

The Total sum of squares (SS_{Total}) is calculated from

$$SS_{\text{Total}} = \sum \sum \sum (x_{ijp} - m)^2, \quad (31)$$

where x_{ijp} is the score provided by participant *p* in condition *i* of IV₁ and condition *j* of IV₂. A simpler description is that it is the sum of the squared deviations of each participant's score from the overall mean.

The sum of squares for the first IV (SS_A) is calculated from

$$SS_A = \sum [n_i \times (m_i - m)^2], \quad (32)$$

where

n_i is the sample size in condition *i* of the first IV

m_i is the mean in condition *i* of the first IV

m is the overall mean.

The sum of squares for the second IV (SS_B) is calculated from

$$SS_B = \sum [n_j \times (m_j - m)^2], \quad (33)$$

where

n_j is the sample size in condition j of the second IV

m_j is the mean of condition j of the second IV

m is the overall mean.

The interaction sum of squares (SS_{AB}) can be found via finding the between-cells sum of squares ($SS_{b,cells}$), where a cell refers to the combination of conditions in the two IVs: for example, first condition of IV₁ and the first condition of IV₂. $SS_{b,cells}$ is found from

$$SS_{b,cells} = \sum \sum [n_{ij} \times (m_{ij} - m)^2], \quad (34)$$

where n_{ij} is the sample size in the combination of condition j of IV₂ and condition i of IV₁,

m_{ij} is the mean of the participants in the combination of condition j of IV₂ and condition i of IV₁,

m is the overall mean.

$$SS_{AB} = SS_{b,cells} - (SS_A + SS_B). \quad (35)$$

The sum of squares for the residual (SS_{res}) can be found from

$$SS_{res} = SS_{Total} - (SS_A + SS_B + SS_{AB}). \quad (36)$$

Degrees of freedom

The total degrees of freedom (df_{Total}) are found from

$$df_{Total} = N - 1, \quad (37)$$

where N is the total sample size.

The degrees of freedom for each main effect (for example df_A) are found from

$$df_A = k - 1, \quad (38)$$

where k is the number of conditions in that IV.

The interaction degrees of freedom (df_{AB}) are found from

$$df_{AB} = df_A \times df_B, \quad (39)$$

where df_A and df_B are the degrees of freedom of the two IVs.

The degrees of freedom for the residual (df_{res}) are calculated from

$$df_{res} = df_{Total} - (df_A + df_B + df_{AB}). \quad (40)$$

Mean Squares

Each mean square is found by dividing the sum of squares by the appropriate df. For example, the mean square for the interaction (MS_{AB}) is found from

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}}. \quad (41)$$

F-ratios

Each *F*-ratio is found by dividing a given mean square by the mean square for the residual. For example, the *F*-ratio for the interaction is found from

$$F = \frac{MS_{AB}}{MS_{res}}. \quad (42)$$

Example

Twenty participants are divided equally between the four combinations of two independent variables, each of which has two conditions (Table 3).

Overall mean = 8.2

Sums of squares

$$SS_{Total} = 139.2$$

$$SS_A = 51.2$$

$$SS_B = 5.0$$

$$SS_{b,cells} = 56.4$$

$$SS_{AB} = 56.4 - (51.2 + 5.0) = 0.2$$

$$SS_{res} = 139.2 - (51.2 + 5.0 + 0.2) = 82.8$$

Degrees of freedom

$$df_{Total} = 20 - 1 = 19$$

$$df_A = 2 - 1 = 1$$

$$df_B = 2 - 1 = 1$$

Table 3 The scores and group means of participants in a 2-way, between-subjects design

IV ₁ (A)	1		2	
IV ₂ (B)	1	2	1	2
	9	7	6	4
	11	10	4	4
	10	14	9	10
	10	10	5	9
	7	10	6	9
Means	9.4	10.2	6	7.2

8 Catalogue of Parametric Tests

$$df_{AB} = 1 \times 1 = 1$$

$$df_{res} = 19 - (1 + 1 + 1) = 16$$

Mean squares

$$MS_A = \frac{51.2}{1} = 51.2$$

$$MS_B = \frac{5}{1} = 5$$

$$MS_{AB} = \frac{0.2}{1} = 0.2$$

$$MS_{res} = \frac{82.8}{16} = 5.175$$

F-ratios

$$F_{A(1,16)} = \frac{51.2}{5.175} = 9.89$$

The critical value for *F* at $\alpha = 0.05$ with df of 1 and 16 is 4.49.

Decision: Reject H_0

$$F_{B(1,16)} = \frac{5}{5.175} = 0.97.$$

Decision: Fail to reject H_0

$$F_{AB(1,16)} = \frac{0.2}{5.175} = 0.04.$$

Decision: Fail to reject H_0 .

Multiway Within-subjects ANOVA. As with other within-subjects ANOVAs, this test partitions the variance into that which is between-subjects and that which is within-subjects. The latter is further partitioned in such a way that the variance for IVs, the interaction between IVs, and the error terms are all identified. Unlike the between-subjects equivalent, there is a different error term for each IV and the interaction (*see Repeated Measures Analysis of Variance*).

Sums of squares

The total sum of squares (SS_{Total}) is calculated from

$$SS_{Total} = \sum \sum \sum (x_{ijp} - m)^2, \quad (43)$$

where x_{ijp} is the score provided by participant p in condition i of IV₁ and condition j of IV₂

m is the overall mean.

A simpler description is that it is the sum of the squared deviations of each participant's score in each condition of each IV from the overall mean.

The between-subjects sum of squares (SS_S) is calculated from

$$SS_S = k_1 \times k_2 \times \sum (m_p - m)^2, \quad (44)$$

where

k_1 and k_2 are the number of conditions in each IV
 m_p is the mean for participant p across all the conditions
 m is the overall mean.

The within-subjects sum of squares (SS_{ws}) is calculated from

$$SS_{ws} = \sum \sum \sum (x_{ijp} - m_p)^2, \quad (45)$$

where

x_{ijp} is the value provided by participant p in condition i of IV₁ and condition j of IV₂
 m_p is the mean of participant p across all the conditions.

In words, for each participant, find the deviation between that person's score on each condition from that person's mean score. Square the deviations and sum them for that person. Find the sum of the sums. The sum of squares for the main effect of IV₁ (SS_A) is calculated from

$$SS_A = n \times k_2 \times \sum (m_i - m)^2, \quad (46)$$

where

n is the sample size
 k_2 is the number of conditions in IV₂
 m_i is the mean for condition i of IV₁
 m is the overall mean.

The sum of squares for IV₁ by subjects cell ($SS_{A \text{ by } s \text{ cell}}$) is calculated from

$$SS_{A \text{ by } s \text{ cell}} = k_1 \times \sum (m_{ip} - m)^2, \quad (47)$$

where

k_1 is the number of conditions in IV₁
 m_{ip} is the mean for participant p for condition i of IV₁.

The sum of squares for IV₁ by subjects (SS_{AS}), which is the error term, is calculated from

$$SS_{AS} = SS_{A \text{ by } s \text{ cell}} - (SS_A + SS_S). \quad (48)$$

(SS_B and its error term SS_{BS} are calculated in an analogous fashion.)

The sum of squares for cells ($SS_{b,cells}$) is found from

$$SS_{b,cells} = n \times \sum \sum (m_{ij} - m)^2, \quad (49)$$

where

- n is the sample size
- m_{ij} is the mean for the combination of condition j of IV_1 and condition i of IV_2
- m is the overall mean.

The sum of squares for the interaction between IV_1 and IV_2 (SS_{AB}) is calculated from

$$SS_{AB} = SS_{b,cells} - (SS_A + SS_B). \quad (50)$$

The sum of squares for the error term for the interaction IV_1 by IV_2 by subjects (SS_{ABS}) is found from

$$SS_{ABS} = SS_{Total} - (SS_A + SS_B + SS_{AB} + SS_{AS} + SS_{BS} + SS_S). \quad (51)$$

Degrees of freedom

The total degrees of freedom is found from

$$df_{Total} = n \times k_1 \times k_2 - 1, \quad (52)$$

where n is the sample size

k_1 and k_2 are the number of conditions in IV_1 and IV_2 respectively

$$df_S = n - 1. \quad (53)$$

$$df_A = k_1 - 1. \quad (54)$$

$$df_B = k_2 - 1. \quad (55)$$

$$df_{AB} = df_A \times df_B. \quad (56)$$

$$df_{AS} = df_A \times df_S. \quad (57)$$

$$df_{BS} = df_B \times df_S. \quad (58)$$

$$df_{ABS} = df_{AB} \times df_S. \quad (59)$$

Mean squares

$$MS_A = \frac{SS_A}{df_A}, \quad MS_B = \frac{SS_B}{df_B}, \quad MS_{AB} = \frac{SS_{AB}}{df_{AB}},$$

$$MS_{AS} = \frac{SS_{AS}}{df_{AS}}, \quad MS_{BS} = \frac{SS_{BS}}{df_{BS}}, \quad (60)$$

$$MS_{ABS} = \frac{SS_{ABS}}{df_{ABS}}. \quad (61)$$

Table 4 The scores and group means of participants in a 2-way, within-subjects design

IV ₁ (A)	1		2	
IV ₂ (B)	1	2	1	2
1	12	15	15	16
2	10	12	17	14
3	15	12	12	19
4	12	14	14	19
5	12	11	13	14
Mean	12.2	12.8	14.2	16.4

F -ratios

$$F_A = \frac{MS_A}{MS_{AS}}, \quad F_B = \frac{MS_B}{MS_{BS}}, \quad F_{AB} = \frac{MS_{AB}}{MS_{ABS}}. \quad (62)$$

Example

Five participants provide scores for two different IVs each of which has two conditions (Table 4).

$$SS_{Total} = 115.8, \quad SS_S = 15.3, \quad SS_A = 39.2, \quad SS_{AS} = 5.3, \quad SS_B = 9.8, \quad SS_{BS} = 10.7, \quad SS_{AB} = 3.2, \quad SS_{ABS} = 32.3$$

$$df_S = 5 - 1 = 4, \quad df_A = 2 - 1 = 1, \quad df_{AS} = 1 \times 4 = 4, \quad df_B = 2 - 1 = 1, \quad df_{BS} = 1 \times 4 = 4, \quad df_{AB} = 1 \times 1 = 1, \quad df_{ABS} = 1 \times 4 = 4$$

$$MS_A = \frac{39.2}{1} = 39.2, \quad MS_{AS} = \frac{5.3}{4} = 1.325,$$

$$MS_B = \frac{9.8}{1} = 9.8, \quad MS_{BS} = \frac{10.7}{4}$$

$$= 2.675, \quad MS_{AB} = \frac{3.2}{1}, \quad MS_{ABS} = \frac{32.3}{4}$$

$$= 8.075$$

$$F_{A(1,4)} = \frac{39.2}{1.325} = 29.58.$$

The critical value of F at $\alpha = 0.05$ with df of 1 and 4 is 7.71.

Decision: Reject H_0

$$F_{B(1,4)} = \frac{9.8}{2.675} = 3.66.$$

Decision: Fail to reject H_0

$$F_{AB(1,4)} = \frac{3.2}{8.075} = 0.56.$$

Decision: Fail to reject H_0 .

Multiway Mixed ANOVA. Mixed has a number of meanings within statistics. Here it is being used to

mean designs that contain both within- and between-subjects independent variables. For the description of the analysis and the example data, the first independent variable will be between-subjects and the second within-subjects.

The overall variance can be partitioned into that which is between-subjects and that which is within-subjects. The first part is further subdivided into the variance for IV_1 and its error term (within groups). The second partition is subdivided into the variance for IV_2 , for the interaction between IV_1 and IV_2 and the error term for both (IV_1 by IV_2 by subjects).

Sums of squares

The total sum of squares is given by

$$SS_{\text{Total}} = \sum \sum (x_{jp} - m)^2, \quad (63)$$

where x_{jp} is the score provided by participant p in condition j of IV_2 .

A simpler description is that it is the sum of the squared deviations of each participant's score in each condition of IV_2 from the overall mean.

The between-subjects sum of squares (SS_S) is calculated from

$$SS_S = k_2 \times \sum (m_p - m)^2, \quad (64)$$

where

k_2 is the number of conditions in IV_2

m_p is the mean for participant p across all the conditions in IV_2

m is the overall mean.

The within-subjects sum of squares (SS_{ws}) is calculated from

$$SS_{ws} = \sum \sum (x_{jp} - m_p)^2, \quad (65)$$

where

x_{jp} is the value provided by participant p in condition j of IV_2

m_p is the mean of participant p across all the conditions.

The sum of squares for IV_1 (SS_A) is given by

$$SS_A = k_2 \times \sum [n_i \times (m_i - m)^2], \quad (66)$$

where

k_2 is the number of conditions in IV_2

n_i is the size of the sample in condition i of IV_1
 m_i is the mean for all the scores in condition i of IV_1
 m is the overall mean.

The sum of squares for within groups (SS_{wg}) is given by:

$$SS_{wg} = SS_S - SS_A. \quad (67)$$

The sum of squares for between cells (SS_{bc}) is given by

$$SS_{bc} = \sum \left[n_i \times \sum (m_{ij} - m)^2 \right], \quad (68)$$

where

n_i is the sample size in condition i of IV_1
 m_{ij} is the mean of the combination of condition i of IV_1 and condition j of IV_2
 m is the overall mean.

The sum of squares for the interaction between IV_1 and IV_2 (SS_{AB}) is given by

$$SS_{AB} = SS_{bc} - (SS_A + SS_B). \quad (69)$$

The sum of squares for IV_2 by subjects within groups ($SS_{B.s(gps)}$) is given by

$$SS_{B.s(gps)} = SS_{ws} - (SS_B + SS_{AB}). \quad (70)$$

Degrees of freedom

$df_{\text{Total}} = N \times k_2 - 1$, where N is the sample size,

$df_A = k_1 - 1$, $df_B = k_2 - 1$, $df_{AB} = df_A \times df_B$,

$df_{wg} = \sum (n_i - 1)$, where n_i is the sample in condition i of IV_1 , $df_{B.s(gps)} = df_B \times df_{wg}$. (71)

Mean squares

$$\begin{aligned} MS_A &= \frac{SS_A}{df_A}, MS_{wg} = \frac{SS_{wg}}{df_{wg}}, MS_B = \frac{SS_B}{df_B}, \\ MS_{AB} &= \frac{SS_{AB}}{df_{AB}}, MS_{bg} = \frac{SS_{bg}}{df_{bg}}, MS_{B.s(gps)} \\ &= \frac{SS_{B.s(gps)}}{df_{B.s(gps)}}. \end{aligned} \quad (72)$$

F-ratios

$$F_A = \frac{MS_A}{MS_{wg}}, F_B = \frac{MS_B}{MS_{B.s(gps)}}, F_{AB} = \frac{MS_{AB}}{MS_{B.s(gps)}}. \quad (73)$$

Table 5 The scores of participants in a 2-way, mixed design

Participant	Condition of IV ₁ (A)	Condition of IV ₂ (B)	
		1	2
1	1	11	9
2	1	13	8
3	1	12	11
4	1	10	11
5	1	10	10
6	2	10	12
7	2	11	12
8	2	13	10
9	2	9	13
10	2	10	11

Example

Five participants provided data for condition 1 of IV₁ and five provided data for condition 2 of IV₁. All participants provided data for both conditions of IV₂ (Table 5).

Sum of squares

$$\begin{aligned} SS_S &= 6.2, SS_{ws} = 31, SS_A = 1.8, SS_{wg} = 4.4, \\ SS_{b,cell} &= 9.2, SS_B = 0.2, SS_{AB} = 9.2 - \\ &(1.8 + 0.2) = 7.2, SS_{B.s(gps)} = 31 - (0.2 + \\ &7.2) = 23.6 \end{aligned}$$

Degrees of freedom

$$\begin{aligned} df_A &= 2 - 1 = 1, \quad df_{wg} = (5 - 1) + (5 - 1) = 8, \\ df_B &= 2 - 1 = 1, \quad df_{AB} = 1 \times 1 = 1, \quad df_{B.s(gps)} \\ &= 1 \times 8 = 8 \end{aligned}$$

Mean squares

$$\begin{aligned} MS_A &= 1.8, MS_{wg} = 0.55, MS_B = 0.2, MS_{AB} = 7.2, \\ MS_{B.s(gps)} &= 2.95 \end{aligned}$$

F-ratios

$$F_{A(1,8)} = \frac{1.8}{0.55} = 3.27$$

The critical value for F at $\alpha = 0.05$ with df of 1 and 8 is 5.32.

Decision: Fail to reject H_0

$$F_{B(1,8)} = \frac{0.2}{2.95} = 0.07.$$

Decision: Fail to reject H_0

$$F_{AB(1,8)} = \frac{7.2}{2.95} = 2.44.$$

Decision: Fail to reject H_0 .

Extensions of ANOVA. The analysis of variance can be extended in a number of ways. Effects with several degrees of freedom can be decomposed into component effects using comparisons of treatment means, including trend tests. **Analyses of covariance** can control statistically for the effects of potentially confounding variables. **Multivariate analyses of variance** can simultaneously test effects of the same factors on different dependent variables.

Comparing Variances

F test for Difference Between Variances

Two Independent Variances. This test compares two variances from different samples to see whether they are significantly different. An example of its use could be where we want to see whether a sample of people's scores on one test were more variable than a sample of people's scores on another test.

The equation for the F test is

$$F = \frac{s_1^2}{s_2^2}, \quad (74)$$

where the variance in one sample is divided by the variance in the other sample.

If the research hypothesis is that one particular group will have the larger variance, then that should be treated as group 1 in this equation. As usual, an F -ratio close to 1 would suggest no difference in the variances of the two groups. A large F -ratio would suggest that group 1 has a larger variance than group 2, but it is worth noting that a particularly small F -ratio, and therefore a probability close to 1, would suggest that group 2 has the larger variance.

Degrees of freedom

The degrees of freedom for each variance are 1 fewer than the sample size in that group.

Example

Group 1
Variance: 16
Sample size: 100

Group 2
Variance: 11
Sample size: 150

$$F = \frac{16}{11} = 1.455$$

Degrees of freedom

Group 1 df = 100 - 1 = 99; group 2 df = 150 - 1 = 149

The critical value of F with df of 99 and 149 for $\alpha = 0.05$ is 1.346.

Decision: Reject H_0 .

Two Correlated Variances. This version of the test compares two variances that have been derived from the same group. The F test has to take account of the degree to which the two sets of data are correlated. Although other equations exist, which would produce an equivalent probability, the version here gives a statistic for an F -ratio.

The F test uses the equation:

$$F = \frac{(s_1^2 - s_2^2)^2 \times (n - 2)}{4 \times s_1^2 \times s_2^2 \times (1 - r_{12}^2)}, \quad (75)$$

where

s^2 refers to variance

r_{12} is the correlation between the two variables

n is the sample size.

Degrees of freedom

The test has df 1 and $n - 2$.

Example

Sample size: 50

Variance in variable 1: 50

Variance in variable 2: 35

Correlation between the two variables: 0.7

Error df = 50 - 2 = 48

$$F_{(1,48)} = \frac{(50 - 35)^2 \times (50 - 2)}{4 \times 50 \times 35 \times (1 - (.7)^2)} = 3.025$$

The critical value for F with df = 1 and 48 for $\alpha = 0.05$ is 4.043.

Decision: Fail to reject H_0 .

K Independent Variances. The following procedure was devised by Bartlett [2] to test differences between the variances from more than two independent groups.

The finding of the statistic B (which can be tested with the chi-squared distribution) involves a number of stages. The first stage is to find an estimate of the

overall variance (S^2). This is achieved by multiplying each sample variance by its related df, which is one fewer than the size of that sample, summing the results and dividing that sum by the sum of all the df:

$$S^2 = \frac{\sum [(n_i - 1) \times s_i^2]}{N - k}, \quad (76)$$

where

n_i is the sample for the i th group

s_i^2 is the variance in group i

N is the total sample size

k is the number of groups.

Next, we need to calculate a statistic known as C , using

$$C = 1 + \frac{1}{3 \times (k - 1)} \times \left[\sum \left(\frac{1}{n_i - 1} \right) - \frac{1}{N - k} \right]. \quad (77)$$

We are now in a position to calculate B :

$$B = \frac{2.3026}{C} \times \left[(N - k) \times \log(S^2) - \sum \{ (n_i - 1) \times \log(s_i^2) \} \right], \quad (78)$$

where log means take the logarithm to the base 10.

Degrees of freedom

df = $k - 1$, where k is the number of groups.

(79)

Kanji [5] cautions against using the chi-square distribution when the sample sizes are smaller than 6 and provides a table of critical values for a statistic derived from B when this is the case.

Example

We wish to compare the variances of three groups: 2.62, 3.66, and 2.49, with each group having the same sample size: 10.

$$N = 30, k = 3, N - k = 27, C = 0.994$$

$$S^2 = 2.923, \log(s_1^2) = 0.418, \log(s_2^2) = 0.563,$$

$$\log(s_3^2) = 0.396$$

$$\log(S^2) = 0.466,$$

$$(N - k) \times \log(S^2) = 12.579$$

$$\sum [(n_i - 1) \times \log(s_i^2)] = 12.402$$

$$B = 0.4098$$

$$df = 3 - 1 = 2$$

The critical value of chi-squared for $\alpha = 0.05$ for $df = 2$ is 5.99.

Decision: Fail to reject H_0

Correlations and Regression

t Test for a Single Correlation Coefficient. This test can be used to test the statistical significance of a correlation between two variables, for example, scores on two tests of mathematical ability. It makes the assumption under the null hypothesis that there is no correlation between these variables in the population. Therefore, the null hypothesis is $\rho = 0$, where ρ is the correlation in the population. The equation for this test is

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}, \quad (80)$$

where

r is the correlation in the sample
 n is the sample size.

Degrees of freedom

In this version of the t Test, $df = n - 2$, where n is the sample size.

Example

$$\begin{aligned} r &= 0.4, n = 15 \\ df &= 13 \\ t_{(13)} &= 1.57 \end{aligned}$$

Critical t for a two-tailed test with $\alpha = 0.05$ and $df = 13$ is 2.16.

Decision: fail to reject H_0 .

t Test for Partial Correlation. A partial correlation involves removing the shared variance between two variables, which may be explained by a third (or even more) other variable(s) (*see Partial Correlation Coefficients*). An example would be if the relationship between two abilities were being assessed in a sample of children. If the sample involves a range of ages, then any link between the two abilities may be an artifact of the link between each of them and age. Partial correlation would allow the relationship

between the two abilities to be evaluated with the relationship each has with age controlled for. The equation for a partial correlation is

$$r_{12.3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}, \quad (81)$$

where $r_{12.3}$ is the correlation between variables 1 and 2 with variable 3 partialled out, r_{12} is the Pearson Product Moment correlation coefficient between variables 1 and 2, r_{13} and r_{23} are the correlations between the two main variables and variable 3.

The process can be extended to partial out more than one variable [3]. Sometimes, partial correlations are described as being of a particular order. If one variable is being partialled out, the correlation is of order 1. This leads to the expression zero-order correlation, which is sometimes used to describe a correlation with no variables partialled out.

When testing the statistical significance of a partial correlation the null hypothesis is that there is no relationship between the two variables of interest when others have been partialled out. Thus for a partial correlation of order 1 the null hypothesis is $\rho_{12.3} = 0$.

The t Test for a partial correlation is

$$t = \frac{r \times \sqrt{n-2-k}}{\sqrt{1-r^2}}, \quad (82)$$

where

r is the partial correlation coefficient
 n is the sample size
 k is the number of variables that have been partialled out (the order of the correlation).

Degrees of freedom

For each variable to be partialled out, an extra degree of freedom is lost from the degrees of freedom, which would apply to the t Test for a correlation when no variable is being partialled out. Therefore, when one variable is being partialled out, $df = n - 2 - 1 = n - 3$.

Example

$$\begin{aligned} r_{12} &= 0.5, r_{13} = 0.3, r_{23} = 0.2 \\ n &= 20 \\ r_{12.3} &= 0.47 \end{aligned}$$

$$df = 20 - 2 - 1 = 17$$

$$t_{(17)} = 2.2$$

The critical value of t with $df = 17$ for a two-tailed test with $\alpha = 0.05$ is 2.11.

Decision: reject H_0 .

z-test for Correlation Where ρ Not Equal to 0.

This test compares a correlation in a sample with that in a population. A complication with this version of the test is that when the population value for the correlation (ρ) is not 0, the distribution of the statistic is not symmetrical. **Fisher** devised a transformation for ρ and for the correlation in the sample (r), which allows a z -test to be performed.

The equation for this test is

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{n-3}}}, \quad (83)$$

where

r' is the Fisher's transformation of r , the correlation in the sample

ρ' is the Fisher's transformation of ρ the correlation in the population

n is the sample size.

Example

$$r = 0.6$$

$$\rho = 0.4$$

$$n = 20$$

Fisher's transformation of r (r') = 0.693147

Fisher's transformation of ρ (ρ') = 0.423649

$$z = \frac{0.693147 - 0.423649}{\sqrt{\frac{1}{20-3}}} = 1.11$$

The critical value for z with $\alpha = 0.05$ for a two-tailed test is 1.96.

Decision: Fail to reject H_0 .

z-test for Comparison of Two Independent Correlations. As with the previous example of a z -test, this version requires the correlations to be transformed using Fisher's transformation.

The equation for z is

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}}}, \quad (84)$$

where r'_1 and r'_2 are the Fisher's transformations of the correlations in the two samples

n_1 and n_2 are the sample sizes of the two samples

Example

Sample 1, $n = 30$, $r = 0.7$, Fisher's transformation = 0.867301

Sample 2, $n = 25$, $r = 0.5$, Fisher's transformation = 0.549306

$$z = \frac{0.867301 - 0.549306}{\sqrt{\frac{1}{30-1} + \frac{1}{25-1}}} = 1.15$$

The critical value for z with $\alpha = 0.05$ for a one-tailed test is 1.64.

Decision: Fail to reject H_0 .

z-test for Comparison of Two Nonindependent Correlations.

When the correlations to be compared are not from independent groups, the equations become more complicated. There are two types of correlation that we might want to compare. The first is where the two correlations involve one variable in common, for example, when we wish to see whether a particular variable is more strongly related to one variable than another. The second is where the variables in the two correlations are different. An example of the latter could be where we are interested in seeing whether two variables are more strongly related than two other variables. This version could also be used if we were measuring the same two variables on two different occasions to see whether the strength of their relationship had changed between the occasions. The tests given here come from Steiger [7].

One Variable in Common. Given a correlation matrix of three variables as in Table 6:

where we wish to compare the correlation between variables 1 and 2 (r_{21}) with that between variables 1 and 3 (r_{31}).

Initially, we need to find what is called the determinant of this matrix, usually shown as $|R|$

Table 6 The correlation matrix for three variables showing the symbol for each correlation

Variable	Variable		
	1	2	3
1	1		
2	r_{21}	1	
3	r_{31}	r_{32}	1

where

$$|R| = [1 - (r_{21})^2 - (r_{31})^2 - (r_{32})^2] + (2 \times r_{21} \times r_{31} \times r_{32}). \quad (85)$$

Next, we need the mean of the two correlations that we are comparing ($\bar{r}$)

These values can now be put into a t Test:

$$t = \sqrt{\frac{(n-1) \times (1+r_{32})}{2 \times \left(\frac{n-1}{n-3}\right) \times |R| + \bar{r}^2 \times (1-r_{32})^3}}. \quad (86)$$

where n is the sample size.

Degrees of freedom

$$df = n - 3. \quad (87)$$

Example

Data on three variables are collected from 20 people. The correlation matrix is shown in Table 7:

$$\begin{aligned} |R| &= 0.832 \\ \bar{r} &= 0.136 \\ n &= 20 \\ df &= 20 - 3 = 17 \\ t_{(17)} &= -1.46 \end{aligned}$$

The critical value of t with $\alpha = 0.05$, using a two-tailed test, for $df = 17$ is 2.11.

Decision: Fail to reject H_0 .

Table 7 The correlation matrix of three variables

	1	2	3
1	1		
2	-0.113	1	
3	0.385	-0.139	1

Table 8 The correlation matrix for four variables showing the symbol for each correlation

Variable	Variable			
	1	2	3	4
1	1			
2	r_{21}	1		
3	r_{31}	r_{32}	1	
4	r_{41}	r_{42}	r_{43}	1

No Variables in Common. For this version of the test, we will have a correlation matrix involving four variables (Table 8):

Imagine that we are comparing r_{21} with r_{43} .

We need to find the mean of the two correlations that we are comparing ($\bar{r}$).

Next we find a statistic for the two correlations ($\Psi_{12.34}$) from

$$\begin{aligned} \Psi_{12.34} &= 0.5 \times \{[(r_{31} - (\bar{r} \times r_{32})) \times (r_{42} \times (r_{32} \times \bar{r}))] \\ &\quad + [(r_{41} - (r_{31} \times \bar{r})) \times (r_{32} \times (\bar{r} \times r_{31}))] \\ &\quad + [(r_{31} - (r_{41} \times \bar{r})) \times (r_{42} \times (\bar{r} \times r_{41}))] \\ &\quad + [(r_{41} - (\bar{r} \times r_{42})) \times (r_{32} \times (r_{42} \times \bar{r}))]\}. \end{aligned} \quad (88)$$

Next we need the covariance between the two correlations ($\bar{s}_{12.34}$) from

$$\bar{s}_{12.34} = \frac{\Psi_{12.34}}{(1-\bar{r})^2}. \quad (89)$$

We next need to transform the correlations we wish to compare using Fisher's transformation r'_{21} and r'_{43} . We can now put the values into the following equation for a z -test:

$$z = (r'_{21} - r'_{43}) \times \left[\frac{\sqrt{n-3}}{\sqrt{2 - (2 \times \bar{s}_{12.34})}} \right], \quad (90)$$

where n is the sample size

The probability of this z -value can be found from the standard normal distribution.

Example

Twenty people each provide scores on four variables. The correlations between the variables are shown in Table 9:

Comparing r_{21} with r_{43}

Table 9 The correlation matrix of four variables

	1	2	3	4
1	1			
2	-0.113	1		
3	0.385	-0.139	1	
4	0.008	-0.284	-0.111	1

$$\begin{aligned}\bar{r} &= -0.112 \\ \Psi_{12,34} &= -0.110 \\ \bar{s}_{12,34} &= -0.113 \\ r'_{21} &= -0.113 \\ r'_{43} &= -0.112 \\ n &= 20 \\ z &= -0.004\end{aligned}$$

The critical value of z with $\alpha = 0.05$ for a two-tailed test is 1.96.

Decision: fail to reject H_0 .

t Test for Regression Coefficient. This tests the size of an unstandardized regression coefficient (see **Multiple Linear Regression**). In the case of simple regression, where there is only one predictor variable, the null hypothesis is that the regression in the population is 0; that is, that the variance in the predictor variable does not account for any of the variance in the criterion variable. In **multiple linear regression**, the null hypothesis is that the predictor variable does not account for any variance in the criterion variable, which is not accounted for by the other predictor variables.

The version of the t Test is

$$t = \frac{b}{s.e.}, \quad (91)$$

where b is the unstandardized regression coefficient
 $s.e.$ is the standard error for the regression coefficient
In simple regression, the standard error is found from

$$s.e. = \sqrt{\frac{MS_{\text{res}}}{SS_p}}, \quad (92)$$

where MS_{res} is the mean squares of the residual for the regression

SS_p is the sum of squares for the predictor variable.
For multiple regression, the standard error takes into account the interrelationship between the predictor variable for which the $s.e.$ is being calculated and the other predictor variables in the regression (see [3]).

Degrees of freedom

The degrees of freedom for this version of the t Test are based on p (the number of predictor variables) and n (the sample size): $df = n - p - 1$.

Example

In a simple regression ($p = 1$), $b = 1.3$ and the standard error of the regression coefficient is 0.5

$$n = 30$$

$$df = 28$$

$$t_{(28)} = 2.6$$

The critical value for t with $df = 28$, for a two-tailed test with $\alpha = 0.05$ is 2.048.

Decision: reject H_0 .

In multiple regression, it can be argued that a correction to α should be made to allow for multiple testing. This could be achieved by using a Bonferroni adjustment (see **Multiple Comparison Procedures**). Thus, if there were three predictor variables, α would be divided by 3, and the probability for each of the t Tests would have to be 0.0167 or less to be significant.

F Test for a Single R^2 . This is the equivalent of a one-way, between-subjects ANOVA. In this case, the overall variance within the variable to be predicted (the criterion variable, for example, blood pressure) is separated into two sources: that which can be explained by the relationship between the predictor variable(s) (for example, hours of daily exercise, a measure of psychological stress, and daily salt intake) and the criterion variable, and that which cannot be explained by this relationship, the residual.

The equation for this F test is

$$F = \frac{(N - p - 1) \times R^2}{(1 - R^2) \times p}, \quad (93)$$

where N is the sample size

p is the number of predictor variables

R^2 is the squared multiple correlation coefficient

Degrees of freedom

The regression degrees of freedom are p (the number of predictor variables)

The residual degrees of freedom are $N - p - 1$

Example

Number of predictor variables: 3

Sample size: 336

Regression df : 3

Residual df: 332

$$R^2 = 0.03343$$

$$F_{(3,332)} = \frac{(336 - 3 - 1) \times .03343}{(1 - .03343) \times 3} = 3.83$$

Critical F -value with df of 3 and 332 for $\alpha = 0.05$ is 2.63.

Decision: Reject H_0 .

F Test for Comparison of Two R^2 . This tests whether the addition of predictor variables to an existing regression model adds significantly to the amount of variance which the model explains. If only one variable is being added to an existing model, then the information about whether it adds significantly is already supplied by the t Test for the regression coefficient of the newly added variable.

The equation for this F test is

$$F = \frac{(N - p_1 - 1) \times (R_1^2 - R_2^2)}{(p_1 - p_2) \times (1 - R_1^2)}, \quad (94)$$

where N is the sample size.

The subscript 1 refers to the regression with more predictor variables and subscript 2 the regression with fewer predictor variables.

p is the number of predictor variables.

R^2 is the squared multiple correlation coefficient from a regression.

Degrees of freedom

The regression df = $p_1 - p_2$, while the residual df = $N - p_1 - 1$.

Example

$$R_1^2 = 0.047504$$

Number of predictor variables (p_1): 5

$$R_2^2 = 0.03343$$

Number of predictor variables (p_2): 3

Total sample size 336

df for regression = $5 - 3 = 2$

df for residual = $336 - 5 - 1 = 330$

$$F_{(2,330)} = 2.238$$

The critical value of F with df of 2 and 330 for $\alpha = 0.05$ is 3.023.

Decision: Fail to reject H_0 .

Frequencies and Proportions

Chi-square

Chi-square is a distribution against which the results of a number of statistical tests are compared. The two most frequently used tests that use this distribution are themselves called chi-square tests and are used when the data take the form of frequencies. Both types involve the comparison of frequencies that have been found (observed) to fall into particular categories with the frequencies that could be expected if the null hypothesis were correct. The categories have to be mutually exclusive; that is, a case cannot appear in more than one category.

The first type of test involves comparing the frequencies in each category for a single variable, for example, the number of smokers and nonsmokers in a sample. It has two variants, which have a different way of viewing the null hypothesis-testing process. One version is like the example given above, and might have the null hypothesis that the number of smokers in the population is equal to the number of nonsmokers. However, the null hypothesis does not have to be that the frequencies are equal; it could be that they divide the population into particular proportions, for example, 0.4 of the population are smokers and 0.6 are nonsmokers.

The second way in which this test can be used is to test whether a set of data is distributed in a particular way, for example, that they form a normal distribution. Here, the expected proportions in different intervals are derived from the proportions of a normal curve, with the observed mean and standard deviation, that would lie in each interval and H_0 is that the data are normally distributed.

The second most frequent chi-squared test is for **contingency tables** where two variables are involved, for example, the number of smokers and nonsmokers in two different socioeconomic groups.

All the tests described here are calculated in the following way:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}, \quad (95)$$

where f_o and f_e are the observed and expected frequencies, respectively.

The degrees of freedom in these tests are based on the number of categories and not on the sample size.

One assumption of this test is over the size of the expected frequencies. When the degrees of freedom are 1, the assumption is that all the expected frequencies will be at least 5. When the df is greater than 1, the assumption is that at least 20% of the expected frequencies will be 5.

Yates [9] devised a correction for chi-square when the degrees of freedom are 1 to allow for the fact that the chi-squared distribution is continuous and yet when $df = 1$, there are so few categories that the chi-squared values from such a test will be far from continuous; hence, the Yates test is referred to as a *correction for continuity*. However, it is considered that this variant on the chi-squared test is only appropriate when the marginal totals are fixed, that is, that they have been chosen in advance [6]. In most uses of chi-squared, this would not be true. If we were looking at gender and smoking status, it would make little sense to set, in advance, how many males and females you were going to sample, as well as how many smokers and nonsmokers.

Chi-square corrected for continuity is found from

$$\chi^2_{(1)} = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}, \quad (96)$$

Where $|f_o - f_e|$ means take the absolute value, in other words, if the result is negative, treat it as positive.

One-group Chi-square/Goodness of Fit

Equal Proportions. In this version of the test, the observed frequencies that occur in each category of a single variable are compared with the expected frequencies, which are that the same proportion will occur in each category.

Degrees of freedom

The df are based on the number of categories (k); $df = k - 1$.

Example

A sample of 45 people are placed into three categories, with 25 in category A, 15 in B, and 5 in C. The expected frequencies are calculated by dividing the total sample by the number of categories. Therefore, each category would be expected to have 15 people in it.

$$\chi^2 = 13.33$$

$$df = 2$$

The critical value of the chi-squared distribution with $df = 2$ and $\alpha = 0.05$ is 5.99.

Decision: reject H_0 .

Test of Distribution. This test is another use of the previous test, but the example will show how it is possible to test whether a set of data is distributed according to a particular pattern. The distribution could be uniform, as in the previous example, or nonuniform.

Example

One hundred scores have been obtained with a mean of 1.67 and a standard deviation of 0.51. In order to test whether the distribution of the scores deviates from being normally distributed, the scores can be converted into z -scores by subtracting the mean from each and dividing the result by the standard deviation. The z -scores can be put into ranges. Given the sample size and the need to maintain at least 80% of the expected frequencies at 5 or more, the width of the ranges can be approximately 1/2 a standard deviation except for the two outer ranges, where the expected frequencies get smaller, the further they go from the mean. At the bottom of the range, as the lowest possible score is 0, the equivalent z -score will be -3.27 . At the top end of the range, there is no limit set on the scale.

By referring to standard tables of probabilities for a normal distribution, we can find out what the expected frequency would be within a given range of z -scores. The following table (Table 10) shows the expected and observed frequencies in each range.

$$\chi^2 = 3.56$$

$$df = 8 - 1 = 7$$

Table 10 The expected (under the assumption of normal distribution) and observed frequencies of a sample of 100 values

From z	to z	f_e	f_o
-3.275	-1.500	6.620	9
-1.499	-1.000	9.172	7
-0.999	-0.500	14.964	15
-0.499	0.000	19.111	22
0.001	0.500	19.106	14
0.501	1.000	14.953	15
1.001	1.500	9.161	11
1.501	5.000	6.661	7

The critical value for chi-squared distribution with $df = 7$ at $\alpha = 0.05$ is 14.07.
Decision: Fail to reject H_0 .

Chi-square Contingency Test

This version of the chi-square test investigates the way in which the frequencies in the levels of one variable differ across the other variable. Once again it is for categorical data. An example could be a sample of blind people and a sample of sighted people; both groups are aged over 80 years. Each person is asked whether they go out of their house in a normal day. Therefore, we have the variable visual condition with the levels blind and sighted, and another variable whether the person goes out, with the levels yes and no. The null hypothesis of this test would be that the proportions of sighted and blind people going out would be the same (which is the same as saying that the proportions staying in would be the same in each group). This can be rephrased to say that the two variables are independent of each other: the likelihood of a person going out is not linked to that person's visual condition.

The expected frequencies are based on the marginal probabilities. In this example, that would be the number of blind people, the number of sighted people, the number of the whole sample who go out, and the number of the whole sample who do not go out. Thus, if 25% of the entire sample went out, then the expected frequencies would be based on 25% of each group going out and, therefore, 75% of each not going out.

The degrees of freedom for this version of the test are calculated from the number of rows and columns in the **contingency table**: $df = (r - 1) \times (c - 1)$, where r is the number of rows and c the number of columns in the table.

Example

Two variables A and B each have two levels. Twenty-seven people are in level 1 of variable A and 13 are in level 2 of variable A. Twenty-one people are in level 1 of variable B and 19 are in level 2 (Table 11).

If variables A and B are independent, then the expected frequency for the number who are in the first level of both variables will be based on the fact that 27 out of 40 (or 0.675) were in level 1 of variable A and 21 out of 40 (or 0.525) were in level 1 of variable B. Therefore, the proportion who would be

Table 11 A contingency table showing the cell and marginal frequencies of 40 participants

		Variable A		Total
		1	2	
Variable B	1	12	9	21
	2	15	4	19
Total		27	13	40

expected to be in level 1 of both variables would be $.675 \times .525 = 0.354375$ and the expected frequency would be $0.354375 \times 40 = 14.175$.

$$\chi^2 = 2.16$$

$$df = (2 - 1) \times (2 - 1) = 1$$

The critical value of chi-square with $df = 1$ and $\alpha = 0.05$ is 3.84.

Decision: Fail to reject H_0 .

z-test for Proportions

Comparison between a Sample and a Population Proportion. This test compares the proportion in a sample with that in a population (or that which might be assumed to exist in a population).

The standard error for this test is $\sqrt{\frac{\pi \times (1 - \pi)}{n}}$

where

π is the proportion in the population

n is the sample size

The equation for the z-test is

$$z = \frac{p - \pi}{\sqrt{\frac{\pi \times (1 - \pi)}{n}}}, \quad (97)$$

where p is the proportion in the sample.

Example

In a sample of 25, the proportion to be tested is 0.7 Under the null hypothesis, the proportion in the population is 0.5

$$z = \frac{.7 - .5}{\sqrt{\frac{.5 \times (1 - .5)}{25}}} = 2$$

The critical value for z with $\alpha = 0.05$ for a two-tailed test is 1.96.

Decision: Reject H_0 .

Comparison of Two Independent Samples. Under the null hypothesis that the proportions in the populations from which two samples have been taken are the same, the standard error for this test is

$$\sqrt{\frac{\pi_1 \times (1 - \pi_1)}{n_1} + \frac{\pi_2 \times (1 - \pi_2)}{n_2}}$$

where

π_1 and π_2 are the proportions in each population
 n_1 and n_2 are the sizes of the two samples.

When the population proportion is not known, it is estimated from the sample proportions.

The equation for the z -test is

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}}}, \quad (98)$$

where p_1 and p_2 are the proportions in the two samples.

Example

Sample 1: $n = 30$, $p = 0.7$

Sample 2: $n = 25$, $p = 0.6$

$$z = \frac{.7 - .6}{\sqrt{\frac{.7 \times (1 - .7)}{30} + \frac{.6 \times (1 - .6)}{25}}} = 0.776$$

Critical value for z at $\alpha = 0.05$ with a two-tailed test is 1.96.

Decision: Fail to reject H_0 .

Comparison of Two Correlated Samples. The main use for this test is to judge whether there has been change across two occasions when a measure was taken from a sample. For example, researchers might be interested in whether people's attitudes to banning smoking in public places had changed after seeing a video on the dangers of passive smoking compared with attitudes held before seeing the video. A complication with this version of the test is over the estimate of the standard error. A number of versions exist, which produce slightly different results.

Table 12 The frequencies of 84 participants placed in two categories and noted at two different times

		After		
		A	B	Total
Before	A	15	25	40
	B	35	9	44
Total		50	34	84

However, one version will be presented here from Agresti [1]. This will be followed by a simplification, which is found in a commonly used test.

Given a table (Table 12).

This shows that originally 40 people were in category A and 44 in category B, and on a second occasion, this had changed to 50 being in category A and 34 in category B. This test is only interested in the cells where change has occurred: the 25 who were in A originally but changed to B and the 35 who were in B originally but changed to A. By converting each of these to proportions of the entire sample, $25/84 = 0.297619$ and $35/84 = 0.416667$, we have the two proportions we wish to compare. The standard error for the test is

$$\sqrt{\frac{(p_1 + p_2) - (p_1 - p_2)^2}{n}}$$

where n is the total sample size

The equation for the z -test is

$$z = \frac{p_1 - p_2}{\sqrt{\frac{(p_1 + p_2) - (p_1 - p_2)^2}{n}}}. \quad (99)$$

Example

Using the data in the table above,

$p_1 = 0.297619$

$p_2 = 0.416667$

$n = 84$

$$z = \frac{.297619 - .416667}{\sqrt{\frac{(.297619 + .416667) - (.297619 - .416667)^2}{84}}} = -1.304$$

The critical z with $\alpha = 0.05$ for a two-tailed test is -1.96.

Decision: Fail to reject H_0 .

When the z from this version of the test is squared, this produces the Wald test statistic.

A simplified version of the standard error allows another test to be derived from the resulting z -test: McNemar's test of change.

In this version, the equation for the z -test is

$$z = \frac{p_1 - p_2}{\sqrt{\frac{(p_1 + p_2)}{n}}} \quad (100)$$

Example

Once again using the same data as that in the previous example,

$$z = -1.291.$$

If this z -value is squared, then the statistic is McNemar's test of change, which is often presented in a further simplified version of the calculations, which produces the same result [3].

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, Hoboken.
- [2] Bartlett, M.S. (1937). Some examples of statistical methods of research in agriculture and applied biology, *Supplement to the Journal of the Royal Statistical Society* **4**, 137–170.
- [3] Clark-Carter, D. (2004). *Quantitative Psychological Research: A Student's Handbook*, Psychology Press, Hove.
- [4] Howell, D.C. (2002). *Statistical Methods for Psychology*, 5th Edition, Duxbury Press, Pacific Grove.
- [5] Kanji, G.K. (1993). *100 statistical tests*, Sage Publications, London.
- [6] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Routledge, London.
- [7] Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix, *Psychological Bulletin* **87**, 245–251.
- [8] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Design*, 3rd Edition, McGraw-Hill, New York.
- [9] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Supplement to the Journal of the Royal Statistical Society* **1**, 217–235.

DAVID CLARK-CARTER

Catalogue of Probability Density Functions

BRIAN S. EVERITT

Volume 1, pp. 228–234

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Catalogue of Probability Density Functions

Probability density functions (PDFs) are mathematical formulae that provide information about how the values of random variables are distributed. For a discrete random variable (one taking only particular values within some interval), the formula gives the probability of each value of the variable. For a continuous random variable (one that can take any value within some interval), the formula specifies the probability that the variable falls within a particular range. This is given by the area under the curve defined by the PDF. Here a list of the most commonly encountered PDFs and their most important properties is provided. More comprehensive accounts of density functions can be found in [1, 2].

Probability Density Functions for Discrete Random Variables

Bernoulli

A simple PDF for a random variable, X , that can take only two values, for example, 0 or 1. Tossing a single coin provides a simple example. Explicitly the density is defined as

$$P(X = x_1) = 1 - P(X = x_0) = p, \quad (1)$$

where x_1 and x_0 are the two values that the random variable can take (often labelled as ‘success’ and ‘failure’) and P denotes probability. The single parameter of the Bernoulli density function is the probability of a ‘success,’ p . The expected value of X is p and its variance is $p(1 - p)$.

Binomial

A PDF for a random variable, X , that is the number of ‘successes’ in a series of n independent Bernoulli variables. The number of heads in n tosses of a coin provides a simple practical example. The binomial density function is given by

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n. \quad (2)$$

The probability of the random variable taking a value in some range of values is found by simply summing the PDF over the required range. The expected value of X is np and its variance is $np(1 - p)$. Some examples of binomial density functions are shown in Figure 1. The binomial is important in assigning probabilities to simple chance events such as the probability of getting three or more sixes in ten throws of a fair die, in the development of simple statistical significance tests such as the sign test (see catalogue of statistical tests) and as the error distribution used in **logistic regression**.

Geometric

The PDF of a random variable X that is the number of ‘failures’ in a series of independent Bernoulli variables before the first ‘success’. An example is provided by the number of tails before the first head, when a coin is tossed a number of times. The density function is given by

$$P(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, \dots \quad (3)$$

The geometric density function possesses a *lack of memory property*, by which we mean that in a series of Bernoulli variables the probability of the next n trials being ‘failures’ followed immediately by a ‘success’ remains the same geometric PDF regardless of what the previous trials were. The mean of X is $1/p$ and its variance is $(1 - p)/p^2$. Some examples of geometric density functions are shown in Figure 2.

Negative Binomial

The PDF of a random variable X that is the number of ‘failures’ in a series of independent Bernoulli variables before the k th ‘success.’ For example, the number of tails before the 10th head in a series of coin tosses. The density function is given by

$$P(X = x) = \frac{(k+x-1)!}{(x-1)!k!} p^k (1-p)^x, \quad x = 0, 1, 2, \dots \quad (4)$$

The mean of X is $k(1 - p)/p$ and its variance is $k(1 - p)/p^2$. Some examples of the negative

2 Catalogue of Probability Density Functions

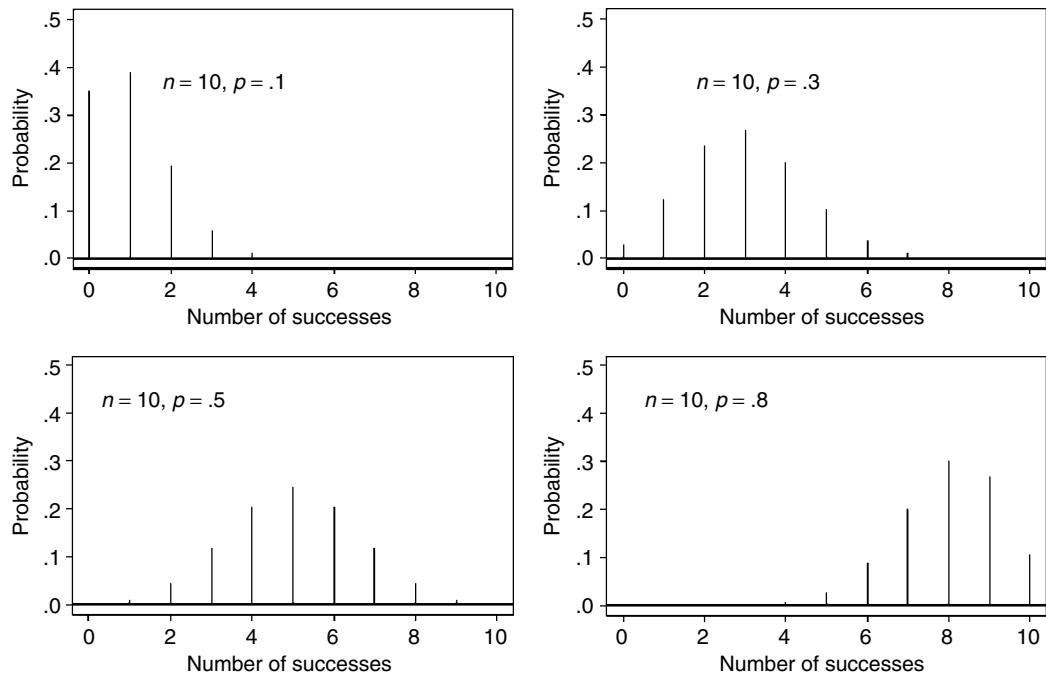


Figure 1 Examples of binomial density functions

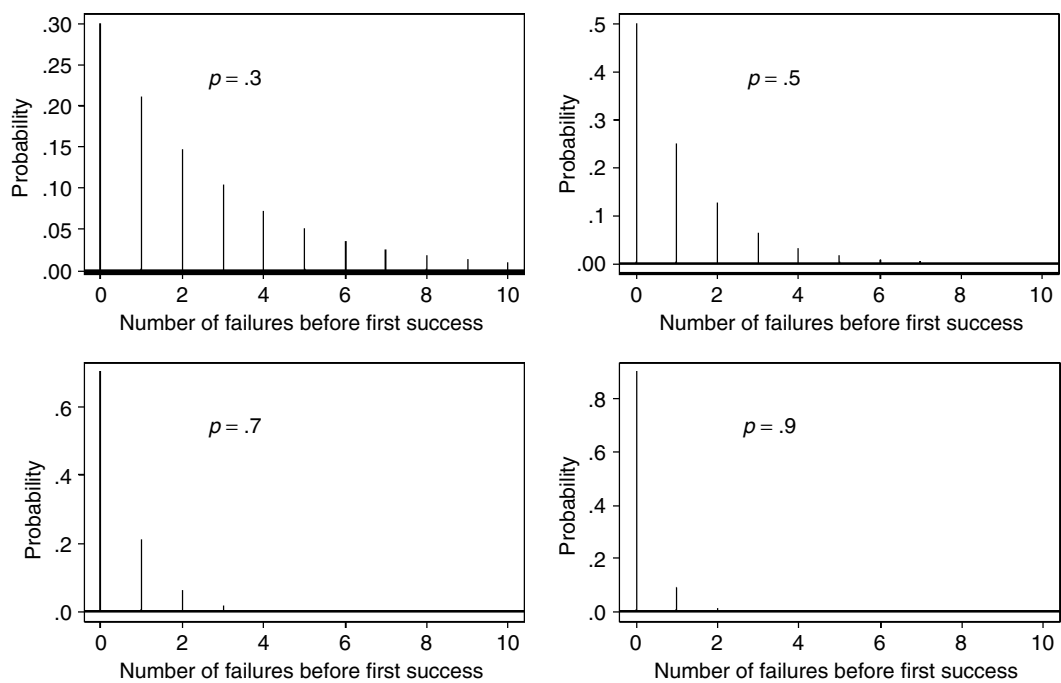


Figure 2 Examples of geometric density functions

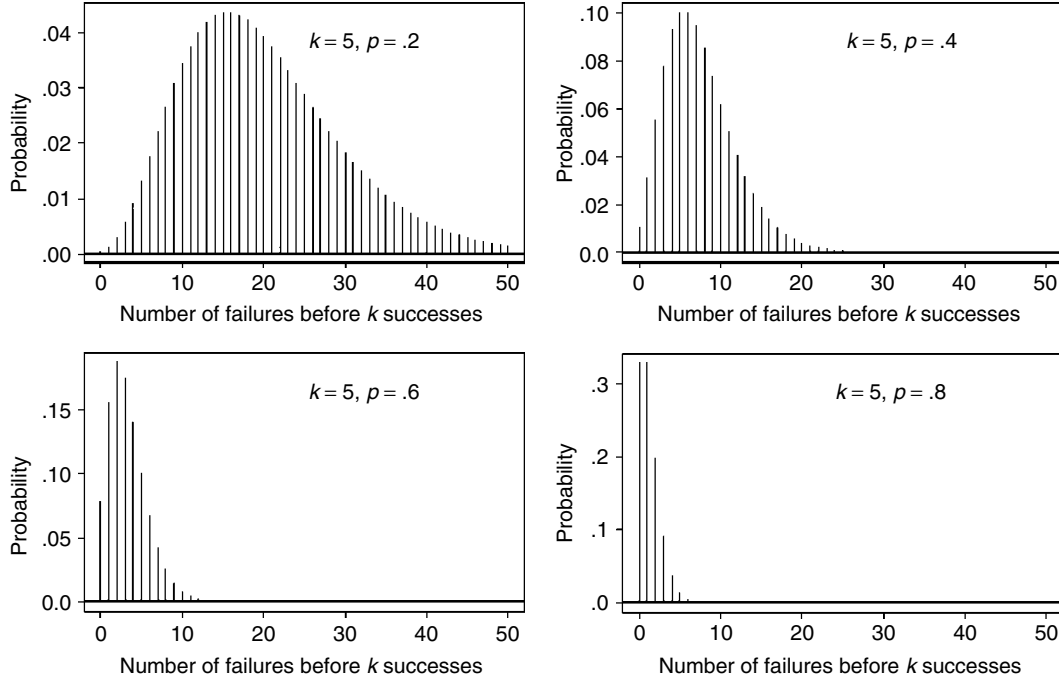


Figure 3 Examples of negative binomial density functions

binomial distribution are given in Figure 3. The density is important in discussions of overdispersion (see **Generalized Linear Models (GLM)**).

Hypergeometric

A PDF associated with *sampling without replacement* from a population of finite size. If the population consists of r elements of one kind and $N-r$ of another, then the hypergeometric is the PDF of the random variable X defined as the number of elements of the first kind when a random sample of n is drawn. The density function is given by

$$P(X = x) = \frac{\frac{r!(N-r)!}{x!(r-x)!(n-x)!(N-r-n+x)!}}{\frac{N!}{n!(N-n)!}}. \quad (5)$$

The mean of X is nr/N and its variance is $(nr/N)(1-r/n)[(N-n)/(N-1)]$. The hypergeometric density function is the basis of Fisher's

exact test used in the analysis of sparse contingency tables (see **Exact Methods for Categorical Data**).

Poisson

A PDF that arises naturally in many instances, particularly as a probability model for the occurrence of rare events, for example, the emission of radioactive particles. In addition, the Poisson density function is the limiting distribution of the binomial when p is small and n is large. The Poisson density function for a random variable X taking integer values from zero to infinity is defined as

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots, \infty. \quad (6)$$

The single parameter of the Poisson density function, λ , is both the expected value and variance, that is, the mean and variance of a Poisson random variable are equal. Some examples of Poisson density functions are given in Figure 4.

4 Catalogue of Probability Density Functions

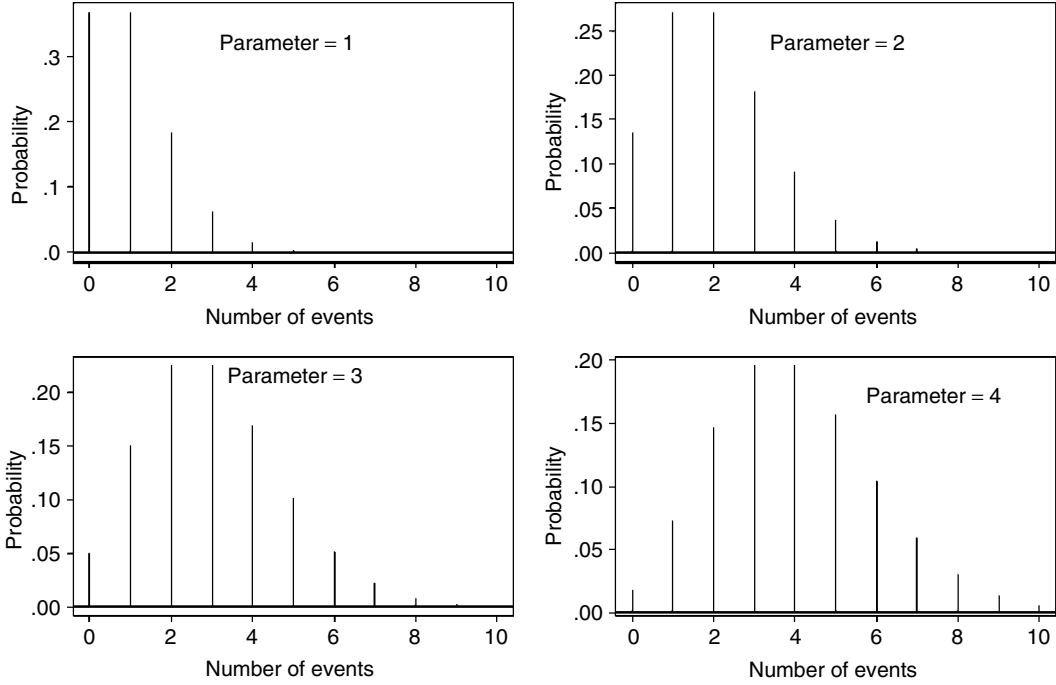


Figure 4 Examples of Poisson density functions

Probability Density Functions for Continuous Random Variables

Normal

The PDF, $f(x)$, of a continuous random variable X defined as follows

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right], \quad -\infty < x < \infty, \quad (7)$$

where μ and σ^2 are, respectively the mean and variance of X . When the mean is zero and the variance one, the resulting density is labelled as the *standard normal*. The normal density function is bell-shaped as is seen in Figure 5, where a number of normal densities are shown.

In the case of continuous random variables, the probability that the random variable takes a particular value is strictly zero; nonzero probabilities can only be assigned to some range of values of the variable. So, for example, we say that $f(x)dx$ gives the probability of X falling in the very small interval, dx , centered on x , and the probability that X falls

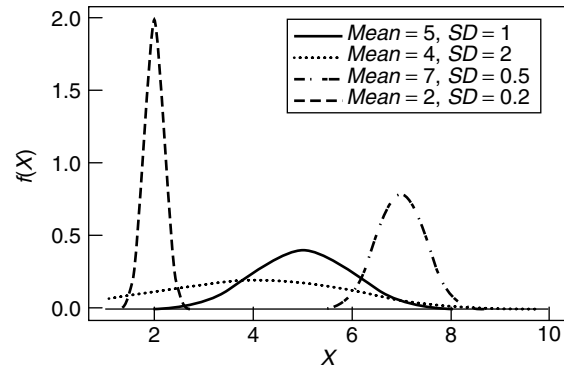


Figure 5 Normal density functions

in some interval, $[A, B]$ say, is given by integrating $f(x)dx$ from A to B .

The normal density function is ubiquitous in statistics. The vast majority of statistical methods are based on the assumption of a normal density for the observed data or for the error terms in models for the data. In part this can be justified by an appeal to the **central limit theorem**. The density function first

appeared in the papers of de Moivre at the beginning of the eighteenth century and some decades later was given by **Gauss** and **Laplace** in the theory of errors and the least squares method. For this reason, the normal is also often referred to as the *Gaussian* or *Gaussian-Laplace*.

Uniform

The mean of such a random variable is having constant probability over an interval. The density function is given by

$$f(x) = \frac{1}{\beta - \alpha}, \quad \alpha < x < \beta. \quad (8)$$

The mean of the density function is $(\alpha + \beta)/2$ and the variance is $(\beta - \alpha)^2/12$. The most commonly encountered version of this density function is one in which the parameters α and β take the values 0 and 1 respectively and is used in generating quasi-random numbers.

Exponential

The PDF of a continuous random variable, X , taking only positive values. The density function is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0. \quad (9)$$

The single parameter of the exponential density function, λ , determines the shape of the density function, as we see in Figure 6, where a number of different exponential density functions are shown. The mean of an exponential variable is $1/\lambda$ and the variance is $1/\lambda^2$.

The exponential density plays an important role in some aspects of **survival analysis** and also gives the distribution of the time intervals between independent consecutive random events such as particles emitted by radioactive material.

Beta

A PDF for a continuous random variable X taking values between zero and one. The density function is defined as

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1} (1-x)^{q-1}, \quad 0 < x < 1, \quad (10)$$

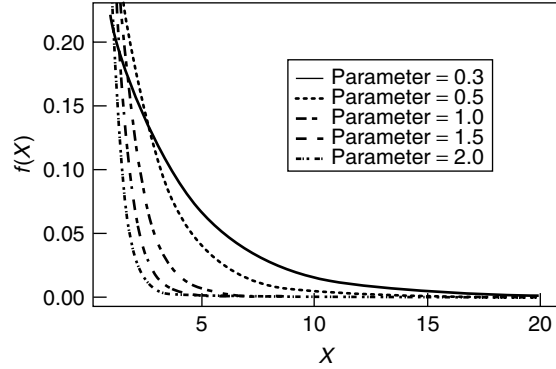


Figure 6 Exponential density functions

where $p > 0$ and $q > 0$ are parameters that define particular aspects of the shape of the beta density and Γ is the gamma function (see [3]). The mean of a beta variable is $p/(p+q)$ and the variance is $pq/(p+q)^2(p+q+1)$.

Gamma

A PDF for a random variable X that can take only positive values. The density function is defined as

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \quad x > 0, \quad (11)$$

where the single parameter, α , is both the mean of X and also its variance.

The gamma density function can often act as a useful model for a variable that cannot reasonably be assumed to have a normal density function because of its positive **skewness**.

Chi-Squared

The PDF of the sum of squares of a number (ν) of independent standard normal variables given by

$$f(x) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2}, \quad x > 0, \quad (12)$$

that is, a gamma density with $\alpha = \nu/2$. The parameter ν is usually known as the *degrees of freedom* of the density. This density function arises in many areas of statistics, most commonly as the null distribution of the chi-squared goodness of fit statistic, for example in the analysis of **contingency tables**.

6 Catalogue of Probability Density Functions

Student's T

The PDF of a variable defined as

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}, \quad (13)$$

where $\bar{X}$ is the arithmetic mean of n observations from a normal density with mean μ and s is the sample standard deviation. The density function is given by

$$f(t) = \frac{\Gamma\left\{\frac{1}{2}(\nu+1)\right\}}{(\nu\pi)^{1/2}\Gamma\left(\frac{1}{2}\nu\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}, \quad -\infty < t < \infty, \quad (14)$$

where $\nu = n - 1$. The shape of the density function varies with ν , and as ν gets larger the shape of the t -density function approaches that of a standard normal. This density function is the null distribution of Student's t -statistic used for testing hypotheses about population means (see **Catalogue of Parametric Tests**).

Fisher's F

The PDF of the ratio of two independent random variables each having a chi-squared distribution, divided by their respective degrees of freedom. The form of the density is that of a beta density with $p = \nu_1/2$ and $q = \nu_2/2$, where ν_1 and ν_2 are, respectively, the degrees of freedom of the numerator and denominator chi-squared variables. Fisher's F density is used to assess the equality of variances in, for example, **analysis of variance**.

Multivariate Density Functions

Multivariate density functions play the same role for *vector random variables* as the density functions described earlier play in the univariate situation. Here we shall look at two such density functions, the multinomial and the multivariate normal.

Multinomial

The multinomial PDF is a multivariate generalization of the binomial density function described earlier.

The density function is associated with a vector of random variables $\mathbf{X}' = [X_1, X_2, \dots, X_k]$ which arises from a sequence of n independent and identical trials each of which can result in one of k possible mutually exclusive and collectively exhaustive events, with probabilities $p_1, p_2, \dots, p_k$. The density function is defined as follows:

$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1!x_2!\dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}, \quad (15)$$

where x_i is the number of trials with outcome i . The expected value of X_i is np_i and its variance is $np_i(1 - p_i)$. The covariance of X_i and X_j is $-np_i p_j$.

Multivariate Normal

The PDF of a vector of continuous random variables, $\mathbf{X}' = [X_1, X_2, \dots, X_p]$ defined as follows

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) = (2\pi)^{-p/2} |\Sigma| \times \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right], \quad (16)$$

where $\boldsymbol{\mu}$ is the mean vector of the variables and Σ is their covariance matrix. The multivariate density function is important in several areas of **multivariate analysis** for example, **multivariate analysis of variance**.

The simplest version of the multivariate normal density function is that for two random variables and known as the *bivariate normal density function*. The density function formula given above now reduces to

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{1-\rho^2} \times \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}\right]\right\}, \quad (17)$$

where $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ are, respectively, the means, standard deviations, and correlation of the two variables. Perspective plots of such density functions are shown in Figure 7.

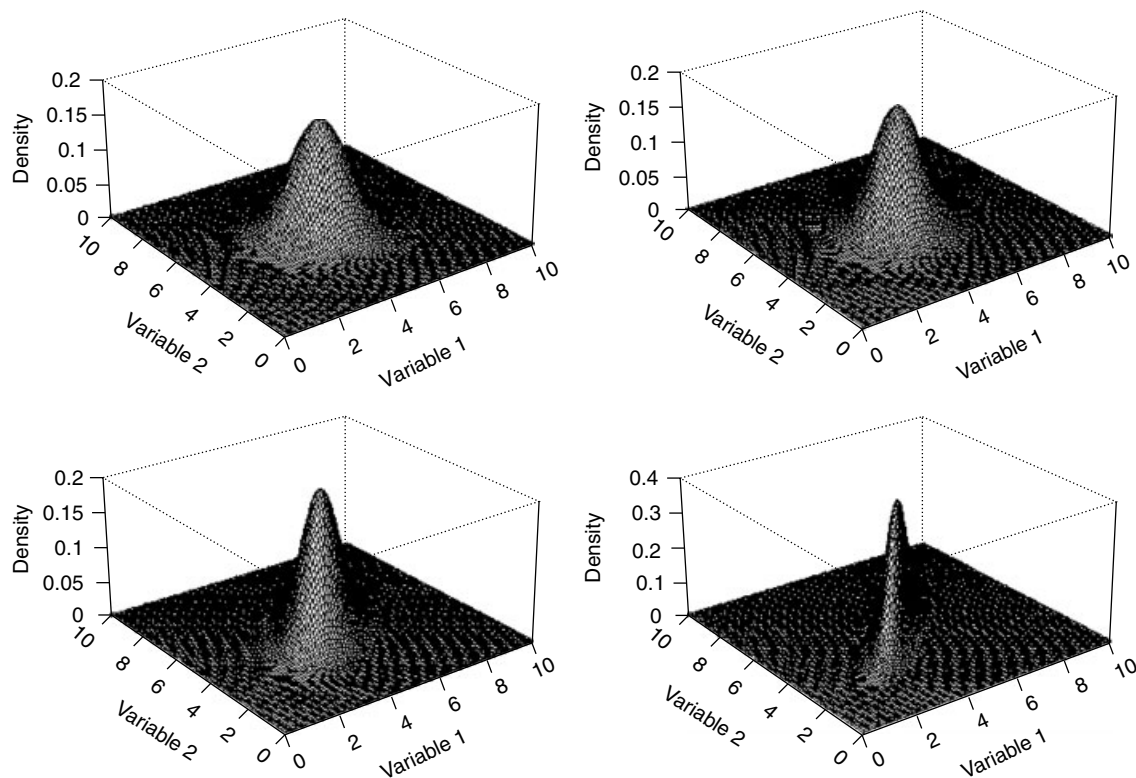


Figure 7 Four bivariate normal density functions with means 5 and standard deviations 1 for variables 1 and 2 and with correlations equal to 0.0, 0.3, 0.6, and 0.9

References

- [1] Balakrishnan, N. & Nevzorov, V.B. (2003). *A Primer on Statistical Distributions*, Wiley, Hoboken.
- [2] Evans, M., Hastings, N. & Peacock, B. (2000). *Statistical Distributions*, 3rd Edition, Wiley, New York.
- [3] Everitt, B.S. (2002). *Cambridge Dictionary of Statistics*, 2nd Edition, Cambridge University Press, Cambridge.

BRIAN S. EVERITT

Catastrophe Theory

E.-J. WAGENMAKERS, H.L.J. VAN DER MAAS AND P.C.M. MOLENAAR

Volume 1, pp. 234–239

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Catastrophe Theory

Introduction

Catastrophe theory describes how small, continuous changes in control parameters (i.e., independent variables that influence the state of a system) can have sudden, discontinuous effects on dependent variables. Such discontinuous, jumplike changes are called *phase-transitions* or *catastrophes*. Examples include the sudden collapse of a bridge under slowly mounting pressure, and the freezing of water when temperature is gradually decreased. Catastrophe theory was developed and popularized in the early 1970s [27, 35]. After a period of criticism [34] catastrophe theory is now well established and widely applied, for instance, in the field of physics, (e.g., [1, 26]), chemistry (e.g., [32]), biology (e.g., [28, 31]), and in the social sciences (e.g., [14]).

In psychology, catastrophe theory has been applied to multistable perception [24], transitions between Piagetian stages of child development [30], the perception of apparent motion [21], sudden transitions in attitudes [29], and motor learning [19, 33], to name just a few. Before proceeding to describe the statistical method required to fit the most popular catastrophe model – the cusp model – we first outline the core principles of catastrophe theory (for details see [9, 22]).

Catastrophe Theory

A key idea in catastrophe theory is that the system under study is driven toward an equilibrium state. This is best illustrated by imagining the movement of a ball on a curved one-dimensional surface, as in Figure 1. The ball represents the state of the system, whereas gravity represents the driving force.

Figure 1, middle panel, displays three possible equilibria. Two of these states are stable states (i.e., the valleys or minima) – when perturbed, the behavior of the system will remain relatively unaffected. One state is unstable (i.e., a hill or maximum) – only a small perturbation is needed to drive the system toward a different state.

Systems that are driven toward equilibrium values, such as the little ball in Figure 1, may be classified

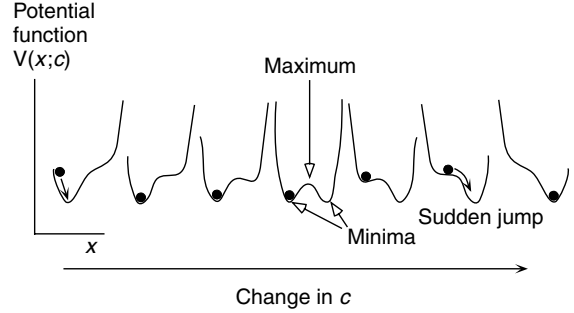


Figure 1 Smooth changes in a potential function may lead to a sudden jump. $V(x; c)$ is the potential function, and c denotes the set of control variables

according to their configuration of critical points, that is, points at which the first or possibly second derivative equals zero. When the configuration of critical points changes, so does the qualitative behavior of the system. For instance, Figure 1 demonstrates how the local minimum (i.e., a critical point) that contains the little ball suddenly disappears as a result of a gradual change in the surface. As a result of this gradual change, the ball will suddenly move from its old position to a new minimum. These ideas may be quantified by postulating that the state of the system, x , will change over time t according to

$$\frac{dx}{dt} = \frac{-dV(x; c)}{dx}, \quad (1)$$

where $V(x; c)$ is the potential function that incorporates the control variables c that affect the state of the system. $V(x; c)$ yields a scalar for each state x and vector of control variables c . The concept of a potential function is very general, – for instance, a potential function that is quadratic in x will yield the ubiquitous normal distribution. A system whose dynamics obey (1) is said to be a gradient dynamical system. When the right-hand side of (1) equals zero, the system is in equilibrium.

As the behavior of catastrophe models can become extremely complex when the number of behavioral and control parameters is increased, we will focus here on the simplest and most often applied catastrophe model that shows discontinuous behavior: the cusp model. The cusp model consists of one *behavioral* variable and only two *control* variables. This may seem like a small number, especially since there are probably numerous variables that exert some kind

2 Catastrophe Theory

of influence on a real-life system, – however, very few of these are likely to qualitatively affect transitional behavior. As will be apparent soon, two control variables already allow for the prediction of quite intricate transitional behavior. The potential function that goes with the cusp model is $V(x; c) = (-1/4)x^4 + (1/2)bx^2 + ax$, where a and b are the control variables. Figure 2 summarizes the behavior of the cusp model by showing, for all values of the control variables, those values of the behavioral variable for which the system is at equilibrium (note that the Figure 2 variable names refer to the data example that will be discussed later). That is, Figure 2 shows the states for which the derivative of the potential function is zero (i.e., $V'(x; c) = -x^3 + bx + a = 0$). Note that one entire panel from Figure 1 is associated with only one (i.e., a minimum), or three (i.e., two minima and one maximum) points on the cusp surface in Figure 2.

We now discuss some of the defining characteristics of the cusp model in terms of a model for attitudinal change [7, 29, 35]. More specifically, we will measure attitude as regards political preference, ranging from left-wing to right-wing. Two control

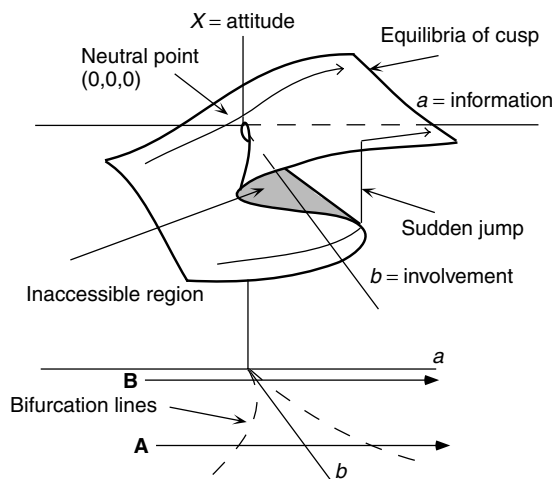


Figure 2 The cusp catastrophe model for attitude change. Of the two control variables, ‘information’ is the *normal* variable, and ‘involvement’ is the *splitting* variable. The behavioral variable is ‘attitude’. The lower panel is a projection of the ‘bifurcation’ area onto the control parameter plane. The bifurcation set consists of those values for ‘information’ and ‘involvement’ combined that allow for more than one attitude. See text for details. Adapted from van der Maas et al. (29)

variables that are important for attitudinal change are *involvement* and *information*. The most distinguishing behavior of the cusp model takes places in the foreground of Figure 2, for the highest levels of involvement. Assume that the lower sheet of the cusp surface corresponds to equilibrium states of being left-wing. As ‘information’ (e.g., experience or environmental effects) more and more favors a right-wing view, not much change will be apparent at first, but at some level of information, a sudden jump to the upper, ‘right-wing’ sheet occurs. When subsequent information becomes available that favors the left-wing view, the system eventually jumps back from the upper sheet unto the lower sheet – but note that this jump does not occur at the same position! The system needs additional impetus to change from one state to the other, and this phenomenon is called *hysteresis*.

Figure 2 also shows that a gradual change of political attitude is possible, but only for low levels of involvement (i.e., in the background of the cusp surface). Now assume one’s political attitude starts out at the neutral point in the middle of the cusp surface, and involvement is increased. According to the cusp model, an increase in involvement will lead to polarization, as one has to move either to the upper sheet or to the lower sheet (i.e., *divergence*), because for high levels of involvement, the intermediate position is *inaccessible*. Hysteresis, divergence, and inaccessibility are three of eight *catastrophe flags* [9], that is, qualitative properties of catastrophe models. Consequently, one method of investigation is to look for the catastrophe flags (i.e., catastrophe detection).

A major challenge in the search of an adequate cusp model is the definition of the control variables. In the cusp model, the variable that causes divergence is called the *splitting variable* (i.e., involvement), and the variable that causes hysteresis is called the *normal variable* (i.e., information). When the normal and splitting variable are correctly identified, and the underlying system dynamics are given by catastrophe theory, this often provides surprisingly elegant insights that cannot be obtained from simple linear models. In the following, we will ignore both the creative aspects of defining appropriate control variables and the qualitative testing of the cusp model using catastrophe flags [30]. Instead, we will focus on the problem of statistically fitting a catastrophe model to empirical data.

Fitting the Cusp Catastrophe Model to Data

Several cusp fitting procedures have been proposed, but none is completely satisfactory (for an overview see [12, 29]). The most important obstacle is that the cusp equilibrium surface is cubic in the dependent variable. This means that for control variables located in the bifurcation area (cf. Figure 2, bottom panel), two values of the dependent variable are plausible (i.e., left-wing/lower sheet and right-wing/upper sheet), whereas one value, corresponding to the unstable intermediate state, is definitely not plausible. Thus, it is important to distinguish between minima of the potential function (i.e., stable states) and maxima of the potential function (i.e., unstable states).

Two methods for fitting the cusp catastrophe models, namely *GEMCAT I and II* [17, 20] and Guastello's polynomial regression technique (see **Polynomial Model**) [10, 11] both suffer from the fact that they consider as the starting point for statistical fitting only those values for the derivative of the potential function that equal zero. The equation $dx/dt = -dV(x; c)/dx = -x^3 + bx + a = 0$ is, however, valid both for minima and maxima – hence, neither GEMCAT nor the polynomial regression technique are able to distinguish between stable equilibria (i.e., minima) and unstable equilibria (i.e., maxima). Obviously, the distinction between stable and unstable states is very important when fitting the cusp model, and neglecting this distinction renders the above methods suspect (for a more detailed critique on the GEMCAT and polynomial regression techniques see [2, 29]).

The most principled method for fitting catastrophe models, and the one under discussion here, is the maximum likelihood method developed by Cobb and coworkers [3, 4, 6]. First, Cobb proposed to make catastrophe theory stochastic by adding a Gaussian white noise driving term $dW(t)$ with standard deviation $D(x)$ to the potential function, leading to

$$dx = \frac{-dV(x; c)}{dx} dt + D(x) dW(t). \quad (2)$$

Equation (2) is a stochastic differential equation (SDE), in which the deterministic term on the right-hand side, $-dV(x; c)/dx$, is called the (*instantaneous*) *drift function*, while $D^2(x)$ is called the (*instantaneous*) *diffusion function*, and $W(t)$ is a

Wiener process (i.e., idealized Brownian motion). The function $D^2(x)$ is the infinitesimal variance function and determines the relative influence of the noise process (for details on SDE's see [8, 15]).

Under the assumption of additive noise (i.e., $D(x)$ is a constant and does not depend on x), it can be shown that the modes (i.e., local maxima) of the empirical probability density function (pdf) correspond to stable equilibria, whereas the antimodes of the pdf (i.e., local minima) correspond to unstable equilibria (see e.g., [15], p. 273). More generally, there is a simple one-to-one correspondence between an additive noise SDE and its stationary pdf. Hence, instead of fitting the drift function of the cusp model directly, it can also be determined by fitting the pdf:

$$p(y|\alpha, \beta) = N \exp \left[-\frac{1}{4}y^4 + \frac{1}{2}\beta y^2 + \alpha y \right], \quad (3)$$

where N is a normalizing constant. In (3), the observed dependent variable x has been rescaled by $y = (x - \lambda)/\sigma$, and α and β are linear functions of the two control variables a and b as follows: $\alpha = k_0 + k_1a + k_2b$ and $\beta = l_0 + l_1a + l_2b$. The parameters λ , σ , k_0 , k_1 , k_2 , l_0 , l_1 , and l_2 can be estimated using maximum likelihood procedures (see **Maximum Likelihood Estimation**) [5].

Although the maximum likelihood method of Cobb is the most elegant and statistically satisfactory method for fitting the cusp catastrophe model to date, it is not used often. One reason may be that Cobb's computer program for fitting the cusp model can sometimes behave erratically. This problem was addressed in Hartelman [12, 13], who outlined a more robust and more flexible version of Cobb's original program. The improved program, Cuspfit, uses a more reliable optimization routine, allows the user to constrain parameter values and to employ different sets of starting values, and is able to fit competing models such as the logistic model. Cuspfit is available at <http://users.fmg.uva.nl/hvandermaas>.

We now illustrate Cobb's maximum likelihood procedure with a practical example on sudden transitions in attitudes [29]. The data set used here is taken from Stouffer et al. [25], and has been discussed in relation to the cusp model in Latané and Nowak [18]. US soldiers were asked their opinion about three issues (i.e., postwar conscription, demobilization, and the Women's Army Corps). An individual attitude score was obtained by combining responses on different questions relating to the

same issue, resulting in an attitude score that could vary between 0 (unfavorable) to 6 (favorable). In addition, respondents indicated the intensity of their opinion on a six-point scale. Thus, this data set consists of one behavioral variable (i.e., attitude) and only one control variable (i.e., the splitting variable ‘intensity’).

Figure 3 displays the histograms of attitude scores for each level of intensity separately. The data show that as intensity increases, the attitudes become polarized (i.e., divergence) resulting in a bimodal histogram for the highest intensities. The dotted line shows the fit of the cusp model. The maximum likelihood method as implemented in Cuspfitt allows for easy model comparison. For instance, one popular model selection method is the Bayesian information criterion (BIC; e.g., [23]), defined as $BIC = -2 \log L + k \log n$, where L is the maximum likelihood, k is the number of free parameters, and n is the number of observations. The BIC implements Occam’s razor by quantifying the trade-off between goodness-of-fit and parsimony, models with lower BIC values being preferable.

The cusp model, whose fit is shown in Figure 3, has a BIC value of 1787. The Cuspfitt program is also able to fit competing models to the data. An example of this is the logistic model, which allows for rapid changes in the dependent variable but cannot handle divergence. The BIC for the logistic model was 1970. To get a feeling for how big this difference really is, one may approximate $P(\text{logistic} | \text{data})$, the probability that the logistic model is



Figure 3 Histograms of attitude scores for five intensities of feeling (data from Stouffer et al., 1950). The dotted line indicates the fit of the cusp model. Both the data and the model show that bimodality in attitudes increases with intensity of feeling. Adapted from van der Maas et al. (29)

true and the cusp model is not, given the data, by $\exp\{(-1/2)(BIC_{\text{logistic}})\} / [\exp\{(-1/2)(BIC_{\text{logistic}})\} + \exp\{(-1/2)(BIC_{\text{cusp}})\}]$ (e.g., [23]). This approximation estimates $P(\text{logistic} | \text{data})$ to be about zero – consequently, the complement $P(\text{cusp} | \text{data})$ equals about one.

One problem of the Cobb method remaining to be solved is that the convenient relation between the pdf and the SDE (i.e., modes corresponding to stable states, antimodes corresponding to unstable states) breaks down when the noise is multiplicative, that is, when $D(x)$ in (2) depends on x . Multiplicative noise is often believed to be present in economic and financial systems (e.g., time series of short-term interest rates, [16]). In general, multiplicative noise arises under nonlinear transformations of the dependent variable x . In contrast, deterministic catastrophe theory is invariant under any smooth and revertible transformation of the dependent variable. Thus, Cobb’s stochastic catastrophe theory loses some of the generality of its deterministic counterpart (see [12], for an in-depth discussion of this point).

Summary and Recommendation

Catastrophe theory is a theory of great generality that can provide useful insights as to how behavior may radically change as a result of smoothly varying control variables. We discussed three statistical procedures for fitting one of the most popular catastrophe models, i.e., the cusp model. Two of these procedures, Guastello’s polynomial regression technique and GEMCAT, are suspect because these methods are unable to distinguish between stable and unstable equilibria. The maximum likelihood method developed by Cobb does not have this problem. The one remaining problem with the method of Cobb is that it is not robust to nonlinear transformations of the dependent variable. Future work, along the lines of Hartelman [12], will have to find a solution to this challenging problem.

References

- [1] Aerts, D., Czachor, M., Gabora, L., Kuna, M., Posiewnik, A., Pykacz, J. & Syty, M. (2003). Quantum morphogenesis: a variation on Thom’s catastrophe theory, *Physical Review E* **67**, 051926.

- [2] Alexander, R.A., Herbert, G.R., Deshon, R.P. & Hanges, P.J. (1992). An examination of least squares regression modeling of catastrophe theory, *Psychological Bulletin* **111**, 366–374.
- [3] Cobb, L. (1978). Stochastic catastrophe models and multimodal distributions, *Behavioral Science* **23**, 360–374.
- [4] Cobb, L. (1981). Parameter estimation for the cusp catastrophe model, *Behavioral Science* **26**, 75–78.
- [5] Cobb, L. & Watson, B. (1980). Statistical catastrophe theory: an overview, *Mathematical Modelling* **1**, 311–317.
- [6] Cobb, L. & Zacks, S. (1985). Applications of catastrophe theory for statistical modeling in the biosciences, *Journal of the American Statistical Association* **80**, 793–802.
- [7] Flay, B.R. (1978). Catastrophe theory in social psychology: some applications to attitudes and social behavior, *Behavioral Science* **23**, 335–350.
- [8] Gardiner, C.W. (1983). *Handbook of Stochastic Methods*, Springer-Verlag, Berlin.
- [9] Gilmore, R. (1981). *Catastrophe Theory for Scientists and Engineers*, Dover, New York.
- [10] Guastello, S.J. (1988). Catastrophe modeling of the accident process: organizational subunit size, *Psychological Bulletin* **103**, 246–255.
- [11] Guastello, S.J. (1992). Clash of the paradigms: a critique of an examination of the polynomial regression technique for evaluating catastrophe theory hypotheses, *Psychological Bulletin* **111**, 375–379.
- [12] Hartelman, P. (1997). *Stochastic catastrophe theory*, Unpublished doctoral dissertation, University of Amsterdam.
- [13] Hartelman, P.A.I., van der Maas, H.L.J. & Molenaar, P.C.M. (1998). Detecting and modeling developmental transitions, *British Journal of Developmental Psychology* **16**, 97–122.
- [14] Holyst, J.A., Kacperski, K. & Schweitzer, F. (2000). Phase transitions in social impact models of opinion formation, *Physica Series A* **285**, 199–210.
- [15] Honerkamp, J. (1994). *Stochastic Dynamical Systems*, VCH Publishers, New York.
- [16] Jiang, G.J. & Knight, J.L. (1997). A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model, *Econometric Theory* **13**, 615–645.
- [17] Lange, R., Oliva, T.A. & McDade, S.R. (2000). An algorithm for estimating multivariate catastrophe models: GEMCAT II, *Studies in Nonlinear Dynamics and Econometrics* **4**, 137–168.
- [18] Latané, B., & Nowak, A. (1994). Attitudes as catastrophes: from dimensions to categories with increasing involvement, in *Dynamical Systems in Social Psychology*, R.R. Vallacher & A. Nowak, eds, Academic Press, San Diego, pp. 219–249.
- [19] Newell, K.M., Liu, Y.-T. & Mayer-Kress, G. (2001). Time scales in motor learning and development, *Psychological Review* **108**, 57–82.
- [20] Oliva, T., DeSarbo, W., Day, D. & Jedidi, K. (1987). GEMCAT: a general multivariate methodology for estimating catastrophe models, *Behavioral Science* **32**, 121–137.
- [21] Ploeger, A., van der Maas, H.L.J. & Hartelman, P.A.I. (2002). Stochastic catastrophe analysis of switches in the perception of apparent motion, *Psychonomic Bulletin & Review* **9**, 26–42.
- [22] Poston, T. & Stewart, I. (1978). *Catastrophe Theory and its Applications*, Dover, New York.
- [23] Raftery, A.E. (1995). Bayesian model selection in social research, *Sociological Methodology* **25**, 111–163.
- [24] Stewart, I.N. & Peregoy, P.L. (1983). Catastrophe theory modeling in psychology, *Psychological Bulletin* **94**, 336–362.
- [25] Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A. & Clausen, J.A. (1950). *Measurement and Prediction*, Princeton University Press, Princeton.
- [26] Tamaki, T., Torii, T. & Meada, K. (2003). Stability analysis of black holes via a catastrophe theory and black hole thermodynamics in generalized theories of gravity, *Physical Review D* **68**, 024028.
- [27] Thom, R. (1975). *Structural Stability and Morphogenesis*, Benjamin-Addison Wesley, New York.
- [28] Torres, J.-L. (2001). Biological power laws and Darwin's principle, *Journal of Theoretical Biology* **209**, 223–232.
- [29] van der Maas, H.L.J., Kolstein, R. & van der Pligt, J. (2003). Sudden transitions in attitudes, *Sociological Methods and Research* **32**, 125–152.
- [30] van der Maas, H.L.J. & Molenaar, P.C.M. (1992). Stages of cognitive development: an application of catastrophe theory, *Psychological Review* **99**, 395–417.
- [31] van Harten, D. (2000). Variable nodding in *Cyprideis torosa* (Ostracoda, Crustacea): an overview, experimental results and a model from catastrophe theory, *Hydrobiologica* **419**, 131–139.
- [32] Wales, D.J. (2001). A microscopic basis for the global appearance of energy landscapes, *Science* **293**, 2067–2070.
- [33] Wimmers, R.H., Savelsbergh, G.J.P., van der Kamp, J. & Hartelman, P. (1998). A developmental transition in prehension modeled as a cusp catastrophe, *Developmental Psychobiology* **32**, 23–35.
- [34] Zahler, R.S. & Sussmann, H.J. (1977). Claims and accomplishments of applied catastrophe theory, *Nature* **269**, 759–763.
- [35] Zeeman, E.C. (1977). *Catastrophe Theory: Selected Papers (1972–1977)*, Addison-Wesley, New York.

(See also **Optimization Methods**)

E.-J. WAGENMAKERS, H.L.J. VAN DER MAAS
AND P.C.M. MOLENAAR

Categorizing Data

VALERIE DURKALSKI AND VANCE W. BERGER

Volume 1, pp. 239–242

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Categorizing Data

Categorizing Continuous Variables

Changing a continuous variable to a categorical form is common in many data analyses because of the impression that categorization makes it easier for clinical interpretation and avoids complex statistical assumptions. Indeed, methods of analysis begin with the most basic form of data, a **2 × 2 contingency table**, which classifies data into one of four unique categories. By splitting a continuous variable such as blood pressure, tumor size, or age into two discrete groups (above and below a given threshold, which may be either predetermined or data-dependent), the data can be presented as the number belonging to each classification (above and below).

Another common form of categorical data is ordered categories (i.e., worse, unchanged, improved). Ordered categorical variables are intermediate between continuous variables and binary variables, and can be the result of categorizing (into more than two categories) a continuous variable. Alternatively, the categories may be defined without reference to any continuous variable. Either way, an ordered categorical variable can be further categorized into another ordered categorical variable with fewer categories, or, in the extreme scenario, into a binary variable. Categorization is thus seen to apply to continuous variables, which may be transformed into ordered categorical variables or to binary variables, and to ordered categorical variables, which may be transformed into ordered categorical variables with fewer categories or to binary variables.

Categorization is attractive for descriptive purposes and for separating patients into risk groups. It also provides a data structure conducive to a familiar, simply executed, and easily interpreted analysis. Given two groups, and data from each group classified into the same two groups (above or below the same threshold), one can produce confidence intervals for the difference between proportions or test the equality, across the groups, of the two success probabilities (e.g., statistical independence between groups and outcomes). However, the advantages and disadvantages need to be understood prior to engaging in this approach.

Methods of Categorization

To categorize a continuous variable or an ordered categorical variable, one needs to determine a cutpoint. This determination may be made based on prior knowledge, or the data at hand. If the data at hand are used, then the determination of the cutpoint may be based on a visual inspection of a graph, the optimization of prediction of or association with some other variable, the maximization of a between-group difference, or the minimization of a *P* value. If the data are to be used to select the cutpoint, then a graphical display may give insight into the appropriate grouping. Inspection of data from the sample using plots can reveal clusters within any given variable and relationships between two variables (i.e., outcome and prognostic variables). These can help to identify potential cutpoints that differentiate patients into distinct groups (i.e., high and low risk). So, for example, if in a given sample of subjects the ages could range from 18 to 80, but as it turns out there is no middle (each subject is either under 30 or over 70), then there is no need to determine the cutpoint beyond specifying that it might fall between 30 and 70.

Relationships between two variables can appear as step-functions, which may be either monotonic or nonmonotonic (i.e., a U-shaped distribution). In such a case, the effect of the independent variable on the dependent variable depends on which cluster the independent variable falls in. Suppose, for example, that systolic blood pressure depends on age, but not in a continuous fashion. That is, suppose that the mean blood pressure for the patients born in a given decade is the same, independent of which year within the decade they were born. If this were the case, then it would provide a clear idea of where the cutpoints should be placed.

Another approach is to choose equally spaced intervals or, if this creates sparsely populated groups, then the median value can be used to obtain roughly equally represented groups [8]. A more systematic approach is to use various cutpoints to determine the one that best differentiates risk groups [11]. Of course, the *P* value that results from a naïve comparison using this optimally chosen cutpoint will be artificially low [3], however, and even if it is adjusted to make it valid, the estimate remains biased [9]. Mazumdar and Glassman [10] offer a nice summary on choosing cutpoints, which includes

a description of the P value adjustment formulae, while Altman [1] discusses creating groups in a regression setting.

Standard cutpoints, which are not influenced by the present data, are useful particularly when comparing results across studies because if the cutpoint changes over time, then it is difficult, if not impossible, to use past information for comparing studies or combining data. In fact, definitions of severity of cancer, or any other disease, that vary over time can lead to the illusion that progress has been made in controlling this disease when in fact it may not have – this is known as *stage migration* [7].

Implications of Chosen Cutpoints

It is illustrated in the literature that categorizing continuous outcome variables can lead to a loss of information, increased chance of misclassification error, and a loss in statistical **power** [1, 8, 12–17]. Categorizing a continuous predictor variable can lead to a reversal in the direction of its effect on an outcome variable [5]. Ragland [13] illustrates with a blood pressure example that the choice of cutpoint affects the estimated measures of association (i.e., proportion differences, prevalence ratios, and odds ratios) and that as the difference between the two distribution means increases, the variations in association measurements and power increase. In addition, Connor [6] and Suissa [18] show that tests based on frequencies of a dichotomous endpoint are, in general, less efficient than tests using mean-value statistics when the underlying distribution is normal (efficiency is defined as the ratio of the expected variances under each model; the model with the lower variance ratio is regarded as more efficient).

Even if the categorization is increased to three or four groups, the relative efficiency (see **Asymptotic Relative Efficiency**) is still less than 90%. This implies that if there really is a difference to detect, then a study using categorized endpoints would require a larger sample size to be able to detect it than would a study based on a continuous endpoint. Because of this, it has recently been suggested that binary variables be ‘reverse dichotomized’ and put together to form the so-called information-preserving composite endpoints, which are more informative than any of the binary endpoints used in their creation (see **Analysis of Covariance: Nonparametric**) [2].

These more informative endpoints are then amenable to a wider range of more powerful analyses that make use of all the categories, as opposed to just Fisher’s exact test on two collapsed categories. More powerful analyses include the **Wilcoxon-Mann Whitney test** and the Smirnov test (see **Kolmogorov-Smirnov Tests**), as well as more recently developed tests such as the convex hull test [4] and adaptive tests [3].

So why would one consider categorizing a continuous variable if it increases the chance of missing real differences, increases the chance of misclassification, and increases the sample size required to detect differences? The best argument thus far seems to be that categorization simplifies the analyses and offers a better approach to understanding and interpreting meaningful results (proportions versus mean-values). Yet one could argue that when planning a research study, one should select the primary response variable that will give the best precision of an estimate or the highest statistical power. In the same respect, prognostic variables (i.e., age, tumor size) should be categorized only according to appropriate methodology to avoid misclassification. Because it is common to group populations into risk groups for analysis and for the purposes of fulfilling eligibility criteria for stratified randomization schemes, categorization methods need to be available.

The best approach may be to collect data on a continuum and then categorize when necessary (i.e., for descriptive purposes). This approach offers the flexibility to conduct the prespecified analyses whether they are based on categorical or continuous data, but it also allows for secondary exploratory and sensitivity analyses to be conducted.

References

- [1] Altman, D.G. (1998). Categorizing continuous variables, in *Encyclopedia of Biostatistics*, 1st Edition, P. Armitage and T. Colton, eds, Wiley, Chichester, pp. 563–567.
- [2] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [3] Berger, V.W. & Ivanova, A. (2002). Adaptive tests for ordered categorical data, *Journal of Modern Applied Statistical Methods* **1**, 269–280.
- [4] Berger, V.W., Permutt, T. & Ivanova, A. (1998). The convex hull test for ordered categorical data, *Biometrics* **54**, 1541–1550.
- [5] Brenner, H. (1998). A potential pitfall in control of covariates in epidemiologic studies, *Epidemiology* **9**(1), 68–71.

-
- [6] Connor, R.J. (1972). Grouping for testing trends in categorical data, *Journal of the American Statistical Association* **67**, 601–604.
 - [7] Feinstein, A., Sosin, D. & Wells, C. (1985). The Will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer, *New England Journal of Medicine* **312**, 1604–1608.
 - [8] MacCallum, R.C., Zhang, S., Preacher, K.J. & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables, *Psychological Methods* **7**, 19–40.
 - [9] Maxwell, S.E. & Delaney, H.D. (1993). Bivariate median splits and spurious statistical significance, *Psychological Bulletin* **113**, 181–190.
 - [10] Mazumdar, M. & Glassman, J.R. (2000). Tutorial in biostatistics: categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments, *Statistics in Medicine* **19**, 113–132.
 - [11] Miller, R. & Siegmund, D. (1982). Maximally selected chi-square statistics, *Biometrics* **38**, 1011–1016.
 - [12] Moses, L.E., Emerson, J.D. & Hosseini, H. (1984). Analyzing data from ordered categories, *New England Journal of Medicine* **311**, 442–448.
 - [13] Ragland, D.R. (1992). Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint, *Epidemiology* **3**, 434–440.
 - [14] Rahrfs, V.W. & Zimmermann, H. (1993). Scores: ordinal data with few categories – how should they be analyzed? *Drug Information Journal* **27**, 1227–1240.
 - [15] Sankey, S.S. & Weissfeld, L.A. (1998). A study of the effect of dichotomizing ordinal data upon modeling, *Communications in Statistics Simulation* **27**(4), 871–887.
 - [16] Senn, S. (2003). Disappointing dichotomies, *Pharmaceutical Statistics* **2**, 239–240.
 - [17] Streiner, D.L. (2002). Breaking up is hard to do: the heartbreak of dichotomizing continuous data, *Canadian Journal of Psychiatry* **47**(3), 262–266.
 - [18] Suissa, S. (1991). Binary methods for continuous outcomes: a parametric alternative, *Journal of Clinical Epidemiology* **44**(3), 241–248.
- (See also **Optimal Design for Categorical Variables**)
- VALERIE DURKALSKI AND VANCE W. BERGER

Cattell, Raymond Bernard

PAUL BARRETT

Volume 1, pp. 242–243

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cattell, Raymond Bernard

Born: March 20, 1905, in West Bromwich, UK.

Died: February 2, 1998, in Honolulu, USA.

At the age of just 19 years, Raymond Bernard Cattell was awarded his first university degree (with first class honors) from University College, London in 1924, reading chemistry. However, during his time here, he was influenced by the thinking and research of three of the most significant figures in mathematical and statistical psychology in the twentieth century, **Cyril Burt**, **Ronald Fisher**, and **Charles Spearman**; all of whom were working within laboratories in close proximity to his own. The magnificent intellectual stimulation of these three, allied to Cattell's own interest in the classification process (exemplified in Dmitri Mendeléeff's Periodic Table of elements as a model for classifying human attributes), led him to pursue a doctorate in human psychology in King's College, London, working on cognition and perception. Spearman was his mentor throughout this period, and in 1929, he was awarded his Ph.D. During the next eight years, owing to the lack of any research positions in the new discipline of psychology, he took a teaching position at Exeter University, UK, within the Education and Psychology department. In 1932, he became Director of the School Psychological Services and Clinic in Leicester, UK. However, Cattell's research career really began in earnest in 1937 with an invitation to join Edward Thorndike in a research associate position within his laboratory at Columbia University, USA. From here, he accepted the G. Stanley Hall professorship at Clark University in Massachusetts, where he developed and announced his theory of fluid versus crystallized intelligence to the 1941 American Psychological Association convention. From 1941 to 1944, he worked within the Harvard University psychology faculty. It was during this time at Harvard that the factor analytic and multivariate psychometrics that was to be his major contribution to human psychology began. In 1945, he was appointed to a newly created research professorship in psychology at the University of Illinois, USA. He remained at Illinois for the next 29 years in the Laboratory of Personality and Group Analysis until his 'official'

retirement in 1973. In 1949, he published the first version of what was to become one of the world's most famous personality tests, the Sixteen Personality Factor test. After his retirement in 1973, Cattell set up the Institute for Research on Morality and Adjustment in Boulder, Colorado, USA. Then, after an existing heart condition was aggravated by the cold mountain conditions and high altitude of his location in Boulder, he moved to Hawaii in 1978 as an unpaid visiting professor within the department of psychology, a position which he held for 20 years until his death. In 1997, one year before his death, he was nominated to receive the prestigious gold medal award for lifetime achievement from the American Psychological Association and the American Psychological Foundation. Unfortunately, owing to an ongoing controversy about his later work, which publicly distorted his actual research and thinking on morality and sociological matters, he declined to accept the award.

Raymond Cattell authored and coauthored 55 books and over 500 scientific articles. Some of these are the most significant publications in human trait psychology and psychometrics of that era. For example, his 1957 book entitled *Personality and Motivation Structure and Measurement* [1] was the handbook from which the major tenets and research outputs of Cattellian Theory were first exposed in a detailed and unified fashion. In 1966, the *Handbook of Multivariate Experimental Psychology* [2] was published, a landmark publication that described research and advances in multivariate statistics, ability, motivation, and personality researches. Another influential book was authored by him in 1971, *Abilities, Their Structure, Growth, and Action* [3]. This book described the entire theory of Fluid and Crystallized Intelligence, along with introducing the hierarchical multivariate model of human abilities that now dominates modern-day psychometric theories of intelligence. His 1978 book on **factor analysis** [4] formed the major summary of his innovations and approach to using this multivariate methodology.

References

- [1] Cattell, R.B. (1957). *Personality and Motivation Structure and Measurement*, World Book, New York.

2 Cattell, Raymond Bernard

- [2] Cattell, R.B., ed. (1966). *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago.
- [3] Cattell, R.B. (1971). *Abilities: Their Structure, Growth, and Action*, Houghton-Mifflin, Boston.
- [4] Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*, Plenum Press, New York.

PAUL BARRETT

Censored Observations

VANCE W. BERGER

Volume 1, pp. 243–244

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Censored Observations

Most common statistical analyses are based on actually observing the values of the observations to be analyzed. However, it is not always the case that an observation is observed in its entirety. Three common types of censoring are right censoring, left censoring, and interval censoring. Each arises in different contexts. An observation can be left censored, for example, if a study is conducted to measure how long it takes for infants to develop a certain skill, but some infants have already acquired the skill by the time they are studied. In such a case, the time to develop this skill is not observed for this infant, but it is known that this time is shorter than the infant's age. So the actual time, though unknown, is known to fall within an interval from zero to the observation time, or everything to the left of the observation time.

Another context in which an observation may be left censored is when an assay has a certain detection limit. That is, a certain agent will be detected only if its concentration exceeds the detection limit. If it is not detected, then this does not mean that it is not present. Rather, it means that the concentration falls between zero and the detection limit, or left of the detection limit. So it is left censored.

Similarly, if all that is known is that an agent was detected but its concentration is not known, then the unobserved concentration is known only to fall to the right of the detection limit, so it is right censored [7]. Right censoring is sufficiently common that it has been further classified into Type I and Type II censoring. Type I censoring describes the situation in which a study is terminated at a particular point in time, even if not all patients have been observed to the event of interest. In this case, the censoring time is fixed (unless some patients dropped out of the trial prior to this common censoring time), and the number of events is a random variable. With Type II censoring, the study would be continued until a fixed number of events are observed. For example, the trial may be stopped after exactly 50 patients die. In this case, the number of events is fixed, and censoring time is the random variable.

In most **clinical trials**, patients could be censored owing to reasons not under the control of the

investigator. For example, some patients may be censored because of loss to follow-up, while others are censored because the study has terminated. This censoring is called *random censoring*.

Besides left and right censoring, there is also interval censoring, in which an observation is known only to fall within a certain interval. If, for example, a subject is followed for the development of cancer, and does not have cancer for the first few evaluations, but then is found to have cancer, then it is known only that the cancer developed (at least to the point of being detectable) between the last 'clean' visit and the first visit during which the cancer was actually detected. Technically speaking, every observation is interval censored, because of the fact that data are recorded to only a finite number of decimal places. An age of 56 years, for example, does not mean that the subject was born precisely 56 years ago, to the second. Rather, it generally means that the true age is somewhere between 55.5 and 56.5 years. Of course, when data are recorded with enough decimal places to make them sufficiently precise, this distinction is unimportant, and so the term interval censoring is generally reserved for cases in which the interval is large enough to be of some concern. That is, the lower endpoint and the upper endpoint might lead to different conclusions.

One other example of interval censoring is for the P value of an exact permutation test (*see Permutation Based Inference*) whose reference distribution is discrete. Consider, for example, Fisher's exact test (*see Exact Methods for Categorical Data*). Only certain P values can be attained because of the discrete distribution (even under the null hypothesis, the P value is not uniformly distributed on the unit interval), and so what is interval censored is the value the P value would have attained without the conservatism. For truth in reporting, it has been suggested that this entire interval be reported as the P value interval [1] (*see Mid- P Values*).

Of course, left censoring, right censoring, and interval censoring do not jointly cover all possibilities. For example, an observation may be known only to fall outside of an interval, or to fall within one of several intervals. However, left censoring, right censoring, and interval censoring tend to cover most of the usual applications.

The analysis of censored data, regardless of the type of censoring, is not as straightforward as the analysis of completely observed data, and special methods have been developed. With censoring, the endpoint may be considered to be measured on only a partially ordered scale, and so methods appropriate for partially ordered endpoints may be used [2]. However, it is more common to use the log-rank test to test if two distributions are the same, Cox proportional hazards regression [3, 4] to assess covariate effects, and Kaplan–Meier curves [5] to estimate the survival time distribution (*see Survival Analysis*). Accelerated failure time methods can also be used. See also [6]. It has been noted that when analyzing censored data, a commonsense criterion called ‘rationality’ may conflict with the desirable property of unbiasedness of a test [8].

Although the concept of censoring may have developed in the biomedical research, censored observations may occur in a number of different areas of research. For example, in the social sciences one may study the ‘survival’ of a marriage. By the end of the study period, some subjects probably may still be married (to the same partner), and hence those subjects represent censored observations. Likewise, in engineering, components have failure times, which may be right censored.

References

- [1] Berger, V.W. (2001). The p-value interval as an inferential tool, *Journal of the Royal Statistical Society D (The Statistician)* **50**(1), 79–85.
- [2] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**(1), 45–55.
- [3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B* **34**, 187–220.
- [4] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London.
- [5] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [6] Kleinbaum, D.G. (1995). *Survival Analysis – A Self-learning Text*, Springer-Verlag, New York.
- [7] Lachenbruch, P.A., Clements, P.J. & He, W. (1995). On encoding values for data recorded as X_iC , *Biometrical Journal* **37**(7), 855–867.
- [8] Sackrowitz, H. & Samuel-Cahn, E. (1994). Rationality and unbiasedness in hypothesis testing, *Journal of the American Statistical Association* **89**(427), 967–971.

(*See also Survival Analysis*)

VANCE W. BERGER

Census

JERRY WELKENHUYSEN-GYBELS AND DIRK HEERWEGH

Volume 1, pp. 245–247

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Census

Definition

In a broad sense, a census can be defined as the gathering of information from all members of any type of population. In this broad sense, the term has, for example, been used to refer to an enumeration of the mountain gorilla population in the vicinity of the Virunga Volcanoes in Central Africa, which established a 17% growth in the population between 1989 and 2004.

In the context of behavioral research, however, the more particular definition issued by the United Nations is used. According to the latter definition, a census can be described as ‘the total process of collecting, compiling, evaluating, analyzing and publishing demographic, economic and social data pertaining, at a specified time, to all persons in a country or in a well-delimited part of a country’ [6]. The latter definition highlights several key features of a census [1, 3]:

1. Government sponsorship: A census is sponsored by the government of the country or area in which the census takes place, because only a government can provide the required resources for the census and the necessary authority and legislation to make public participation to the census enforceable.
2. The census must cover a precisely defined area. Most often, country boundaries define the census area.
3. All persons in the scope of the census are included without omission or duplication.
4. The information is collected at the individual level rather than at some aggregate level, for example, household level. Note, however, that this does not preclude the collection of information about the household to which the individual belongs.
5. The information must be gathered within a specified time interval.
6. The information that is collected via the census must be made publicly available.

Purpose of the Census

Earlier on, we have mentioned that censuses are sponsored by the government, because only a

government can establish the necessary conditions for its conduct. Government involvement is, however, also due to the fact that the information collected by the census is necessary and sometimes vital for government policy and administration. For instance, the main objective of the US census is to perform an enumeration of the total population residing on its territory, allowing the proportional apportionment of federal funds and seats in the US House of Representatives to the different states [7]. Similarly, the South African census provides data to guide the postapartheid Reconstruction and Development Program (RDP). This census is used for the distribution of resources to redress social inequalities by providing expanded opportunities for education, housing, employment, and improving the situation of women and the youth. Additionally, the South African census provides data on fertility, mortality, and the devastation resulting from the HIV/AIDS pandemic [2].

These purposes, of course, also have their impact on the content of the census. Often, the content of the census depends on the country’s needs at the time that the census is administered although some standard information is collected in nearly every census. The United Nations [5] have also approved a set of recommendations for population censuses that includes the following topics: location at time of census, place of usual residence, relation to the head of the household or family, sex, age, marital status, place of birth, citizenship, whether economically active or not, occupation, industry, status (e.g., employer, employee), language, ethnic characteristics, literacy, education, school attendance, and number of children.

A Historical Perspective

Population counts go back a long time in history. There are reports of population counts in ancient Japan, by ancient Egyptians, Greeks, Hebrews, Romans, Persians, and Peruvians [3, 4]. The purpose of these surveys was to determine, for example, the number of men at military age or the number of households eligible for taxes. Because of these goals, early censuses did not always focus on the entire population within a country or area. Modern-day censuses are traced back to the fifteenth and sixteenth century. However, the oldest census that has been

carried out on a regular basis is the US census, which was first held in 1790 and henceforth repeated every ten years. The United Kingdom census also has a long tradition, dating back to 1801 and being held every 10 years from then on (except in 1941). By 1983, virtually every country in the world had taken a census of its population [4].

Some countries, like the United States and the United Kingdom, have a tradition of conducting a census every 5 or 10 years. A lot of countries, however, only hold a census when it is necessary and not at fixed dates.

Methodological Issues

The census data are usually gathered in one of two ways: either by self-enumeration or by direct enumeration. In the former case, the questionnaire is given to the individual from whom one wants to collect information and he/she fills out the questionnaire him/herself. In the case of direct enumeration, the questionnaire is filled out via a face-to-face interview with the individual.

Given the fact that the census data are collected via questionnaires, all methodological issues that arise with respect to the construction and the use of questionnaires also apply in the census context (*see Survey Questionnaire Design*). First of all, it is important that the questions are worded in such a way that they are easily understandable for everyone. For this reason, a pilot study is often administered to a limited sample of individuals to test the questionnaire and detect and solve difficulties with the questions as well as with the questionnaire in general (e.g., too long, inadequate layout, etc.). Secondly, individuals might not always provide correct and/or truthful answers to the questions asked. In the context of a census, this problem is, however, somewhat different than in the context of other questionnaire-based research that is not sponsored by the government. On the one hand, the government usually issues legislations that enforce people to provide correct and truthful answers, which might lead some people to do so more than in any other context. On the other hand, because the information is used for government purposes, some people might be hesitant to convey certain information out of fear that it in some way will be used against them.

Thirdly, in the case of direct enumeration, interviewers are used to gather the census information.

It is well known that interviewers often influence respondents' answers to questions. To minimize this type of interviewer effects, it is important to provide these interviewers with the necessary training with respect to good interview conduct. Furthermore, interviewers also need to be supervised in order to avoid fraud.

Another issue that needs to be considered is non-response (*see Missing Data*). Although governments issue legislation that enforces participation in the census, there are always individuals who can or will not cooperate. Usually, these nonresponders do not represent a random sample from the population, but are systematically different from the responders. Hence, there will be some bias in the statistics that are calculated on the basis of the census. However, given the fact that information about some background characteristics of these nonresponders are usually available from other administrative sources, it can be assessed to what extent they are different from the responders and (to some extent) a correction of the statistics is possible.

Over the course of the years, censuses have become more than population enumerations. The US Census, for example, has come to collect much more information, such as data on manufacturing, agriculture, housing, religious bodies, employment, internal migration, and so on [7]. The contemporary South Africa [2] census (*see above*) also illustrates the wide scope of modern censuses. Expanding the scope of a census could not have happened without evolutions in other domains. Use of computers and optical sensing devices for data input have greatly increased the speed with which returned census forms can be processed and analyzed (*see Computer-Adaptive Testing*). Also, censuses have come to use sampling techniques (*see Survey Sampling Procedures*). First of all, certain questions are sometimes only administered to a random sample of the population to avoid high respondent burden due to long questionnaires. This implies that not all questions are administered to all individuals. Questions could for instance be asked on a cyclical basis: One set of questions for individual 1, another set for individual 2, and so on, until individual 6, who again receives the set of questions from the first cycle [7]. As another example, the 2000 US census used a short and a long form, the latter being administered to a random sample of the population.

Secondly, some countries have recently abandoned the idea of gathering information from the entire population, because a lot of administrative information about a country's population has become available via other databases that have been set up over the years. Belgium, for example, from 2001 onward has replaced the former decennial population censuses by a so-called General Socio-economic Survey that collects data on a large sample (Sample fraction: 20–25%) from the population.

References

- [1] Benjamin, B. (1989). Census, in *Encyclopedia of Statistical Sciences*, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 397–402.
- [2] Kinsella, K. & Ferreira, M. (1997). Aging trends: South Africa, Retrieved March 2, 2004 from <http://www.census.gov/ipc/prod/ib-9702.pdf>.
- [3] Taeuber, C. (1978). Census, in *International Encyclopedia of Statistics*, W.H. Kruskal & J.M. Tanur, eds, Free Press, New York, pp. 41–46.
- [4] Taeuber, C. (1996). Census of population, in *The Social Science Encyclopedia*, A. Kuper & J. Kuper, eds, Routledge, London, pp. 77–79.
- [5] United Nations, Statistical office. (1958). *Handbook of Population Census Methods*, Vol. 1: *General Aspects of a Population Census*, United Nations, New York.
- [6] United Nations. (1980). *Principles and Recommendations for Population and Housing Censuses*, United Nations, New York.
- [7] US Bureau of the Census (South Dakota). Historical background of the United States census, Retrieved March 2, 2004 from <http://www.census.gov/mso/www/centennial/bkgrnd.htm>

JERRY WELKENHUYSEN-GYBELS AND
DIRK HEERWEGH

Centering in Linear Multilevel Models

JAN DE LEEUW

Volume 1, pp. 247–249

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Centering in Linear Multilevel Models

Consider the situation in which we have m groups of individuals, where group j has n_j members. We consider a general multilevel model, (see **Linear Multilevel Models**) that is, a random coefficient model for each group of the form

$$y_{ij} = \beta_{0j} + \sum_{s=1}^p x_{ijs} \beta_{sj} + \epsilon_{ij}, \quad (1)$$

where the coefficients are the outcomes of a second regression model

$$\beta_{sj} = \gamma_{0s} + \sum_{r=1}^q z_{jr} \gamma_{rs} + \delta_{sj}. \quad (2)$$

Both the $n_j \times (p+1)$ matrices X_j of first-level predictors and the $p \times (q+1)$ matrix Z of second-level predictors have a leading column with all elements equal to +1, corresponding with the intercepts of the regression equations. To single out the intercepts in our formulas, our indices for both first- and second-level predictors start at zero. Thus, $0 \leq s \leq p$ and $0 \leq r \leq q$ and $x_{ij0} = z_{j0} = 1$ for all i and j . Observe that the predictors X_j and Z are both non-random, either because they are fixed by design or because we condition our model on their observed values.

The disturbance vectors ϵ_j and δ_j have zero expectations. They are uncorrelated with each other and have covariance matrices $V(\epsilon_j) = \sigma^2 I$, where I is the $n_j \times n_j$ identity matrix, and where $V(\delta_j) = \Omega$. The $p \times p$ matrix Ω has elements ω_{st} . It follows that the expectations are

$$\begin{aligned} E(y_{ij}) &= \gamma_{00} + \sum_{r=1}^q z_{jr} \gamma_{r0} + \sum_{s=1}^p x_{ijs} \gamma_{0s} \\ &+ \sum_{s=1}^p \sum_{r=1}^q x_{ijs} z_{jr} \gamma_{rs}, \end{aligned} \quad (3)$$

and the covariances are

$$C(y_{ij}, y_{kj}) = \omega_{00} + \sum_{s=1}^p (x_{ijs} + x_{kjs}) \omega_{0s}$$

$$+ \sum_{s=1}^p \sum_{t=1}^p x_{ijs} x_{kjt} \omega_{st} + \delta^{ik} \sigma^2. \quad (4)$$

Here δ^{ik} is Kronecker's delta, that is, it is equal to one if $i = k$ and equal to zero otherwise. Typically, we define more restrictive models for the same data by requiring that some of the regression coefficients γ_{rs} and some of the variance and covariance components ω_{st} are zero.

In multilevel analysis, the scaling and centering of the predictors is often arbitrary. Also, there are sometimes theoretical reasons to choose a particular form of centering. See Raudenbush and Bryk 2, p. 31–34 or Kreft et al. [1]. In this entry, we consider the effect on the model of translations. Suppose we replace x_{is} by $\tilde{x}_{ijs} = x_{ijs} - a_s$. Thus, we subtract a constant from each first-level predictor, and we use the same constant for all groups. If the a_s are the predictor means, this means *grand mean centering* of all predictor variables. Using grand mean centering has some familiar interpretational advantages. It allows us to interpret the intercept, for instance, as the expected value if all predictors are equal to their mean value. If we do not center, the intercept is the expected value if all predictors are zero, and for many predictors used in behavioral science zero is an arbitrary or impossible value (think of a zero IQ, a zero income, or a person of zero height).

After some algebra, we see that

$$\begin{aligned} \gamma_{00} &+ \sum_{r=1}^q z_{jr} \gamma_{r0} + \sum_{s=1}^p x_{ijs} \gamma_{0s} \\ &+ \sum_{s=1}^p \sum_{r=1}^q x_{ijs} z_{jr} \gamma_{rs} = \tilde{\gamma}_{00} + \sum_{r=1}^q z_{jr} \tilde{\gamma}_{r0} \\ &+ \sum_{s=1}^p \tilde{x}_{ijs} \gamma_{0s} + \sum_{s=1}^p \sum_{r=1}^q \tilde{x}_{ijs} z_{jr} \gamma_{rs}, \end{aligned} \quad (5)$$

with

$$\tilde{\gamma}_{r0} = \gamma_{r0} + \sum_{s=1}^p \gamma_{rs} a_s \quad (6)$$

for all $0 \leq r \leq q$. Thus, the translation of the predictor can be compensated by a linear transformation of the regression coefficients, and any vector of expected values generated by the untranslated model can also be generated by the translated model. This

2 Centering in Linear Multilevel Models

is a useful type of invariance. But it is important to observe that if we restrict our untranslated model, for instance, by requiring one or more γ_{r0} to be zero, then those same γ_{r0} will no longer be zero in the corresponding translated model. We have invariance of the expected values under translation if the regression coefficients of the group-level predictors are nonzero.

In the same way, we can see that

$$\begin{aligned} \omega_{00} + \sum_{s=1}^p (x_{ijs} + x_{kjs})\omega_{0s} \\ + \sum_{s=1}^p \sum_{t=1}^p x_{ijs}x_{kjt}\omega_{st} = \\ \tilde{\omega}_{00} + \sum_{s=1}^p (\tilde{x}_{ijs} + \tilde{x}_{kjs})\tilde{\omega}_{0s} \\ + \sum_{s=1}^p \sum_{t=1}^p \tilde{x}_{ijs}\tilde{x}_{kjt}\tilde{\omega}_{st} \end{aligned} \quad (7)$$

if

$$\begin{aligned} \tilde{\omega}_{00} &= \omega_{00} + 2 \sum_{s=1}^p \omega_{0s}a_s + \sum_{s=1}^p \sum_{t=1}^p \omega_{st}a_sa_t, \\ \tilde{\omega}_{0s} &= \omega_{0s} + \sum_{t=1}^p \omega_{st}a_t. \end{aligned} \quad (8)$$

Thus, we have invariance under translation of the variance and covariance components as well, but, again, only if we do not require the ω_{0s} , that is, the covariances of the slopes and the intercepts, to be zero. If we center by using the grand mean of the predictors, we still fit the same model, at least in the case in which we do not restrict the γ_{r0} or the ω_{s0} to be zero.

If we translate by $\tilde{x}_{ijs} = x_{ijs} - a_{js}$ and thus subtract a different constant for each group, the situation

becomes more complicated. If the a_{js} are the group means of the predictors, this is *within-group centering*. The relevant formulas are derived in [1], and we will not repeat them here. The conclusion is that separate translations for each group cannot be compensated for by adjusting the regression coefficients and the variance components. In this case, there is no invariance, and we are fitting a truly different model. In other words, choosing between a translated and a nontranslated model becomes a matter of either theoretical or statistical (goodness-of-fit) considerations.

From the theoretical point of view, consider the difference in meaning of a grand mean centered and a within-group mean centered version of a predictor such as grade point average. If two students have the same grade point average (GPA), they will also have the same grand mean centered GPA. But GPA in deviations from the school mean defines a different variable, in which students with high GPAs in good schools have the same corrected GPAs as students with low GPAs in bad schools. In the first case, the variable measures GPA; in the second case, it measures how good the student is in comparison to all students in his or her school. The two GPA variables are certainly not monotonic with each other, and if the within-school variation is small, they will be almost uncorrelated.

References

- [1] Kreft, G.G., de Leeuw, J. & Aiken, L.S. (1995). The effects of different forms of centering in hierarchical linear models, *Multivariate Behavioral Research* **30**, 1–21.
- [2] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods*, 2nd Edition, Sage Publications, Newbury Park.

JAN DE LEEUW

Central Limit Theory

JEREMY MILES

Volume 1, pp. 249–255

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Central Limit Theory

Description

The normal distribution (*see Catalogue of Probability Density Functions*) is an essential and ubiquitous part of statistics, and many tests make an assumption relating to the normal distribution. The reason that the normal distribution is so omnipresent is because of *the central limit theorem*; it has been described, for example, as ‘the reason **analysis of variance** ‘works’ [3].

The assumption of normality of distribution made by many statistical tests is one that confuses many students – we do not (usually) make an assumption about the distribution of the measurements that we have taken, rather we make an assumption about the sampling distribution of those measurements – it is this sampling distribution that we assume to be normal, and this is where the central limit theorem comes in. The theorem says that the **sampling distribution** of sample means will approach normality as the sample size increases, *whatever the shape of the distribution of the measure that we have taken* (we must warn that the measurements must be independent, and identically distributed [i.i.d.]).

The theorem also tells us that given the population mean μ and variance σ^2 , the mean of the sampling distribution (that is, the mean of all of the means) $\mu_{\bar{x}}$ will be equal to the μ (the population mean), and that the variance of the sample means $\sigma_{\bar{x}}^2$ will be equal to σ^2/n (where n is the size of each of the samples). Taking the square root of this equation gives the familiar formula for the standard deviation of the sampling distribution of the mean, or standard error:

$$se_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \quad (1)$$

The mathematical proof of this requires some mathematical sophistication (although less than many mathematical proofs). One can show the moment-generating function (*see Moments*) of a standardized sample mean asymptotes to the moment-generating function of a standard normal distribution as the sample size increases; [2] it can also be seen at http://en.wikipedia.org/wiki/Central_limit_theorem, or <http://mathworld.wolfram.com/CentralLimitTheorem.html>

Demonstration

The idea of the sampling distribution becoming normal at large samples is one that can be difficult to understand (or believe), but a simple demonstration can show that this can occur. Imagine that our measurement is that of dice rolls – purely random numbers, from 1 to 6. If we were to take a sample of size $N = 1$ roll of a die, record the result, and repeat the procedure, a very large number of times, it is clear that the shape of the distribution would be uniform, not normal – we would find an equal proportion of every value.

It is hardly necessary to draw the graph, but we will do it anyway – it is shown in Figure 1.

The mean of this distribution is equal to 3.5, and the mean value that would be calculated from a dice roll is also equal to 3.5.

If we increase the sample to $n = 2$ dice rolled, the distribution will change shape. A mean roll of 1 can be achieved in only 1 way – to roll a 1 on both dice, therefore the probability of a mean of 1 is equal to $1/(6 \times 6) = 0.028$. A mean of 1.5 can be achieved in two ways: a 1 on the first die, followed by a 2 on the second die, or a 2 on the first die, followed by a 1 on the second die, given a probability of 0.056. We could continue with this for every possible mean score. If we were to do this, we would find the distribution shown in Figure 2. This graph is still not showing a normal distribution, but it is considerably closer to the normal distribution than Figure 1.

If I repeat the operation, with a sample size of 7, we can again calculate the probability of any value arising. For example, to achieve a mean score of 1, you would need to roll a 1 on all 7 dice. The probability of this occurring is equal to $1/6^7 = 0.0000036$ (around 1 in 280 000). In a similar way, we can plot the distribution of the measure – this is shown in Figure 3, along with a normal distribution curve with the same mean and standard deviation. It can be seen that the two curves are very close to each other (although they are not identical).

This demonstration has shown that the sampling distribution can indeed approach the normal distribution as the sample size increases, even though the distribution of the variable is not normal – it is uniform, and the sample size was only 7.

However, this variable was symmetrically distributed and that is not always going to be the case.

2 Central Limit Theory

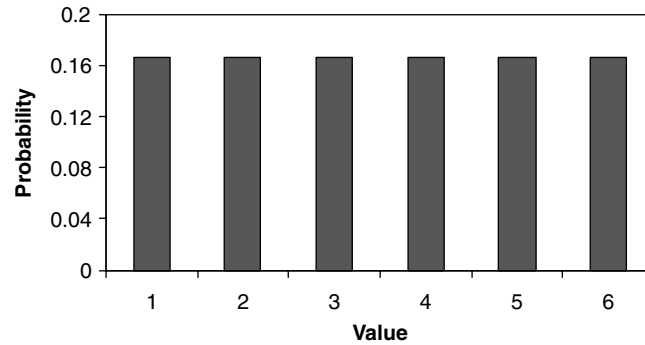


Figure 1 Probability distribution from samples of $N = 1$

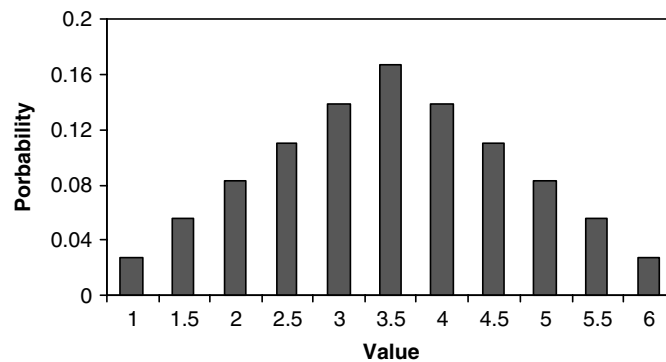


Figure 2 Probability distribution from samples of $N = 2$

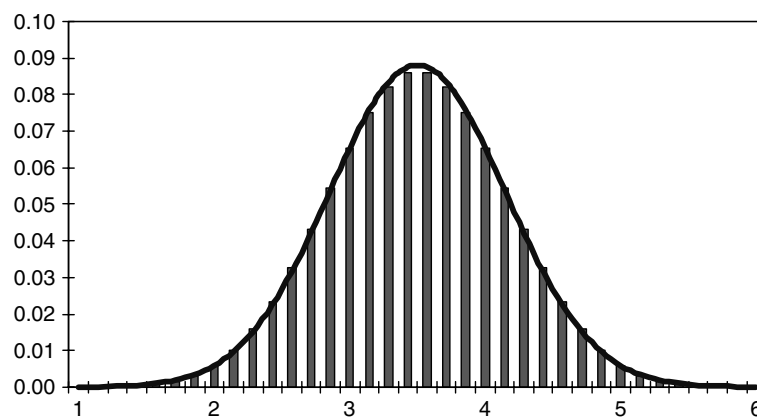


Figure 3 Distribution of mean score of 7 dice rolls (bars) and normal distribution with same mean and SD (line)

Imagine that we have a die, which has the original numbers removed, and the numbers 1, 1, 1, 2, 2, 3, added to it. The distribution of this measure in the sample is going to be markedly (positively) skewed, as is shown in Figure 4.

We might ask whether, in these circumstances, the central limit theorem still holds – we can see if the sampling distribution of the mean is normal. For one die, the distribution is markedly skewed, as we have seen. We can calculate the probability of different

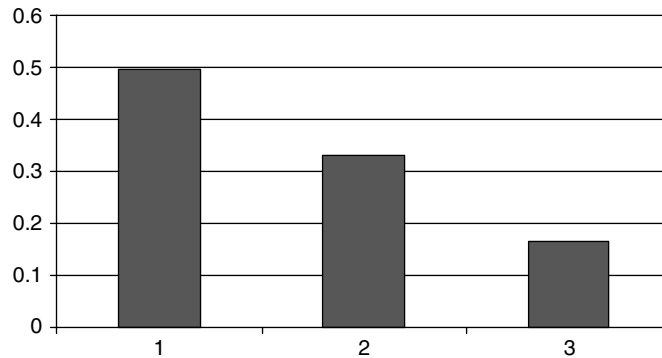


Figure 4 Distribution of die rolls, when die is marked 1, 1, 1, 2, 2, 3, $N = 1$

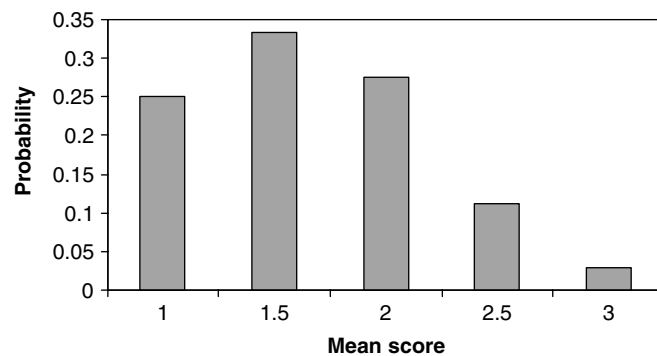


Figure 5 Sampling distribution of dice labelled 1, 1, 1, 2, 2, 3, when $N = 2$

values occurring in larger samples. When $N = 2$, a mean score of 1 can be achieved by rolling a 1 on both die. The probability of this event occurring = $0.5 \times 0.5 = 0.25$. We could continue to calculate the probability of each possible value occurring – these are shown in graphical form in Figure 5. Although we would not describe this distribution as normal, it is closer to a normal distribution than that shown in Figure 4.

Again, if we increase the sample size to 7 (still a very small sample), the distribution becomes a much better approximation to a normal distribution. This is shown in Figure 6.

When the Central Limit Theorem Goes Bad

As long as the sample is sufficiently large, it seems that we can rely on the central limit theorem to ensure

that the sampling distribution of the mean is normal. The usually stated definition of ‘sufficient’ is 30 (see, e.g., Howell [1]). However, this is dependent upon the shape of the distributions. Wilcox [4, 5] discusses a number of situations where, even with relatively large sample sizes, the central limit theorem fails to apply. In particular, the theorem is prone to letting us down when the distributions have heavy tails. This is likely to be the case when the data are derived from a mixed-normal, or contaminated, distribution.

A mixed-normal, or contaminated, distribution occurs when the population comprises of two or more groups, each of which has a different distribution (see **Finite Mixture Distributions**). For example, it may be the case that a measure has a different variance for males and for females, or for alcoholics and nonalcoholics. Wilcox [4] gives an example of a contaminated distribution with two groups. For both subgroups, the mean was equal to zero, for the larger group, which comprised 90% of the population,

4 Central Limit Theory

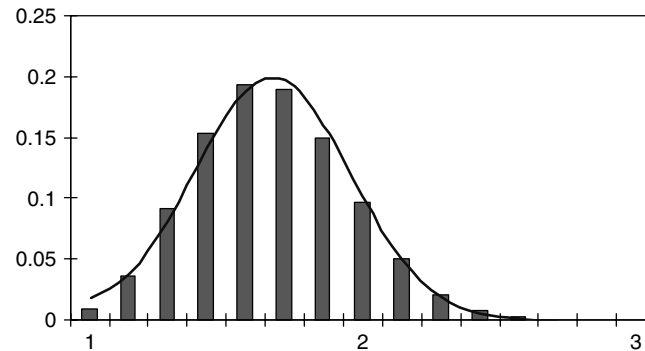


Figure 6 Sampling distribution of the mean, when $N = 7$, and dice are labelled 1, 1, 1, 2, 2, 3. Bars show sampling distribution, line shows normal distribution with same mean and standard deviation

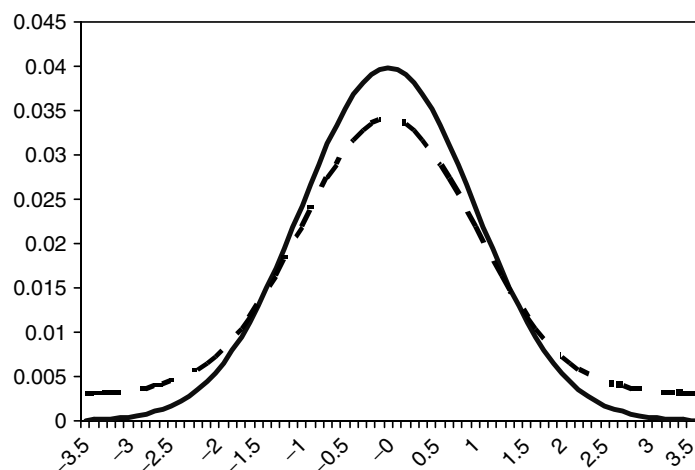


Figure 7 A normal and a mixed-normal (contaminated) distribution. The normal distribution (solid line) has mean = 0, SD = 1, the mixed normal has mean = 0, SD = 1 for 90% of the population, and mean = 0, SD = 10, for 10% of the population

the standard deviation was equal to 1, while for the smaller subgroup, which comprised 10% of the population, the standard deviation was equal to 10. The population distribution is shown in Figure 7, with a normal distribution for comparison. The shape of the distributions is similar, and probably would not give us cause for concern; however, we should note that the tails of the mixed distribution are heavier.

I generated 100 000 samples, of size 40, from this population. I calculated the mean for each sample, in order to estimate the sampling distribution of the mean. The distribution of the sample means is shown

in the left-hand side of Figure 8. Examining that graph ‘by eye’ one would probably say that it fairly closely approximated a normal distribution. On the right-hand side of Figure 8 the same distribution is redrawn, with a normal distribution curve, with the same mean and standard deviation. We can see that these distributions are different shapes – but is this enough to cause us problems?

The standard deviation of the sampling distribution of these means is equal to 0.48. The mean standard error that was calculated in each sample was 0.44. These values seem close to each other.

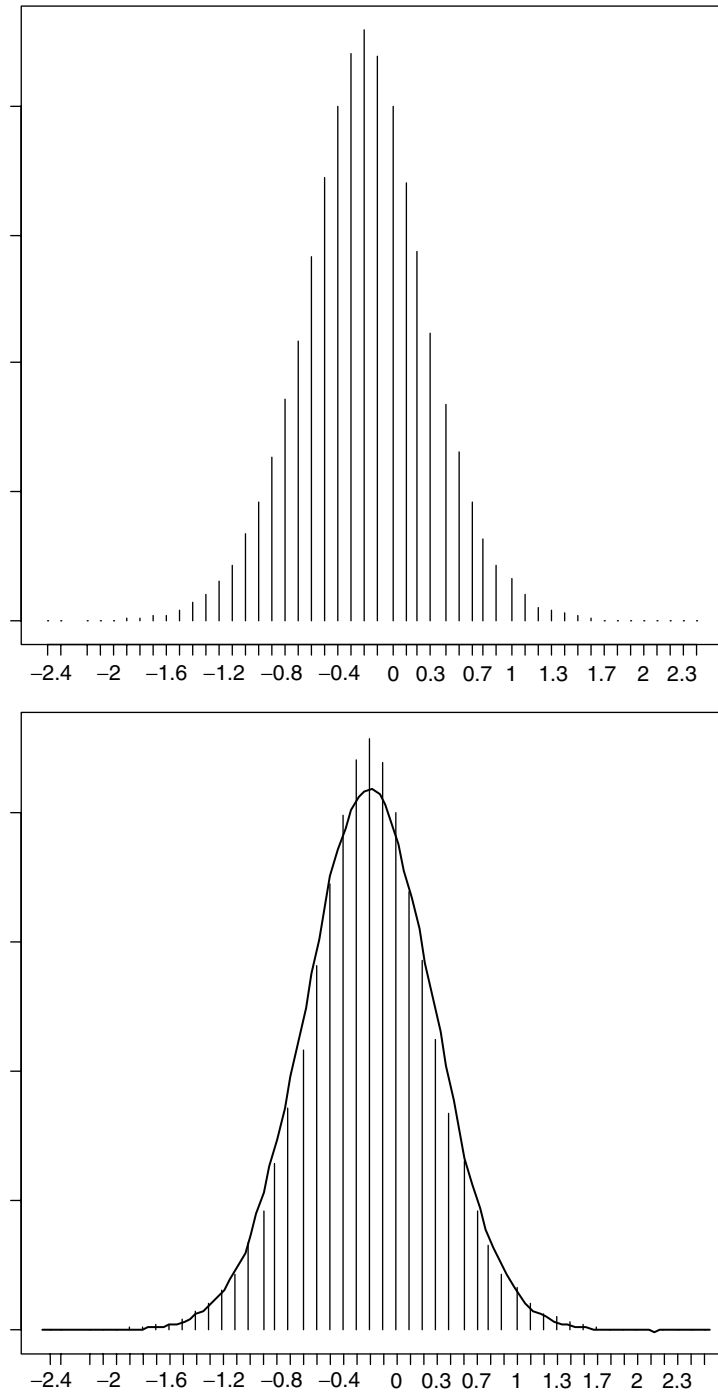


Figure 8 Sampling distribution of the mean, with 100 000 samples

6 Central Limit Theory

I also examined how often the 95% **confidence limits** in each sample excluded 0. According to the central limit theorem asymptotically, we would expect that 95% of the samples would have confidence limits that included zero. In this analysis, the figure was 97% – the overestimation of the standard error has caused our type I error rate to drop to 3%. This drop in type I error rate may not seem such a bad thing – we would all like fewer type I errors. However, it must be remembered that along with a drop in type I errors, we must have an increase, of unknown proportion, in type II errors, and hence a decrease in power (*see Power*). (For further discussion of these issues, and possible solutions, see entry on robust statistics.)

Appendix

The following is an R program for carrying out a small Monte Carlo simulation to examine the effect of contaminated distributions on the sampling distribution of the mean. (R is available for most

computer platforms, and can be freely downloaded from www.r-project.org).

Note that any text following a # symbol is a comment and will be ignored by the program; the <– symbol is the assignment operator – to make the variable ‘x’ equal to 3, I would use `x <– 3`.

The program will produce the graph shown in Figure 8, as well as the proportion of the 95% CIs which include the population value (of zero).

Four values are changeable – the first two lines give the SDs for two groups – in the example that was used in the text, the values were 1 and 10.

The third line is used to give the proportion of people in the population in the group with the higher standard deviation – in the example we used 0.1–10%.

Finally, in lines 4 and 5 the sample size and the number of samples to be drawn from the population are given. Again, these are the same as in the example.

```
lowerSD <- 1           #set SD for group with lower SD
upperSD <- 10          #set SD for group with higher SD
proportionInUpper = 0.1 #set proportion of people in group with higher
sampleSize <- 40        #set size of samples
nSamples <- 100000      #set number of samples
#generate data-one variable of length number of samples * sample size.
data <- rnorm(sampleSize * nSamples, 0, lowerSD)+rnorm(sampleSize * nSamples, 0,
  (upperSD-lowerSD)) * rbinom(n=sampleSize * nSamples, size=1,
  p=proportionInUpper)

#Dimension the data to break it into samples
dim(data) <- c(nSamples, sampleSize)

#calculate the mean in each sample
sampleMeans <- rowSums(data) / sampleSize

#generate the labels for the histogram
xAxisLabels <- seq(min(sampleMeans), max(sampleMeans), length=100)
xAxisLabels <- xAxisLabels - mean(sampleMeans) / sd(sampleMeans)

#generate true normal distribution, with same mean and SD as sampling
  distribution, for comparison
NormalHeights <- dnorm(xAxisLabels, mean=mean(sampleMeans), sd=sd(sampleMeans))

#draw histogram, with normal distribution for comparison.
#hist(sampleMeans, probability=TRUE)
heights <- table(round(sampleMeans, 2))
plot(table(round(sampleMeans, 1)))
lines(xAxisLabels, NormalHeights * (nSamples / 10), col="red")
```

```
#Calculate 95% CIs of mean for each sample.
dataSquared <- data^2
dataRowSums <- rowSums(data)
dataRowSumsSquared <- rowSums(dataSquared)
sampleSD <- sqrt((sampleSize*dataRowSumsSquared - dataRowSums^2) / (sampleSize *
  (sampleSize - 1)))
sampleSE <- sampleSD/sqrt(sampleSize)
lowerCIs <- sampleMeans - (sampleSE*(qt(0.975,sampleSize - 1)))
upperCIs <- sampleMeans + (sampleSE*(qt(0.975,sampleSize - 1)))

#check to see if 95% CIs include 0
within95CIs <- lowerCIs < 0 & upperCIs > 0

#draw table of means.
table(within95CIs)
```

The figures shown in the small table earlier refer to the number of samples in which the 95% CIs included the population value of zero. In this run, 96 860 samples (from 100 000, 96.9%) included zero, 3140 (from 100 000, 3.1%) did not. In a normal distribution, and extremely large sample, these values would be 95% and 5%. It may be a worthwhile exercise to set the SDs to be equal, and check to see if this is the case.

To run the program, paste the text into R.

As well as the graph, the output will show a small table:

```
FALSE TRUE
3140 96860
```

References

- [1] Howell, D.C. (2002). *Statistical Methods for Psychology*, 5th Edition, Duxbury Press, Belmont.

- [2] Mood, A., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Edition, McGraw-Hill, New York.
- [3] Roberts, M.J. & Russo, R. (1999). *A Student's Guide to Analysis of Variance*, Routledge, London.
- [4] Wilcox, R.R. (1996). *Statistics for the Social Sciences*, Academic Press, London.
- [5] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, London.

JEREMY MILES

Children of Twins Design

BRIAN M. D'ONOFRIO

Volume 1, pp. 256–258

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Children of Twins Design

One of the first lessons in all statistics classes for the social sciences is that correlation does not mean causation (*see* **Correlation Issues in Genetics Research**). In spite of this premise, most disciplines have historically assumed that associations between parental and offspring characteristics are due to direct environmental causation [15]. However, third variables, or selection factors, may influence both generations and account for the intergenerational relations. Social scientists typically try to statistically control for environmental selection factors in their studies, but unmeasured confounds may also influence the associations. These cannot be controlled statistically. However, unmeasured selection factors can be taken into account by carefully selecting comparison groups, thus moving researchers closer to being able to make causal statements.

Genetic factors are critical to take into account in intergenerational studies, because parents provide the environmental context and transmit their genes to their offspring. Therefore, any statistical association between parents and children may also be due to similar genetic backgrounds. Genetic confounds in intergenerational associations are referred to as passive **gene-environment correlation (passive rGE)**. Passive rGE occurs when common genotypic factors influence parental behaviors, which are considered to be environmental risk factors, and child outcomes [14, 18]. Although genetic confounds render most typical research, such as family studies, uninterruptible, researchers have largely ignored the role of passive rGE [16].

There are a number of behavior genetic designs that delineate between the genetic and environment processes that are responsible for relations between parental characteristics and child outcomes [3, 7, 17]. The most well-known genetically informed design is the **adoption study** because of the clear separation between the genetic risk transmitted to the child by the biological parents and the environment that is provided by the adopting family. However, the adoption study is becoming increasingly difficult to conduct and suffers from a number of methodological assumptions and limitations [17]. The **co-twin control design** can also help separate genetic and environmental processes through which environmental risk factors influence one's behavior, but the

design cannot include risk factors that both twins share [3]. Therefore, other behavior genetic designs are necessary.

The Children of Twins (CoT) Design is a genetically informed approach that also explores the association between environmental risk factors and outcomes in offspring. The offspring of identical twins have the same genetic relationship with their parents and their parents' cotwin because each twin is genetically the same. However, only one twin provides the environment for his/her children. As a result, the genetic risk associated with the parental behavior can be inferred from the cotwin [7, 13]. Using children of identical twins can determine if a parental characteristic has an environmental association with a child behavior or whether the intergenerational relation is confounded by selection factors. When children of fraternal twins are included, the design is able to reveal whether confounds are genetic or environmental in origin [5].

The CoT Design is best known for its use with studying dichotomous environmental risk factors, such as a diagnosis of a psychiatric disorder [6]. For example, the children of schizophrenic parents are at higher risk for developing the disorder than the general population. In order to elucidate the genetic and environmental mechanisms responsible for the intergenerational association, researchers compare the rates of schizophrenia in the offspring of discordant pairs of twins (one twin is diagnosed with the disorder and one is not). A comparison between the children of affected (diagnosed with schizophrenia) identical twins and their unaffected (no diagnosis) cotwins is the initial step in trying to understand the processes through which the intergenerational risk is mediated. Because offspring of both identical twins share the same genetic risk associated with the parental psychopathology from the twins, any difference between the offspring is associated with environmental processes specifically related to the parental psychopathology (see below for a discussion of the influence of the nontwin parent). Effectively, the CoT Design provides the best control comparison group because children with schizophrenia are compared with their cousins who share the same genetic risk associated with schizophrenia and any environmental conditions that the twins share. If offspring from the unaffected identical twin have a lower prevalence of schizophrenia than offspring of the affected identical twin, the results would suggest that

2 Children of Twins Design

the experience of having schizophrenic parent has a direct environmental impact on one's own risk for schizophrenia.

If the rates of the disorder in the offspring of the affected and unaffected identical cotwins are equal to each other, the direct causal role of the parental psychopathology would be undermined. However, such results do not elucidate whether shared genetic or environmental processes are responsible for the intergenerational transmission. A comparison of the rates of psychopathology in the children of the unaffected identical and fraternal cotwins highlights the nature of the selection factors. Children of the unaffected identical twins only vary with respect to the environmental risk associated with schizophrenia, whereas offspring of the unaffected fraternal twin differs with respect to the environmental *and* genetic risk (lower). Therefore, higher rates of schizophrenia in the children of the unaffected identical cotwins than in children of the unaffected fraternal cotwins suggest that genetic factors account for some of the intergenerational covariation. If the rates are similar for the children in unaffected identical and fraternal families, shared environmental factors would be of most import because differences in the level of genetic risk would not influence the rate of schizophrenia.

The most well-known application of the design explored the intergenerational association of schizophrenia using discordant twins [6]. Offspring of schizophrenic identical cotwins had a morbid risk of being diagnosed with schizophrenia of 16.8, whereas offspring of the unaffected identical cotwins had a morbid risk of 17.4. Although the offspring in this later group did not have a parent with schizophrenia, they had the same risk as offspring with a schizophrenic parent. The results effectively discount the direct causal environmental theory of schizophrenia transmission. The risk in the offspring of the unaffected identical twins was 17.4, but the risk was much lower (2.1) in the offspring of the unaffected fraternal cotwins. This latter comparison suggests that genetic factors account for the association between parental and offspring schizophrenia. Similar findings were reported for the transmission of bipolar depression [1]. In contrast, the use of the CoT to explore transmission of alcohol abuse and dependence from parents to their offspring highlighted role of the family environment [8].

One of the main strengths of the design is its ability to study different phenotypes in the parent and

child generations. For example, a study of divorce using the CoT Design reported results consistent with a direct environmental causal connection between parental marital instability and young-adult behavior and substance abuse problems [4]. Similar conclusions were found with CoT Design studies of the association between harsh parenting and child behavior problems [9] and between smoking during pregnancy and child birth weight [3, 11]. However, a CoT analysis found that selection factors accounted for the lower age of menarche in girls growing up in households with a stepfather, results that suggest the statistical association is not a causal relation [12]. These findings suggest that underlying processes in intergenerational associations are dependent on the family risk factors and outcomes in the offspring.

In summary, selection factors hinder all family studies that explore the association between risk factors and child outcomes. Without the ability to experimentally assign children to different conditions, researchers are unable to determine whether differences among groups (e.g., children from intact versus divorced families) are due to the measured risk factor or unmeasured differences between families. Because selection factors may be environmental or genetic in origin, researchers need to use quasi-experimental designs that pull apart the co-occurring genetic and environmental risk processes [17]. The CoT Design is a behavior genetic approach that can explore intergenerational associations with limited methodological assumptions compared to other designs [3]. However, caution must be used when interpreting the result of studies using the CoT Design. Similar to all nonexperimental studies, the design cannot definitely prove causation. The results can only be consistent with a causal hypothesis because environmental processes that are correlated with the risk factor and only influence one twin and their offspring may actually be responsible for the associations.

The CoT Design can also be expanded in a number of ways. The design can include continuously distributed risk factors [3, 7] and measures of family level environments. Associations between parental characteristics and child outcomes may also be due to reverse causation, but given certain assumptions, the CoT Design can delineate between parent-to-child and child-to-parent processes [19]. Because the design is also a quasi-adoption study, the differences in genetic and environmental risk in the approach

provides the opportunity to **gene-environment interaction** [8]. When the spouses of the adult twins are included in the design, the role of **assortative mating** and the influence of both spouses can be considered, an important consideration for accurately describing the processes involved in the intergenerational associations [7]. Finally, the CoT Design can be combined with other behavior genetic designs to test more complex models of parent-child relations [2, 10, 20]. Overall, the CoT Design is an important genetically informed methodology that will continue to highlight the mechanisms through which environmental and genetic factors act and interact.

References

- [1] Bertelsen, A. & Gottesman, I.I. (1986). Offspring of twin pairs discordant for psychiatric illness, *Acta Geneticae Medicae et Gemellologia* **35**, 110.
- [2] D'Onofrio, B.M., Eaves, L.J., Murrelle, L., Maes, H.H. & Spilka, B. (1999). Understanding biological and social influences on religious affiliation, attitudes and behaviors: a behavior-genetic perspective, *Journal of Personality* **67**, 953–984.
- [3] D'Onofrio, B.M., Turkheimer, E., Eaves, L.J., Corey, L.A., Berg, K., Solaas, M.H. & Emery, R.E. (2003). The role of the children of twins design in elucidating causal relations between parent characteristics and child outcomes, *Journal of Child Psychology and Psychiatry* **44**, 1130–1144.
- [4] D'Onofrio, B.M., Turkheimer, E., Emery, R.E., Slutske, W., Heath, A., Madden, P. & Martin, N. (submitted). A genetically informed study of marital instability and offspring psychopathology, *Journal of Abnormal Psychology*.
- [5] Eaves, L.J., Last, L.A., Young, P.A. & Martin, N.B. (1978). Model-fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.
- [6] Gottesman, I.I. & Bertelsen, A. (1989). Confirming unexpressed genotypes for schizophrenia, *Archives of General Psychiatry* **46**, 867–872.
- [7] Heath, A.C., Kendler, K.S., Eaves, L.J. & Markell, D. (1985). The resolution of cultural and biological inheritance: informativeness of different relationships, *Behavior Genetics* **15**, 439–465.
- [8] Jacob, T., Waterman, B., Heath, A., True, W., Buchholz, K.K., Haber, R., Scherrer, J. & Quiang, F. (2003). Genetic and environmental effects on offspring alcoholism: new insights using an offspring-of-twins design, *Archives of General Psychiatry* **60**, 1265–1272.
- [9] Lynch, S.K., Turkheimer, E., Emery, R.E., D'Onofrio, B.M., Mendle, J., Slutske, W. & Martin, N.G. (submitted). A genetically informed study of the association between harsh punishment and offspring behavioral problems.
- [10] Maes, H.M., Neale, M.C. & Eaves, L.J. (1997). Genetic and environmental factors in relative body weight and human adiposity, *Behavior Genetics* **27**, 325–351.
- [11] Magnus, P., Berg, K., Bjerkedal, T. & Nance, W.E. (1985). The heritability of smoking behaviour in pregnancy, and the birth weights of offspring of smoking-discordant twins, *Scandinavian Journal of Social Medicine* **13**, 29–34.
- [12] Mendle, J., Turkheimer, E., D'Onofrio, B.M., Lynch, S.K. & Emery, R.E. (submitted). Stepfather presence and age at menarche: a children of twins approach.
- [13] Nance, W.E. & Corey, L.A. (1976). Genetic models for the analysis of data from the families of identical twins, *Genetics* **83**, 811–826.
- [14] Plomin, R., DeFries, J.C. & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior, *Psychological Bulletin* **84**, 309–322.
- [15] Rutter, M. (2000). Psychosocial influences: critiques, findings, and research needs, *Development and Psychopathology* **12**, 375–405.
- [16] Rutter, M., Dunn, J., Plomin, R., Simonoff, E., Pickles, A., Maughan, B., Ormel, J., Meyer, J. & Eaves, L. (1997). Integrating nature and nurture: implications of person-environment correlations and interactions for developmental psychopathology, *Development and Psychopathology* **9**, 335–364.
- [17] Rutter, M., Pickles, A., Murray, R. & Eaves, L.J. (2001). Testing hypotheses on specific environmental causal effects on behavior, *Psychological Bulletin* **127**, 291–324.
- [18] Scarr, S. & McCartney, K. (1983). How people make their own environments: a theory of genotype-environment effects, *Child Development* **54**, 424–435.
- [19] Silberg, J.L. & Eaves, L.J. (2004). Analyzing the contribution of genes and parent-child interaction to childhood behavioural and emotional problems: a model for the children of twins, *Psychological Medicine* **34**, 347–356.
- [20] Truett, K.R., Eaves, L.J., Walters, E.E., Heath, A.C., Hewitt, J.K., Meyer, J.M., Silberg, J., Neale, M.C., Martin, N.G. & Kendler, K.S. (1994). A model system for analysis of family resemblance in extended kinships of twins, *Behavior Genetics* **24**, 35–49.

BRIAN M. D'ONOFRIO

Chi-Square Decomposition

DAVID RINDSKOPF

Volume 1, pp. 258–262

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Chi-Square Decomposition

When analyzing the relationship between two categorical predictors, traditional methods allow one of two decisions: Either the two variables are related or they are not. If they are related, the researcher is not led to further consider the nature of the relationship. This contrasts with the case of a continuous outcome variable, where a significant F in an **analysis of variance (ANOVA)** would be followed up by post hoc tests (*see Multiple Comparison Procedures*). Similarly, with an ANOVA one can do preplanned comparisons if there are hypotheses about where group differences lie. For categorical data, the analogous method is the partition (or decomposition) of chi-square.

Example 1 Relationship Between Two Variables. As a first example, consider the data, reprinted here as Table 1, on race and party identification from the entry **contingency tables** in this encyclopedia. A test of independence reported in that entry led to the conclusion that race and party identification were related; examination of residuals indicated an excess (compared to what would be expected if race and party were independent) of Blacks who were Democrat, and Whites who were Republican. To examine the nature of the relationship using a partition of chi-square, we first calculate the row proportions, reported here in Table 2.

These results are consistent with the previous findings based on residual analysis. The proportion of Blacks who are Democrat is much larger than the proportion of Whites who are Democrat, and the proportion of Whites who are Republican is much larger than the proportion of Blacks who are Republican. Notice also that the proportion of

Table 1 Party identification and race (from the 1991 General Social Survey)

Race	Party identification		
	Democrat	Independent	Republican
White	341	105	405
Black	103	15	11

Table 2 Proportions of each race who have each party identification

Race	Party identification		
	Democrat	Independent	Republican
White	0.401	0.123	0.476
Black	0.798	0.116	0.085

each race who are Independent seems to be about equal.

Although the results here are so obvious that a formal partition of chi-square is somewhat superfluous, I will use this as a simple introduction to the technique. The first step is to introduce another measure used to test models in contingency tables, the likelihood ratio chi-square (*see Contingency Tables*). In the entry on contingency tables, the expected frequencies from the model of independence were compared to the observed frequencies using a measure often represented as X^2 . Here we will call this the Pearson fit statistic. The likelihood ratio fit statistic also compares the observed to the expected frequencies, but using a slightly less obvious formula $G^2 = 2 \sum_i O_i \ln(O_i/E_i)$, where the index i indicates the cell of the table. (The use of a single subscript allows this to apply to tables of any size, including nonrectangular tables.) The utility of this measure of fit will be seen shortly. For the race and party identification data, $G^2 = 90.331$ with 2 degrees of freedom (df); the Pearson statistic was 79.431.

To return to the data, we might hypothesize that (a) Blacks and Whites do not differ in the proportion who register as Independents, and (b) among those who have a party affiliation, Whites are more likely to be Republican than are Blacks. In making these hypotheses, we are dividing the two degrees of freedom for the test of independence into two one-degree-of-freedom parts; we will therefore be partitioning (decomposing) the original fit statistic into two components, each of which corresponds to one of these hypotheses.

The first hypothesis can be tested by using the data in Table 3, where the original three categories of party identification are collapsed into two categories: Independents, and members of a major party.

The value of G^2 for this table is .053 (which is the same as the value of X^2 ; this is often the case

2 Chi-Square Decomposition

Table 3 Race by party identification (independent versus major party)

Race	Party identification	
	Democrat + republican	Independent
White	746	105
Black	114	15

for models that fit well); with 1 df, it is obvious that there is no evidence for a relationship, so we can say that the first hypothesis is confirmed.

We now use the data only on those who are members of a major party to answer the second question. The frequencies are reproduced in Table 4.

Here we find that $G^2 = 90.278$, and $X^2 = 78.908$, each with 1 df. Independence is rejected, and we conclude that among those registered to a major party, there is a large difference between Blacks and Whites in party registration.

Why is this called a partition of chi-square? If we add the likelihood ratio statistics for each of the one df hypotheses, we get $0.053 + 90.278 = 90.331$; this is equal to the likelihood ratio statistic for the full 3×2 table. (The same does not happen for the Pearson statistics; they are still valid tests of these hypotheses, but the results aren't as 'pretty' because they do not produce an exact partition.) We can split the overall fit for the model of independence into parts, testing hypotheses within segments of the overall table.

To get a partition, we must be careful of how we select these hypotheses. Detailed explanations can be found in [3], but a simple general rule is that they correspond to selecting orthogonal contrasts in an ANOVA. For example, the contrast coefficients for testing the first hypothesis (independents versus major party) would be $(-1, 2, -1)$, and the coefficients for comparing Democrats to Republicans would be $(1, 0, -1)$. These two sets of coefficients are orthogonal.

Table 4 Race by party identification (Major party members only)

Race	Party identification	
	Democrat	Republican
White	341	405
Black	103	11

Table 5 Depression in adolescents, by age and sex (Seriously emotionally disturbed only)

Age	Sex	Depression		$P(\text{High})$
		Low	High	
12–14	Male	14	5	.26
	Female	5	8	.62
15–16	Male	32	3	.09
	Female	15	7	.32
17–18	Male	36	5	.12
	Female	12	2	.14

Example 2 Relationships Among Three Variables. Although partitioning is usually applied to two-way tables, it can also be applied to tables of higher dimension. To illustrate a more complex partitioning, I will use some of the data on depression in adolescents from Table 3 of the entry on contingency tables. In that example, there were two groups; I will use only the data on children classified as SED (seriously emotionally disturbed). The remaining variables (age, sex, and depression) give a $3 \times 2 \times 2$ table; my conceptualization will treat age and sex as predictors, and depression as an outcome variable. The data are reproduced in Table 5.

If we consider the two predictors, there are six Age $\times$ Sex groups that form the rows of the table. Therefore, we can think of this as a 6×2 table instead of as a $3 \times 2 \times 2$ table. If we test for independence between the (six) rows and the (two) columns, the likelihood ratio statistic is 18.272 with 5 df. So there is some evidence of a relationship between group and depression. But what is the nature of this relationship? We will start by breaking the overall table into three parts, testing the following three hypotheses:

- (1) the depression rate for males is constant; that is, it does not change with age;
- (2) the depression rate for females is constant;
- (3) the rate of depression in males is the same as that for females.

To test hypothesis (1), we consider Table 6, which contains only the males.

For this table, a test of independence results in a value $G^2 = 3.065$, with 2 df. This is consistent with the hypothesis that the proportion depressed among males does not change with age.

Table 6 Depression in adolescents (Males only)

Age	Depression		<i>P</i> (High)
	Low	High	
12–14	14	5	.26
15–16	32	3	.09
17–18	36	5	.12

Table 7 Depression in adolescents (Females only)

Age	Depression		<i>P</i> (High)
	Low	High	
12–14	5	8	.62
15–16	15	7	.32
17–18	12	2	.14

To test hypothesis (2) we consider Table 7, which is identical in structure to Table 6, but contains only females.

The test of independence here has $G^2 = 6.934$, $df = 2$; this is just significant at the 0.05 level, but the relationship between age and depression is weak. We can further partition the chi-square for this table into two parts. It appears that the highest rate of depression by far among the females is in the youngest age group. Therefore, we will determine whether the 15–16 year-olds have the same level of depression as the 17–18 year-olds, and whether the youngest group differs from the older groups.

To compare 15–16 year-olds with 17–18 year-olds we look only at the last two rows of Table 7, and see whether depression is related to age if we use only those two groups. The value of G^2 is 1.48 with 1 df , so there is no difference in the rate of depression between the two groups. Next we combine these two groups and compare them to the youngest group. The value of G^2 for this table is 5.451 with 1 df , showing that the source of the Age $\times$ Depression relationship in girls is due solely to the youngest group having a higher rate of depression than the two older groups.

To test hypothesis (3), we use the data in Table 8, which is collapsed over age. A test here has $G^2 = 8.27$ with 1 df , so we clearly reject the notion that males and females have equal chances of being depressed. From the proportions in the table, we see that females are more frequently at the high level of depression than males.

Table 8 Depression in adolescents (Collapsed over age)

Sex	Depression		<i>P</i> (High)
	Low	High	
Male	82	13	.137
Female	32	17	.347

To summarize the analyses of these data, the rate of depression in males is constant across the ages measured in this study, but for females there is a somewhat higher rate in the youngest group compared to the two older groups. Finally, males have a lower rate than do females.

Some History and Conclusions

The use of partitioning goes back at least to **Fisher** [1], but was never generally adopted by researchers. Later, partitioning of chi-square was mentioned sporadically through the literature, but often in a mechanical way rather than in a way that corresponded to testing hypotheses of interest to researchers. Further, attempts were made to try to adjust the Pearson chi-square tests so that partitioning was exact; this was unnecessary because (a) the likelihood ratio test will partition exactly, and (b) even without an exact partition, the Pearson fit statistic tests the right hypothesis, and is a valid tool.

Partitioning has an advantage of being extremely simple to understand and implement; the main disadvantage is that some hypotheses cannot easily be tested within this framework. For example, the data on girls in Table 7 indicates the proportion depressed may not change abruptly with age, but instead might decline steadily with age. To test this requires a more general technique called nonstandard **loglinear models** (see e.g., [2]). This technique would also allow us to test all of the hypotheses of interest in one model. A final issue is whether (and how) to adjust probability levels if these techniques are used in a post hoc fashion (see **Multiple Comparison Procedures**). The most frequent suggestion is to use an adjustment similar to a Scheffé test; this would be equivalent to using the critical value for the table as a whole when evaluating any of the subtables in a partition. As with the use of the Scheffé procedure in ANOVA, this is a very conservative approach.

In spite of its limitations, partitioning chi-square allows researchers to test in a clear and simple way

4 Chi-Square Decomposition

many substantive research hypotheses that cannot be tested using traditional methods. For this reason, it deserves to be in the arsenal of all researchers analyzing categorical data.

References

- [1] Fisher, R.A. (1930). *Statistical Methods for Research Workers*, 3rd Edition, Oliver and Boyd, Edinburgh.

- [2] Rindskopf, D. (1990). Nonstandard loglinear models, *Psychological Bulletin* **108**, 150–162.

- [3] Rindskopf, D. (1996). Partitioning chi-square, in *Categorical Variables in Developmental Research*, C.C. Clogg & A. von Eye, eds, Academic Press, San Diego.

(See also **Log-linear Models**)

DAVID RINDSKOPF

Cholesky Decomposition

STACEY S. CHERNY

Volume 1, pp. 262–263

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cholesky Decomposition

The Cholesky decomposition is a model with as many latent factors as there are observed measures, with no measure-specific variances. The first factor loads on all measures, the second factor loads on all but the first measure, the third factor loads on all but the first two measures, continuing down to the last factor, which loads on only the last measure. This model has the advantage of being of full rank; that is, there are as many parameters being estimated as there are data points. Therefore, a perfect fit to the data will result.

The parameters of the model are contained in the matrix Λ , where,

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \end{pmatrix}. \quad (1)$$

The parameters are estimated using **direct maximum-likelihood estimation** for the analysis of raw data. The likelihood function, which is maximized with respect to the Cholesky parameters, is

$$LL = \sum_{i=1}^N \left(-\frac{1}{2} \ln |\Sigma_i| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \Sigma_i^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right), \quad (2)$$

where

- $\mathbf{x}_i$ = vector of scores for a given case (e.g., an individual or twin pair or family)
- Σ_i = appropriate expected covariance matrix for that case
- N = total number of case
- $\boldsymbol{\mu}$ = vector of estimated means

and where

$$2(LL_1 - LL_2) = \chi^2 \quad (3)$$

for testing the difference between two alternative models.

This model can be extended to genetically informative data, such as those obtained from twin pairs, where instead of simply estimating phenotypic covariance structure, we can partition the variance and covariance structure among a twin pair

into genetic ($\mathbf{G}$) and shared ($\mathbf{C}$) and nonshared ($\mathbf{E}$) environmental influences and use the Cholesky decomposition to estimate those partitioned covariance matrices:

$$\Sigma = \begin{pmatrix} \mathbf{G} + \mathbf{C} + \mathbf{E} & r\mathbf{G} + \mathbf{C} \\ r\mathbf{G} + \mathbf{C} & \mathbf{G} + \mathbf{C} + \mathbf{E} \end{pmatrix}, \quad (4)$$

where the top left and bottom right submatrices represent covariances among the multiple measures made within the first and second member of the twin pair, respectively, and the top right and bottom left submatrices represent covariances between those measures taken on twin 1 and with those taken on twin 2. In the above equation, r is 1 for MZ twin pairs and 1/2 for DZ twin pairs. Each quadrant is a function of genetic ($\mathbf{G}$), shared environmental ($\mathbf{C}$) and nonshared environmental ($\mathbf{E}$) covariance matrices, each estimated using a Cholesky decomposition involving $\Lambda_G \Lambda'_G$, $\Lambda_C \Lambda'_C$, and $\Lambda_E \Lambda'_E$, respectively.

Since the maximum-likelihood estimation procedures are performed on unstandardized data, the resulting expected covariance matrices are typically standardized for ease of interpretability. Each of the genetic, shared environmental, and unique environmental covariance matrices, $\mathbf{G}$, $\mathbf{C}$, and $\mathbf{E}$ can be standardized as follows:

$$\mathbf{G}^* = \mathbf{hR}_G \mathbf{h} = \text{diag}(\Sigma_P)^{-\frac{1}{2}} \mathbf{G} \text{diag}(\Sigma_P)^{-\frac{1}{2}} \quad (5)$$

$$\mathbf{C}^* = \mathbf{cR}_C \mathbf{c} = \text{diag}(\Sigma_P)^{-\frac{1}{2}} \mathbf{C} \text{diag}(\Sigma_P)^{-\frac{1}{2}} \quad (6)$$

$$\mathbf{E}^* = \mathbf{eR}_E \mathbf{e} = \text{diag}(\Sigma_P)^{-\frac{1}{2}} \mathbf{E} \text{diag}(\Sigma_P)^{-\frac{1}{2}}, \quad (7)$$

where Σ_P is the expected phenotypic covariance matrix, simply the upper left or bottom right quadrants of either the expected MZ or DZ covariance matrices (since all those quadrants are expected to be equal).

This results in a partitioning of the phenotypic correlation structure among multiple measures, such that these three components, when summed, equal the expected phenotypic correlation matrix. The diagonal elements of the standardized matrices ($\mathbf{G}^*$, $\mathbf{C}^*$, and $\mathbf{E}^*$) contain the heritabilities (*see Heritability*) and proportions of shared and nonshared environmental variance, respectively, for each of the measures. The off-diagonal elements contain what are commonly referred to as the phenotypically standardized genetic

2 Cholesky Decomposition

and environmental covariances. As shown, $\mathbf{G}^*$ can be decomposed into the genetic correlation matrix, $\mathbf{R}_G$, pre- and postmultiplied by the square roots of the heritabilities in a diagonal matrix, $\mathbf{h}$. This

can similarly be done for the shared and nonshared environmental components.

STACEY S. CHERNY

Classical Statistical Inference Extended: Split-Tailed Tests

RICHARD J. HARRIS

Volume 1, pp. 263–268

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Classical Statistical Inference Extended: Split-Tailed Tests

As pointed out in the entry on **classical statistical inference: practice versus presentation**, classical statistical inference (CSI) as practiced by most researchers is a very useful tool in determining whether sufficient evidence has been marshaled to establish the direction of a population effect. However, classic two-tailed tests, which divide alpha evenly between the rejection regions corresponding to positively and negatively signed values of the population effect, do not take into account the possibility that logic, expert opinion, existing theories, and/or previous empirical studies of this same population effect might strongly (or weakly, for that matter) favor the hypothesis that θ (the population parameter under investigation) is greater than θ_0 (the highly salient dividing line – often zero – between positive and negative population effects) over the alternative hypothesis that $\theta < \theta_0$, or vice versa. The *split-tailed test*, introduced (but not so labeled) by Kaiser [4] and introduced formally by Braver [1], provides a way to incorporate prior evidence into one's significance test so as to increase the likelihood of finding statistical significance in the correct direction (provided that the researchers' assessment of the evidence has indeed pointed her in the correct direction).

The remainder of this article describes the decision rules (DRs) employed in conducting split-tailed significance tests and constructing corresponding confidence intervals. It also points out that the classic one-tailed test (100% of α devoted to the predicted tail of the sampling distribution) is simply an infinitely biased split-tailed test.

Decision Rule(s) for Split-tailed Tests

Split-tailed Statistical Inference (Correlation Coefficient Example)

We test $H_0: \rho_{XY} = 0$ against $H_1: \rho_{XY} > 0$ and $H_2: \rho_{XY} < 0$

by comparing $p_{>} = \Pr(r_{XY} > r_{XY}^* \text{ if } H_0 \text{ were true})$ to $\alpha_{>}$

$$\begin{aligned} \text{and } p_{<} &= \Pr(r_{XY} < r_{XY}^* \text{ if } H_0 \text{ were true}) \\ &= 1 - p_{>} \text{ to } \alpha_{<}. \end{aligned}$$

where r_{XY}^* is the observed value of the correlation coefficient calculated for your sample;

r_{XY} is a random variable representing the values of the sample correlation coefficient obtained from an infinitely large number of independent random samples from a population in which the true population correlation coefficient is precisely zero;

$p_{>}$ is, as the formula indicates, the percentage of the correlation coefficients computed from independent random samples drawn from a population in which $\rho_{XY} = 0$ that are as large as or larger than the one you got for your single sample from that population;

$p_{<}$ is defined analogously as $\Pr(r_{XY} < r_{XY}^* \text{ if } H_0 \text{ were true})$;

$\alpha_{>}$ is the criterion you have set (*before* examining your data) as to how low $p_{>}$ has to be to convince you that the population correlation is, indeed, positive (high values of Y tending to go along with high values of X and low, with low); and

$\alpha_{<}$ is the criterion you have set as to how low $p_{<}$ has to be to convince you that the population correlation is, indeed, negative.

These two comparisons ($p_{<}$ to $\alpha_{<}$ and $p_{>}$ to $\alpha_{>}$) are then used to decide between H_1 and H_2 , as follows:

Decision rule: If $p_{>} < \alpha_{>}$, conclude that (or, at least for the duration of this study, act as if)

H_1 is true (i.e., accept the hypothesis that $\rho_{XY} > 0$).

If $p_{<} < \alpha_{<}$, conclude that (or act as if)

H_2 is true (i.e., accept H_2 that $\rho_{XY} < 0$).

If neither of the above is true (i.e., if $p_{>} \geq \alpha_{>}$ and $p_{<} \geq \alpha_{<}$, which is equivalent to the condition that $p_{>}$ be greater than $\alpha_{>}$ but less than $1 - \alpha_{<}$), conclude that we do not have enough evidence to decide between H_1 and H_2 (i.e., fail to reject H_0).

Split-tailed Statistical Inference (about Other Population Parameters)

As is true for classic (equal-tailed) significance tests, the same basic logic holds for split-tailed tests of any

2 Classical Statistical Inference Extended: Split-Tailed Tests

population effect: the difference between two population means or between a single population mean and its hypothesized value, the difference between a population correlation coefficient and zero, or between two (independent or dependent) correlation coefficients, and so on.

Alternative, but Equivalent Decision Rules

The split-tailed versions of DRs 2 through 3 presented in classical statistical inference: practice versus presentation should be clear. In particular, the rejection-region-based DR 3 can be illustrated as follows for the case of a large-sample t Test of the difference between two sample means (i.e., an attempt to establish the direction of the difference between the two corresponding population means), where the researcher requires considerably stronger evidence to be convinced that $\mu_1 < \mu_2$ than to be convinced that $\mu_1 > \mu_2$ (Figure 1).

Let us modify the example used in the companion entry [3] in which we tested whether interest in Web surfing (Y) increases or decreases with age (X), that is, whether the correlation between Web surfing and age (r_{XY}) is positive (>0) or negative (<0). *Before* collecting any data, we decide that, on the basis of prior data and logic, it will be easier to convince us that the overall trend is negative (i.e., that the overall tendency is for Web-surfing interest to decline with age) than the reverse, so we set $\alpha_<$ to 0.04 and $\alpha_>$ to 0.01. We draw a random sample of 43 US residents five years of age or older, determine each sampled individual's age and interest in Web surfing (measured on a 10-point quasi-interval scale on which high scores represent high interest in Web surfing), and compute the sample correlation between our two measures, which turns out to be -0.31 .

Using Victor Bissonette's statistical applet for computation of the P values associated with sample correlation coefficients (<http://fsweb.berry.edu/academic/education/vbissonnette/applets/sigr.html>), we find that $p_<$ (the probability that an r_{XY} computed on a random sample of size 43 and, thus, $df = 43 - 2 = 41$, drawn from a population in which $\rho_{XY} = 0$, would be less than or equal to -0.31) equals 0.02153. Since this is smaller than 0.04, we reject H_0 in favor of $H_2: \rho_{XY} < 0$.

Had our sample yielded an r_{XY} of $+0.31$, we would have found that $p_> = 0.022$, which is greater than 0.01 (the value to which we set $\alpha_>$), so we would *not* reject H_0 (though we would certainly not accept it), and would conclude that we had insufficient evidence to determine whether ρ_{XY} is positive or negative. Had we simply conducted a symmetric (aka two-tailed) test of H_0 —that is, had we set $\alpha_> = \alpha_< = 0.025$, we would have found in this case that $p_> < \alpha_>$ and we would, therefore, have concluded that $\rho_{XY} > 0$. This is, of course, the price one pays in employing unequal values of $\alpha_>$ and $\alpha_<$: If our pretest assessment of the relative plausibility of H_2 versus H_1 is correct, we will have higher power to detect differences in the predicted direction but lower power to detect true population differences opposite in sign to our expectation than had we set identical decision criteria for positive and negative sample correlations.

A Recommended Supplementary Calculation: The Confidence Interval

The **confidence interval** for the case where we have set $\alpha_<$ to 0.04 and $\alpha_>$ to 0.01, but obtained an r_{XY} of $+0.31$, is instructive on two grounds: First, the

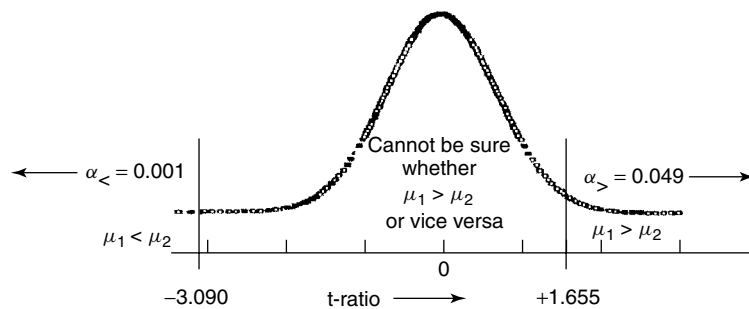


Figure 1 An example: Establishing the sign of a (population) correlation coefficient

bias toward negative values of ρ_{XY} built into the difference between $\alpha_{<}$ and $\alpha_{>}$ leads to a confidence interval that is shifted toward the negative end of the -1 to $+1$ continuum, namely, $-0.047 < \rho_{XY} < 0.535$. Second, it also yields a confidence interval that is wider (range of 0.582) than the symmetric-alpha case (range of 0.547). Indeed, a common justification for preferring $\alpha_{>} = \alpha_{<}$ is that splitting total alpha equally leads to a narrower confidence interval than any other distribution of alpha between $\alpha_{>}$ and $\alpha_{<}$. However, Harris and Vigil ([3], briefly described in Chapter 1 of [2]), have found that the PICI (Prior Information Confidence Interval) around the mean yields asymmetric-case confidence intervals that are, over a wide range of true values of μ , narrower than corresponding symmetric-case intervals. (The PICI is defined as the set of possible values of the population mean that would not be rejected by a split-tailed test, where the $\alpha_{>}$ to $\alpha_{<}$ ratio employed decreases as an exponential function of the particular value of μ being tested, asymptotically approaching unity and zero for infinitely small and infinitely large values of μ , and equals 0.5 for the value of μ that represents the investigator's *a priori* estimate of μ .) Since the CI for a correlation is symmetric around the sample correlation when expressed in Fisher- z -transformed units, it seems likely that applying PICIs around correlation coefficients would also overcome this disadvantage of splitting total alpha unequally.

Power of Split-tailed versus One- and Two (Equal)-Tailed Tests

A common justification for using one-tailed tests is that they have greater power (because they employ lower critical values) than do two-tailed tests. However, that is true only if the researcher's hypothesis about the direction of the population effect is indeed correct; if the population effect differs in sign from that hypothesized, the power of his one-tailed test of his hypothesis cannot exceed $\alpha/2$ – and all of even that low power is actually Type III error, that is, represents cases where the researcher comes to the incorrect conclusion as to the direction of the population effect. Further, the power advantage of a one-tailed test (i.e., a split-tailed test with infinite bias in favor of one's research hypothesis) over a split-tailed test with, say, a 50-to-1 bias is miniscule when the research hypothesis is correct, and the latter

test has much higher power than the one-tailed test when the researcher is mistaken about the sign of the population effect.

To demonstrate the above points, consider the case where IQ scores are, for the population under consideration, distributed normally with a population mean of 105 and a population standard deviation of 15. Assume further that we are interested primarily in establishing whether this population's mean IQ is above or below 100, and that we propose to test this by drawing a random sample of size 36 from this population. If we conduct a two-tailed test of this effect, our power (the probability of rejecting H_0) will be 0.516, and the width of the 95% CI around our sample mean IQ will be 9.8 IQ points. If, on the other hand, we conduct a one-tailed test of H_r that $\mu > 100$, our power will be 0.639 – a substantial increase. If, however, our assessment of prior evidence, expert opinion, and so on, has led us astray and we, instead, hypothesize that $\mu < 100$, our power will be a miniscule 0.00013 – and every bit of that power will come from cases where we have concluded incorrectly that $\mu < 100$. Further, since no sample value on the nonpredicted side of 100 can be rejected by a one-tailed test, our confidence interval will be infinitely wide. Had we instead conducted a split-tailed test with 'only' a 49-to-1 bias in favor of our H_r , our power would have been 0.635 (0.004 lower than the one-tailed test's power) if our prediction was correct and 0.1380 if our prediction was incorrect (over a thousand times higher than the one-tailed test's power in that situation); only 0.0001 of that power would have been attributable to Type III error; and our confidence intervals would each have the decidedly noninfinite width of 11.9 IQ points.

In my opinion, the very small gain in power from conducting a one-tailed test of a correct H_r , rather than a split-tailed test, is hardly worth the risk of near-zero power (all of it actually Type III error), the certainty of an infinitely wide confidence interval, and the knowledge that one has violated the principle that scientific hypotheses must be disconfirmable that come along with the use of a one-tailed test. Small surprise, then, that the one-tailed test was *not* included in the companion entry's description of Classical Statistical Inference in scientifically sound practice (see **Classical Statistical Inference: Practice versus Presentation**).

4 Classical Statistical Inference Extended: Split-Tailed Tests

Practical Computation of Split-tailed Significance Tests and Confidence Intervals

Tables of critical values and some computer programs are set up for one-tailed and two-tailed tests at conventional (usually 0.05 and 0.01 and, sometimes, 0.001) alpha levels. This makes sense; it would, after all, be impossible to provide a column for every possible numerical value that might appear in the numerator or denominator of the $\alpha_{>} / \alpha_{<}$ ratio in a split-tailed test. If the computer program you use to compute your test statistic does not provide the P value associated with that test statistic (which would, if provided, permit applying DR 1), or if you find yourself relying on a table or program with only 0.05 and 0.01 levels, you can use the 0.05 one-tailed critical value (aka the 0.10-level two-tailed critical value) in the predicted direction and the 0.01-level one-tailed test in the nonpredicted direction for a 5-to-1 bias in favor of your research hypothesis, and a total alpha of 0.06 – well within the range of uncertainty about the true alpha of a nominal 0.05-level test, given all the assumptions that are never perfectly satisfied. Or, you could use the 0.05-level and 0.001-level critical values for a 50-to-1 bias and a total alpha of 0.051.

Almost all statistical packages and computer sub-routines report only symmetric confidence intervals. However, one can construct the CI corresponding to a split-tailed test by combining the lower bound of a $(1-2\alpha_{>})$ -level symmetric CI with the upper bound of a $(1-2\alpha_{<})$ -level symmetric CI. This also serves as a rubric for hand computations of CIs to accompany split-tailed tests. For instance, for the example used to demonstrate relative power ($H_0 : \mu_Y = 100$, Y distributed normally with a *known* population standard deviation of 15 IQ points, $\alpha_{>} = 0.049$ and $\alpha_{<} = 0.001$, sample size = 36 observations), if our sample mean equals 103, the corresponding CI will extend from $103 - 1.655(2.5) = 98.8625$ (the lower bound of a 0.902-level symmetric CI around 103) to $103 + 3.090(2.5) = 110.725$ (the upper bound of the 0.998-level symmetric CI). For comparison, the CI corresponding to a 0.05-level one-tailed test with $H_r: \mu_Y > 100$ (i.e., an infinitely biased split-tailed test with $\alpha_{>} = 0.05$ and $\alpha_{<} = 0$) would extend from $103 - 1.645(2.5) = 98.89$ to $103 + \infty(2.5) = +\infty$ – or, perhaps, from 100 to $+\infty$, since the primary presumption of the one-tailed

test is that we do not accept the possibility that μ_Y could be < 100 .

Choosing an $\alpha_{>}/\alpha_{<}$ Ratio

The choice of how strongly to bias your test of the sign of the population effect being estimated in favor of the direction you believe to be implied by logic, theoretical analysis, or evidence from previous empirical studies involving this same parameter is ultimately a subjective one, as is the choice as to what overall alpha ($\alpha_{>} + \alpha_{<}$) to employ. One suggestion is to consider how strong the evidence for an effect opposite in direction to your prediction must be before you would feel compelled to admit that, under the conditions of the study under consideration, the population effect is indeed opposite to what you had expected it to be. Would a sample result that would have less than a one percent chance of occurring do it? Would a $p_{<}$ of 0.001 or less be enough? Whatever the breaking point for your prediction is (i.e., however low the probability of a result as far from the value specified in H_0 in the direction opposite to prediction as your obtained results has to be to get you to conclude that you got it wrong), make that the portion of alpha you assign to the nonpredicted tail.

Alternatively and as suggested earlier, you could choose a ratio that is easy to implement using standard tables, such as 0.05/0.01 or 0.05/0.001, though that requires accepting a slightly higher overall alpha (0.06 or 0.051, respectively, in these two cases) than the traditional 0.05.

Finally, as pointed out by section editor Randal Macdonald (personal communication), you could leave to the reader the choice of the $\alpha_{>}/\alpha_{<}$ ratio by reporting, for example, that the effect being tested would be statistically significant at the 0.05 level for all $\alpha_{>}/\alpha_{<}$ ratios greater than 4.3. Such a statement can be made for some finite ratio if and only if $p_{>} < \alpha$ – more generally, if and only if the obtained test statistic exceeds the one-tailed critical value for the obtained direction. For instance, if a t Test of $\mu_1 - \mu_2$ yielded a positive sample difference (mean for first group greater than mean for second group) with an associated $p_{>}$ of 0.049, this would be considered statistically significant evidence that $\mu_1 > \mu_2$ by any split-tailed test with an overall alpha of 0.05 and an $\alpha_{>}/\alpha_{<}$ ratio of $0.049/0.001 = 49$ or greater. If the obtained $p_{>}$ were 0.003, then the difference would be declared statistically significant evidence

that $\mu_1 > \mu_2$ by any split-tailed test with an overall alpha of 0.05 and an $\alpha_{>}/\alpha_{<}$ ratio of $0.003/0.997 = 0.00301$ or more – that is, even by readers whose bias in favor of the hypothesis that $\mu_1 < \mu_2$ led them to employ an $\alpha_{<}/\alpha_{>}$ ratio of $0.997/0.003 = 332.3$ or less. But if the obtained $p_{>}$ were 0.052, no split-tailed test with an overall alpha of 0.05 would yield statistical significance for this effect, no matter how high the preferred $\alpha_{>}/\alpha_{<}$ ratio. The one-tailed critical value, thus, can play a role as the basis for a preliminary test of whether the difference might be statistically significant by any split-tailed test – provided (in my opinion) that the researcher does not ‘buy into’ the associated logic of a one-tailed test by employing an infinitely large ratio of predicted to nonpredicted rejection-region area.

Wrapping Up

Finally, this entry should *not* be taken as an endorsement of split-tailed tests in preference to two-tailed tests. Indeed, my personal preference and habit is to rely almost exclusively on two-tailed tests and, thereby, let the data from the study in hand completely determine the decision as to which of that

study’s results to consider statistically significant. On the other hand, if you find yourself tempted to carry out a one-tailed test because of strong prior evidence as to the sign of a population parameter, a split-tailed test is, in my opinion, a far sounder approach to giving into that temptation than would be a one-tailed test.

References

- [1] Braver, S.L. (1975). On splitting the tails unequally: a new perspective on one- versus two-tailed tests, *Educational & Psychological Measurement* **35**, 283–301.
- [2] Harris, R.J. (2001). *A Primer of Multivariate Statistics*, 3rd Edition, Lawrence Erlbaum Associates, Mahwah.
- [3] Harris, R.J. & Vigil, B.V. (1998, October). Peekies: prior information confidence intervals, *Presented at meetings of Society for Multivariate Experimental Psychology*, Woodside Lake.
- [4] Kaiser, H.F. (1960). Directional statistical decisions, *Psychological Review* **67**, 160–167.

RICHARD J. HARRIS

Classical Statistical Inference: Practice versus Presentation

RICHARD J. HARRIS

Volume 1, pp. 268–278

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Classical Statistical Inference: Practice versus Presentation

Classical statistical inference as practiced by most researchers is a very useful tool in determining whether sufficient evidence has been marshaled to establish the direction of a population effect. Classical statistical inference (CSI) as described in almost all textbooks forces the researcher who takes that description seriously to choose among affirming a truism, accepting a falsehood on scant evidence, or violating one of the most fundamental tenets of scientific method by declaring one's research hypothesis impervious to disconfirmation. Let us start with the positive side of CSI.

As has often been pointed out (e.g., [7]) CSI has evolved over many decades as a blending of Fisherian and Neyman-Pearson approaches. I make no attempt in this entry to identify the traces of those traditions in current practice, but focus instead on the consensus that has emerged. (But see [7, 14], and **Deductive Reasoning and Statistical Inference; Neyman-Pearson Inference.**)

CSI in (Sound) Practice

Purpose

The test of statistical significance that is the core tool of CSI is a test of the sign or direction of some population effect: the sign of a population correlation coefficient; the slope (positive versus negative) of a population regression coefficient; which of two population means is larger (and thus whether the sign of the difference between the two means is positive or negative); whether attitudes towards a politician are, on average and in the population, positive or negative (and thus whether the population mean attitude score minus the midpoint of the scale is positive or negative); whether the students in a given state have a mean IQ above or below 100 (and thus whether mean IQ minus 100 is positive or negative); and so on (see **Catalogue of Parametric Tests**).

Underlying Logic

We attempt to establish the sign of the population effect via the also-classic method of *reductio ad absurdum*. We set up a straw-man null hypothesis that the true population effect size is precisely zero (or equivalently, that the population parameter representing the effect has a value that precisely straddles the dividing line between positive and negative or between increasing and decreasing or between $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$) and then proceed to marshal sufficient evidence to disprove this least interesting possibility.

If we cannot disprove the hypothesis that the population effect size is exactly zero, we will also be unable to disprove either the hypothesis that it equals $+10^{-24}$ (pounds or meters or units on a 10-point scale) or the hypothesis that it equals -10^{-31} . In other words, both positive and negative (increasing and decreasing, etc.) values of the population effect size could have generated our data, and we thus cannot come to any conclusion about the sign or direction of the population effect. If, on the other hand, we are able to reject the straw-man null hypothesis – that is, our observed sample effect size is too much greater or less than zero to have been generated plausibly via sampling from a population in which the null hypothesis is exactly true – then we will also be able to reject all hypotheses that the true population effect size is even more distant from zero than our observed sample effect size. In other words, all plausible values of the effect size will have the same sign as our sample effect size, and we will be safe in concluding that the population and sample effect sizes have the same sign.

Put another way, failure to reject the null hypothesis implies that the **confidence interval** around our sample effect size includes both positive and negative values, while rejecting H_0 implies that the confidence interval includes only negative or only positive values of the effect size.

Ordinarily, we should decide before we begin gathering the data on which to base our test of statistical significance what degree of implausibility will be sufficient to reject the null hypothesis or, equivalently, what confidence level we will set in establishing a confidence interval around the observed sample effect size. The decision as to how stringent a criterion to set for any particular significance test may hinge on what other tests we plan to perform on other

2 Classical Statistical Inference: Practice versus Presentation

aspects of the data collected in the present study (*see Multiple Comparison Procedures.*)

Decision Rule(s), Compactly Presented

The last few paragraphs can be summarized succinctly in the following more formal presentation of the case in which our interest is in whether the correlation between two variables, X and Y , is positive or negative.

Scientifically Sound Classical Statistical Inference (Correlation Coefficient Example).

We test $H_0: \rho_{XY} = 0$ against $H_1: \rho_{XY} > 0$ and $H_2: \rho_{XY} < 0$

by comparing $p_{>} = P(r_{XY} > r_{XY}^* \text{ if } H_0 \text{ were true})$ to $\alpha/2$ and

$p_{<} = P(r_{XY} < r_{XY}^* \text{ if } H_0 \text{ were true}) = 1 - p_{>}$ to $\alpha/2$, where

r_{XY}^* is the observed value of the correlation coefficient calculated for your sample;

r_{XY} is a random variable representing the values of the sample correlation coefficient obtained from an infinitely large number of independent random samples from a population in which the true population correlation coefficient is precisely zero;

$p_{>}$ is, as the formula indicates, the percentage of the correlation coefficients computed from independent random samples drawn from a population in which $\rho_{XY} = 0$ that are as large as or larger than the one you got for your single sample from that population;

$p_{<}$ is defined analogously as $P(r_{XY} < r_{XY}^* \text{ if } H_0 \text{ were true})$; and

α is the criterion you have set (*before* examining your data) as to how low $p_{>}$ or $p_{<}$ has to be to convince you that the population correlation is, indeed, positive (high values of Y tending to go along with high values of X and low, with low).

These two comparisons ($p_{<}$ to $\alpha_{<}$ and $p_{>}$ to $\alpha_{>}$) are then used to decide between H_1 and H_2 , as follows:

Decision rule: If $p_{>} < \alpha/2$, conclude that (or, at least for the duration of this study, act as if)

H_1 is true. (I.e., accept the hypothesis that $\rho_{XY} > 0$.)

If $p_{<} < \alpha/2$, conclude that (or act as if)

H_2 is true. (I.e., accept H_2 that $\rho_{XY} < 0$.)

If neither of the above is true (i.e., if $p_{>}$ and $p_{<}$ are both $\geq \alpha/2$ which is

equivalent to the condition that $p_{>}$ be greater than $\alpha/2$ but less than $1 - \alpha/2$),

conclude that

we don't have enough evidence to decide between H_1 and H_2 .

(i.e., fail to reject H_0 .)

Scientifically Sound Classical Statistical Inference About Other Population Parameters.

To test whether any other population parameter θ (e.g., the population mean, the difference between two population means, the difference between two population correlations) is greater than or less than some especially salient value θ_0 that represents the dividing line between a positive and negative effect (e.g., 100 for mean IQ, zero for the difference between two means or the difference between two correlations), simply substitute θ (the parameter of interest) for ρ_{XY} , $\hat{\theta}^*$ (your sample estimate of the parameter of interest, that is, the observed value in your sample of the statistic that corresponds to the population parameter) for r_{XY}^* , and θ_0 for zero (0) in the above.

The above description is silent on the issue of how one goes about computing $p_{>}$ and/or $p_{<}$. This can be as simple as conducting a single-sample z test or as complicated as the lengthy algebraic formulae for the test of the significance of the difference between two correlation coefficients computed on the same sample [17]. See your local friendly statistics textbook or journal or the entries on various significance tests in this encyclopedia for details (*see Catalogue of Parametric Tests*).

An Example: Establishing the Sign of a (Population) Correlation Coefficient

For example, let's say that we wish to know whether interest in Web surfing (Y) increases or decreases with age (X), that is, whether the correlation between Web surfing and age (r_{XY}) is positive (>0) or negative (<0). *Before* collecting any data we decide to set α to 0.05. We draw a random sample of 43 US residents 5 years of age or older, determine each sampled individual's age and interest in Web surfing (measured on a 10-point quasi-interval scale on which high scores represent high interest in Web surfing), and compute the sample correlation between our two measures, which turns out to be -0.31 . (Of course, this relationship is highly likely to be curvilinear, with few 5-year-olds expressing much interest in Web

surfing; we're testing only the overall linear trend of the relationship between age and surfing interest. Even with that qualification, before proceeding with a formal significance test we should examine the scatterplot of Y versus X for the presence of **outliers** that might be having a drastic impact on the slope of the best-fitting straight line and for departures from normality in the distributions of X and Y so extreme that transformation and/or nonparametric alternatives should be considered.)

Using Victor Bissonette's statistical applet for computation of the P values associated with sample correlation coefficients (<http://fsweb.berry.edu/academic/education/vbissonnette/applets/sigr.html>), we find that $p_<$ (the probability that an r_{XY} computed on a random sample of size 43 and thus $df = 43 - 2 = 41$, drawn from a population in which $\rho_{XY} = 0$, would be less than or equal to -0.31) equals 0.02153. Since this is smaller than $0.05/2 = 0.025$, we reject H_0 in favor of H_2 : $\rho_{XY} < 0$.

Had our sample yielded an r_{XY} of $+0.28$ we would have found that $p_< = 0.965$ and $p_> = 0.035$, so we would *not* reject H_0 (though we would certainly not accept it) and would conclude that we had insufficient evidence to determine whether ρ_{XY} is positive or negative. (It passeth all plausibility that it could be precisely 0.000... to even a few hundred decimal places.)

A Recommended Supplementary Calculation: The Confidence Interval

It is almost always a good idea to supplement any test of statistical significance with the confidence interval around the observed sample difference or correlation (Harlow, Significance testing introduction and overview, in [9]). The details of how and why to do this are covered in the confidence interval entry in this encyclopedia (see **Confidence Intervals**). To reinforce its importance, however, we'll display the confidence intervals for the two subcases mentioned above.

First, for the sample r_{XY} of $+0.31$ with $\alpha = 0.05$, the 95% confidence interval (CI) around (well, attempting to capture) ρ_{XY} is $0.011 < \rho_{XY} < 0.558$. (It can be obtained via a plug-in program on Richard Lowry's VassarStats website, <http://faculty.vassar.edu/lowry/rho.html>.) The lower bound of this interval provides a useful caution against

confusing statistical significance with magnitude or importance, in that it tells us that, while we can be quite confident that the population correlation is positive, it could plausibly be as low as 0.011. (For example, our data are insufficient to reject the null hypothesis that $\rho_{XY} = 0.02$, that is that the two variables share only 0.04% of their variance).

The case where we obtained an r_{XY} of $+0.28$ yields a 95% CI of $-0.022 < \rho_{XY} < 0.535$. No value contained within the 95% CI can be rejected as a plausible value by a 0.05-level significance test, so that it is indeed true that we cannot rule out zero as a possible value of ρ_{XY} , which is consistent with our significance test of r_{XY} . On the other hand, we also can't rule out values of -0.020 , $+0.200$, or even $+0.50$ – a population correlation accounting for 25% of the variation in Y on the basis of its linear relationship to X . The CI thus makes it abundantly clear how foolish it would be to accept (rather than fail to reject) the null hypothesis of a population correlation of precisely zero on the basis of statistical nonsignificance.

Alternative, but Equivalent Decision Rules

DR (Decision Rule) 2

If either $p_>$ or $p_<$ is $< \alpha/2$ and the sample estimate of the population effect is positive ($\hat{\theta} > \theta_0$, for example, r_{XY}^* positive or sample mean IQ > 100) reject H_0 and accept H_1 that $\theta > \theta_0$ (e.g., that $\rho_{XY} > 0$).

If either $p_>$ or $p_<$ is $< \alpha/2$ and the sample estimate of the population effect is negative ($\hat{\theta} < \theta_0$, for example, $r_{XY}^* < 0$ or sample mean IQ < 100) reject H_0 and accept H_1 that $\theta < \theta_0$ (e.g., that the population correlation is negative or that the population mean IQ is below 100).

DR 3. First, select a test statistic T that, for fixed α and sample size, is monotonically related to the discrepancy between $\hat{\theta}$ and θ_0 . (Common examples would be the z - or t ratio for the difference between two independent or correlated means and the chi-square test for the difference between two independent proportions.) Compute T^* (the observed value of T for your sample). By looking it up in a table or using a computer program (e.g., any of the widely available online statistical applets, such as those on Victor Bissonnette's site, cited earlier),

4 Classical Statistical Inference: Practice versus Presentation

determine either the two-tailed P value associated with T^* or T_{crit} , the value of T^* that would yield a two-tailed P value of exactly α . (The ‘two-tailed P value’ equals twice the smaller of $p_>$ or $p_<$ – that is, it is the probability of observing a value of T as large as or larger than T^* in absolute value in repeated random samplings from a population in which H_0 is true.) Then, if $p < \alpha$ or if $|T^*|$ (the absolute value of T^* , that is, its numerical value, ignoring sign) is greater than T_{crit} , conclude that the population effect has the same sign as the sample effect – that is, accept whichever of H_1 or H_2 is consistent with the data. If instead $|T^*| < T_{\text{crit}}$, conclude that we don’t have enough evidence to decide whether $\theta > \theta_0$ or $< \theta_0$ (e.g., whether ρ_{XY} is positive or negative).

This decision rule is illustrated below in Figure 1 for the case of a large-sample t Test of the difference between two sample means (i.e., an attempt to establish the direction of the difference between the two corresponding population means) where the researcher has set α to 0.05.

DR 4. Construct the $(1 - \alpha)$ -level confidence interval (CI) for θ corresponding to your choice of α and to $\hat{\theta}^*$, the value of $\hat{\theta}$ you obtained for your sample of observations. (See the entry in this encyclopedia on confidence intervals for details of how to do this.) Then

- If the CI includes only values that are $> \theta_0$, conclude that $\theta > \theta_0$.
- If the CI includes only values that are $< \theta_0$, conclude that $\theta < \theta_0$.
- Otherwise (i.e., if the CI includes some values that are $> \theta_0$ and some that are $< \theta_0$), conclude that we

don’t have enough evidence to decide whether $\theta > \theta_0$ or $< \theta_0$.

As applied to testing a single, one-degree-of-freedom hypothesis the above six decision rules are logically and algebraically equivalent and therefore lead to identical decisions when applied to any given sample of data. However, Decision Rule DR 4 (based on examination of the confidence interval around the sample estimate) actually encompasses an infinity of significance tests, since it neatly partitions the real line into values of θ that our data disconfirm (to within the ‘reasonable doubt’ quantified by α) and those that are not inconsistent with our data. This efficiency adds to the argument that confidence intervals could readily replace the use of significance tests as represented by DRs 1–3. However, this efficiency comparison is reversed when we consider multiple-degree-of-freedom (aka ‘overall’) significance tests, since an appropriate overall test tells us whether any of the infinite number of single- df contrasts is (or would be, if tested) statistically significant.

Criticisms of and Alternatives to CSI

The late 1990’s saw a renewal of a recurring cycle of criticism of classical statistical inference, including a call by Hunter [10] for a ban on the reporting of null-hypothesis significance tests (NHSTs) in APA journals. The history of this particular cycle is recounted by Fidler [6]; a compilation of arguments for and against NHST is provided by [10].

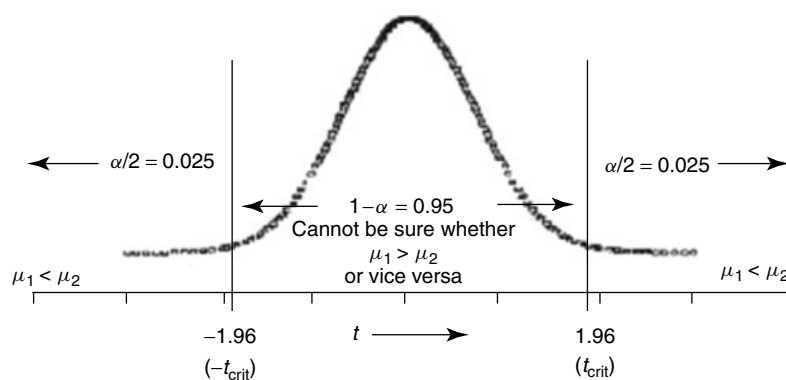


Figure 1 An example: Establishing the direction of the difference between two population means

Briefly, the principal arguments offered against NHST (aka CSI) are that

- (a) It wastes the researcher's time testing an hypothesis (the null hypothesis) that is never true for real variables.
- (b) It is misinterpreted and misused by many researchers, the most common abuse being treating a nonsignificant result (failure to reject H_0) as equivalent to accepting the null hypothesis.
- (c) It is too rigid, treating a result whose P value is just barely below α (e.g., $p = .0498$) as much more important than one whose P value is just barely above α (e.g., $p = .0501$).
- (d) It provides the probability of the data, given that H_0 is true, when what the researcher really wants to know is the probability that H_0 is true, given the observed data.
- (e) There are alternatives to NHST that have much more desirable properties, for example (i) Bayesian inference (see **Bayesian Statistics**), (ii) effect sizes (see **Effect Size Measures**), (iii) confidence intervals, and (iv) **meta-analysis**.

Almost equally briefly, the counterarguments to the above arguments are that

- (a) As pointed out earlier, we test H_0 , not because anyone thinks it might really be true, but because, if we can't rule out θ_0 (the least interesting possible value of our population parameter, θ), we also can't rule out values of θ that are both greater than and smaller than θ_0 – that is, we won't be able to establish the sign (direction) of our effect in the population, which is what significance testing (correctly interpreted) is all about.
- (b) That NHST is often misused is an argument for better education, rather than for abandoning CSI. We certainly do need to do a better job of presenting CSI to research neophytes – in particular, we need to expunge from our textbooks the traditional, antiscientific presentation of NHST as involving a choice between two, rather than three, conclusions.
- (c) Not much of a counterargument to this one. It seems compelling that confidence in the direction of an effect should be a relatively smooth, continuous function of the strength of the evidence against the null hypothesis (and thus,

more importantly, against the hypothesis that the population effect is opposite in sign to its sample estimate). In practice (as Estes [5] explicitly states for himself but also opines is true for most journal editors), a result with an associated P value of 0.051 is unlikely to lead to an automatic decision not to publish the report – unless the editor feels strongly that the most appropriate alpha level for translating the continuous confidence function into a discrete, 'take it seriously' versus 'require more data and/or replication' decision is 0.01 or 0.001. The adoption as a new 'standard' of an overall alpha ($\alpha_+ + \alpha_-$) greater than 0.05 (say, 0.055) is unlikely, however, for at least two reasons: First, such a move would simply transfer the frustration of 'just missing' statistical significance and the number of complaints about the rigidity of NHST from researchers whose P values have come out to 0.51 or 0.52 to those with P values of 0.56 or 0.57. Second, as a number of authors (e.g., Wainer & Robinson [18]) have documented, 0.05 is already a more liberal alpha level than the founders of CSI had envisioned and than is necessary to give reasonable confidence in the replicability of a finding (e.g., Greenwald, et al. [8], who found that a P value of 0.005 – note the extra zero – provides about an 80% chance that the finding will be statistically significant at the 0.05 level in a subsequent exact replication).

- (d) This one is just flat wrong. Researchers are *not* interested in (or at least shouldn't be interested in) the probability that H_0 is true, since we know *a priori* that it is almost certainly *not* true. As suggested before, most of the ills attributed to CSI are due to the misconception that we can *ever* collect enough evidence to demonstrate that H_0 is true to umpteen gazillion decimal places.
- (e)
 - (i) Bayesian statistical inference (BSI) is in many ways a better representation than CSI of the way researchers integrate data from successive studies into the belief systems they have built up from a combination of logic, previous empirical findings, and perhaps unconscious personal biases. BSI requires that the researcher make her belief

system explicit by spelling out the prior probability she attaches to every possible value of the population parameter being estimated and then, once the data have been collected, apply Bayes' Theorem to modify that distribution of prior probabilities in accordance with the data, weighted by their strength relative to the prior beliefs. However, most researchers (though the size of this majority has probably decreased in recent years) feel uncomfortable with the overt subjectivity involved in the specification of prior probabilities. CSI has the advantage of limiting subjectivity to the decision as to how to distribute total alpha between $\alpha_> + \alpha_<$. Indeed, split-tailed tests (cf. **Classical Statistical Inference Extended: Split-Tailed Tests**) can be seen as a 'back door' approach to Bayesian inference.

- (ii) Confidence intervals are best seen as complementing, rather than replacing significance tests. Even though the conclusion reached by a significance test of a particular θ_0 is identical to that reached by checking whether θ_0 is or is not included in the corresponding CI, there are nonetheless aspects of our evaluation of the data that are much more easily gleaned from one or the other of these two procedures. For instance, the CI provides a quick, easily understood assessment of whether a non-significant result is a result of a population effect size that is very close to zero (e.g., a CI around mean IQ of your university's students that extends from 99.5 to 100.3) or is instead due to a sloppy research design (high variability and/or low sample size) that has not narrowed down the possible values of θ_0 very much (e.g., a CI that states that population mean IQ is somewhere between 23.1 and 142.9). On the other hand, the P value from a significance test provides a measure of the confidence you should feel that you've got the sign of the population effect right. Specifically, the probability that you have declared $\theta - \theta_0$ to have the wrong sign is at most half of the P value for a two-tailed significance test. The P value is also directly related

to the likelihood that an exact replication would yield statistical significance in the same direction [8]. Neither of these pieces of information is easily gleaned from a CI except by rearranging the components from which the CI was constructed so as to reconstruct the significance test. Further, the P value enables the reader to determine whether to consider an effect statistically significant, regardless of the alpha level he or she prefers, while significance can be determined from a confidence interval only for the alpha level chosen by the author. In short, we need both CIs and significance tests, rather than either by itself.

- (iii) In addition to the issue of the sign or direction of a given population effect, we will almost always be interested in how large an effect is, and thus how important it is on theoretical and/or practical/clinical grounds. That is, we will want to report a measure of *effect size* for our finding. This will often be provided implicitly as the midpoint of the range of values included in the confidence interval – which, for symmetrically distributed $\hat{\theta}$, will also be our point estimate of that parameter. However, if the units in which the population parameter being tested and its sample estimate are expressed are arbitrary (e.g., number of items endorsed on an attitude inventory) a *standardized* measure of effect size, such as Cohen's d , may be more informative. (Cohen's d is, for a single- or two-sample t Test, the observed difference divided by the best available estimate of the standard deviation of $\hat{\theta}$) However, worrying about size of effect (e.g., how *much* a treatment helps) is usually secondary to establishing *direction* of effect (e.g., whether the treatment helps or harms).
- (iv) The 'wait for the *Psych Bull* article' paradigm was old hat when I entered the field lo those many decades ago. This paradigm acknowledges that any one study will have many unique features that render generalization of its findings hazardous, which is the basis for the statement earlier in this entry that a 'conclusion' based on a significance test is limited in scope to the

present study. For the duration of the report of the particular study in which a given significance test occurs we agree to treat statistically significant effects as if they indeed matched the corresponding effect in sign or direction, even though we realize that we may have happened upon the one set of unique conditions (e.g., the particular on/off schedule that yielded ulcers in the classic ‘executive monkey’ study [2]) that yields this direction of effect. We gain much more confidence in the robustness of a finding if it holds up under replication in different laboratories, with different sources of respondents, different researcher biases, different operational definitions, and so on. Review articles (for many years but no longer a near-monopoly of *Psychological Bulletin*) provide a summary of how well a given finding or set of findings holds up under such scrutiny. In recent decades the traditional ‘head count’ (tabular review) of what proportion of studies of a given effect yielded statistical significance under various conditions has been greatly improved by the tools of meta-analysis, which emphasizes the recording and analysis of an effect-size measure extracted from each reviewed study, rather than simply the dichotomous measure of statistical significance or not. Thus, for instance, one study employing a sample of size 50 may find a statistically nonsignificant correlation of 0.14, while another study finds a statistically significant correlation of 0.12, based on a sample size of 500. Tabular review would count these studies as evidence that about half of attempts to test this relationship yield statistical significance, while meta-analysis would treat them as evidence that the effect size is in the vicinity of 0.10 to 0.15, with considerable consistency (low variability in obtained effect size) from study to study. Some authors [e.g., [16]] have extolled meta-analysis as an alternative to significance testing. However, meta-analytic results, while based on larger sample sizes than any of the single studies

being integrated, are not immune to sampling variability. One would certainly wish to ask whether effect size, averaged across studies, is statistically significantly higher or lower than zero. In other words, conducting a meta-analysis does not relieve one of the responsibility of testing the strength of the evidence that the population effect has a particular sign or direction.

CSI As Presented (Potentially Disastrously) in Most Textbooks

Here’s where I attempt to document the ‘dark side’ of classical statistical inference, namely, the overwhelming tendency of textbooks to present its logic in a way that forces the researcher who takes it seriously to choose between vacuous versus scientifically unsound conclusions. Almost all such presentations consider two DRs: the *two-tailed test* (related to the two-tailed P value defined earlier, as well as to DRs 1–3) and the *one-tailed test*, in which the sign of the population effect is considered to be known *a priori*.

Two-tailed Significance Test (Traditional Textbook Presentation)

DR 2T. Compute (or, more likely, look up) the two-tailed critical value of symmetrically distributed test statistic T , T_{crit} . This is the $100(1-\alpha/2)$ th percentile of the sampling distribution of T , that is the value of T such that $100(\alpha/2)\%$ of the samples drawn from a population in which H_0 is true would yield a value of T that is as far from θ_0 as (or even farther from θ_0 than) that critical value, in either direction. Then

If $|T^*|$ (the absolute value of the observed value of T in your sample) is $>cv_{\alpha/2}$, accept H_1 that $\theta \neq \theta_0$.
If $|T^*|$ is $<T_{\text{crit}}$, that is, if $-T_{\text{crit}} < T^* < T_{\text{crit}}$, do not reject H_0 that $\theta = \theta_0$.

An even more egregious but unfortunately common variation of DR 2T replaces the ‘do not reject H_0 ’ decision with the injunction to ‘accept H_0 ’.

This decision rule can be illustrated for the same case (difference between two means) that was used to illustrate DR 3 as follows (Figure 2):

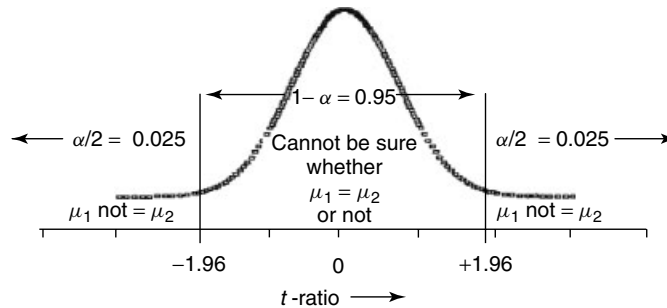


Figure 2 Two-tailed test as presented in (almost all) textbooks

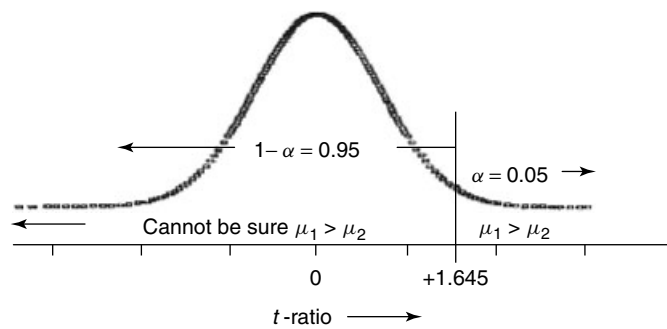


Figure 3 One-tailed Test

One-tailed Significance Test (Traditional Textbook Presentation)

DR 1T: Compute (or, more likely, look up) the one-tailed critical value of symmetrically distributed test statistic T , cv_α . This is either the negative of the α th percentile or the $100(1-\alpha)$ th percentile of the sampling distribution of T , that is the value of T such that $100(\alpha)\%$ of the samples drawn from a population in which H_0 is true would yield a value of T that is as far from θ_0 as (or even farther from θ_0 than) that critical value, in the *hypothesized* direction. Then, if the researcher has (*before* looking at the data) hypothesized that $\theta > \theta_0$ (a mirror-image decision rule applies if the *a priori* hypothesis is that $\theta < \theta_0$),

If $|T^*|$ is $> cv_\alpha$ and $\hat{\theta}^*$ (the observed sample estimate of θ) $> \theta_0$ (i.e., the sample statistic came out on the predicted side of θ_0), accept H_1 that $\theta > \theta_0$.

If $|T^*|$ is $< cv_\alpha$ or $\hat{\theta}^* < \theta_0$ (i.e., either the sample estimate was on the nonpredicted side of θ_0 or the sample result, while in the hypothesized

direction, wasn't discrepant enough from θ_0 to yield statistical significance), fail to reject H_0 that $\theta \leq \theta_0$ (i.e., conclude that the data provide insufficient evidence to prove that the researcher's hypothesis is correct).

This decision rule can be illustrated for the same case (difference between two means) that was used to illustrate DR 3 as follows (Figure 3):

One-tailed or Unidirectional? The labeling of the two kinds of tests described above as 'one-tailed' and 'two-tailed' is a bit of a misnomer, in that the crucial logical characteristics of these tests are not a function of which tail(s) of the sampling distribution constitute the rejection region, but of the nature of the alternative hypotheses that can be accepted as a result of the tests. For instance, the difference between two means can be tested (via 'two-tailed' logic) at the 0.05 level by determining whether the square of t^* for the difference falls in the right-hand tail (i.e., beyond the 95th percentile) of the F distribution with one df for numerator and the same

df for denominator as the t Test's df. Squaring t^* has essentially folded both tails of the t distribution into a single tail of the F distribution. It's what one does after comparing your test statistic to its critical value, not how many tails of the sampling distribution of your chosen test statistic are involved in that determination, that determines whether you are conducting a 'two-tailed' significance test as presented in textbooks or a 'two-tailed' significance test as employed by sound researchers or a 'one-tailed' test. Perhaps it would be better to label the above two kinds of tests as *bidirectional* versus *unidirectional* tests.

The Unpalatable Choice Presented by Classical Significance Tests as Classically Presented. Now, look again at the conclusions that can be reached under the above two decision rules. Nary a hint of direction of effect appears in either of the two conclusions (θ could $= \theta_0$ or $\theta \neq \theta_0$ in the general case, μ_1 could $= \mu_2$ or $\mu_1 \neq \mu_2$ in the example) that could result from a two-tailed test.

Further, as hinted earlier, there are no true null hypotheses except by construction. No two populations have precisely identical means on any real variable; no treatment to which we can expose the members of any real population leaves them utterly unaffected; and so on. Thus, even a statistically significant two-tailed test provides no new information. Yes, we can be 95% confident that the true value of the population parameter doesn't precisely equal θ_0 to 45 or 50 decimal places – but, then, we were 100% confident of that before we ever looked at a single datum!

For the researcher who wishes to be able to come to a conclusion (at least for purposes of the discussion of this study's results) about the *sign* (direction) of the difference between θ and θ_0 , textbook-presented significance testing leaves only the choice of conducting a one-tailed test. Doing so, however, requires not only that she make an *a priori* prediction as to the direction of the effect being tested (i.e., as to the sign of $\theta - \theta_0$), but that she declare that hypothesis to be impervious to any empirical evidence to the contrary. (If we conduct a one-tailed test of the hypothesis that $\mu_1 > \mu_2$, we can never come to the conclusion that $\mu_1 < \mu_2$ no matter how much larger $\bar{Y}_1$ is than $\bar{Y}_2$ and no matter how close to negative infinity our t ratio for the difference gets.)

In my opinion, this may be a satisfying way to run a market test of your product ('We tested every

competitor's product against ours, and not a one performed statistically significantly better than our product' – because we did one-tailed tests, and we sure weren't going to predict better performance for the competitor's product), but in my opinion it's a terrible way to run a science.

In short, and to reiterate the second sentence of this entry, classical statistical inference as described in almost all textbooks forces the researcher who takes that description seriously to choose among affirming a truism, accepting a falsehood on scant evidence, or violating one of the most fundamental tenets of scientific method by declaring one's research hypothesis impervious to disconfirmation.

Fortunately, most researchers *don't* take the textbook description seriously. Rather, they conduct two-tailed tests, but with the three possible outcomes spelled out in DR 1 through DR 4 above. Or they pretend to conduct a one-tailed test but abandon that logic if the evidence is overwhelmingly in favor of an effect opposite in direction to their research hypothesis, thus effectively conducting a split-tailed test such as those described in the entry, *Classical Statistical Inference Extended: Split-tailed Tests* [9], but with a somewhat unconventional alpha level. (E.g., if you begin by planning a one-tailed test with $\alpha = 0.05$ but revert to a two-tailed test if t comes out large enough to be significant in the direction opposite to prediction by a 0.05-level two-tailed test, you are effectively conducting a split-tailed test with an alpha of 0.05 in the predicted direction and 0.025 in the nonpredicted direction, for a total alpha of 0.075. See [1] for an example of a research report in which exactly that procedure was followed.)

However, one does still find authors who explicitly state that you must conduct a one-tailed test if you have any hint about the direction of your effect (e.g., [13], p. 136); or an academic department that insists that intro sections of dissertations should state all hypotheses in null form, rather than indicating the direction in which you predict your treatment conditions will differ (see [15] and <http://www.blackwell-synergy.com/links/doi/10.1111/j.1365-2648.2004.03074.x/abs/;jsessionid=11twdxnTU-ze> for examples of this practice and <http://etleads.csuhayward.edu/6900.html> and http://www.edb.utexas.edu/coe/depts/sped/syllabi/Spring%2003'/Parker_sed387_2nd.htm for examples of dissertation guides that enshrine it); or an

article in which the statistical significance of an effect is reported, with no mention of the direction of that effect (see http://www.gerardkeegan.co.uk/glossary/gloss_repwrit.htm and <http://web.hku.hk/~rytyeung/nurs2509b.ppt> for examples in which this practice is held up as a model for students to follow); or a researcher who treats a huge difference opposite to prediction as a nonsignificant effect – just as textbook-presented logic dictates. (Lurking somewhere in, but as yet unrecovered from my 30+ years of notes is a reference to a specific study that committed that last-mentioned sin.)

There are even researchers who continue to champion one-tailed tests. As pointed out earlier, many of these (fortunately) do not really follow the logic of one-tailed tests. For instance, after expressing concern about and disagreement with this entry's condemnation of one-tailed tests, section editor Ranald Macdonald (email note to me) mentioned that, of course, should a study for which a one-tailed test had been planned yield a large difference opposite to prediction he would consider the assumptions of the test violated and acknowledge the reversal of the predicted effect – that is, the decision rule he applies is equivalent to a split-tailed test with a somewhat vague ratio of predicted to nonpredicted alpha. Others, though (e.g., Cohen, in the entry on **Directed Alternatives in Testing** and many of the references on ordinal alternatives cited in that entry) explicitly endorse the logic of one-tailed tests.

Two especially interesting examples are provided by Burke [4] and Lohnes & Cooley [13]. Burke, in an early salvo of the 1950s debate with Jones on one-tailed versus two-tailed tests (which Leventhal [12] reports was begun because of Burke's concern that some researchers were coming to directional conclusions on the basis of two-tailed tests) concedes that a two-tailed rejection region could be used to support the directional hypothesis that the experimental-condition μ is greater than the control-condition μ – but then goes on to say that if a researcher did so 'his position would be unenviable. For following the rules of (such a) test, he would have to reject the (null hypothesis) in favor of the alternative ($\mu_E > \mu_C$), even though an observed difference ($\bar{Y}_1 - \bar{Y}_2$) was a substantial negative value.' Such are the consequences of the assumption that any significance test can have only two outcomes, rather than three: If your alternative hypothesis is that $\mu_E > \mu_C$, that leaves only $H_0: \mu_E \leq \mu_C$ as the other possible conclusion.

Lohnes and Cooley [13] follow their strong endorsement of one-tailed tests by an even stronger endorsement of traditional, symmetric confidence intervals: 'The great value of (a CI) is that it dramatizes that *all* values of μ within these limits are tenable in the light of the available evidence.' However, those nonrejected values include every value that would be rejected by a one-tailed test but not by a two-tailed test. More generally, it is the set of values that would not be rejected by a two-tailed test that match up perfectly with the set of values that lie within the symmetric confidence interval. Lohnes and Cooley thus manage to strongly denigrate two-tailed tests and to strongly endorse the logically equivalent symmetric confidence-interval procedure within a three-page interval of their text.

Few researchers would disagree that it is eminently reasonable to temper the conclusions one reaches on the basis of a single study with the evidence available from earlier studies and/or from logical analysis. However, as I explain in the companion entry, one can use **split-tailed tests** (Braver [3], Kaiser [11]) to take prior evidence into account and thereby increase the power of your significance tests without rending your directional hypothesis disconfirmable – and while preserving, as the one-tailed test does not, some possibility of reaching the correct conclusion about the sign of the population parameter when you have picked the wrong directional hypothesis.

References

- [1] Biller, H.R. (1968). A multiaspect investigation of masculine development in kindergarten age boys, *Genetic Psychology Monographs* **78**, 89–138.
- [2] Brady, J.V., Porter, R.W., Conrad, D.G. & Mason, J.W. (1958). Avoidance behavior and the development of gastroduodenal ulcers, *Journal of the Experimental Analysis of Behavior* **1**, 69–73.
- [3] Braver, S.L. (1975). On splitting the tails unequally: a new perspective on one- versus two-tailed tests, *Educational & Psychological Measurement* **35**, 283–301.
- [4] Burke, C.J. (1954). A rejoinder on one-tailed tests, *Psychological Bulletin* **51**, 585–586.
- [5] Estes, W.K. (1997). Significance testing in psychological research: some persisting issues, *Psychological Science* **8**, 18–20.
- [6] Fidler, F. (2002). The fifth edition of the APA publication manual: why its statistics recommendations are so controversial, *Educational & Psychological Measurement* **62**, 749–770.

-
- [7] Gigerenzer, G. & Murray, D.J. (1987). *Cognition as Intuitive Statistics*, Lawrence Erlbaum Associates, Hillsdale.
 - [8] Greenwald, A.G., Gonzalez, R. & Harris, R.J. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology* **33**, 175–183.
 - [9] Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (1997). *What if there were no Significance Tests?* Lawrence Erlbaum Associates, Mahwah, pp. 446.
 - [10] Hunter, J.E. (1997). Needed: a ban on the significance test, *Psychological Science* **8**, 3–7.
 - [11] Kaiser, H.F. (1960). Directional statistical decisions, *Psychological Review* **67**, 160–167.
 - [12] Leventhal, L. (1999). Updating the debate on one-versus two-tailed tests with the directional two-tailed test, *Psychological Reports* **84**, 707–718.
 - [13] Lohnes, P.R. & Cooley, W.W. (1968). *Introduction to Statistical Procedures*, Wiley, New York.
 - [14] Macdonald, R.R. (1997). On statistical testing in psychology, *British Journal of Psychology* **88**, 333–347.
 - [15] Ryan, M.G. (1994). *The effects of computer-assisted instruction on at-risk technical college students*, Doctoral dissertation, University of South Carolina, ProQuest order #ABA95-17308.
 - [16] Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers, *Psychological Methods* **1**, 115–129.
 - [17] Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix, *Psychological Bulletin* **87**, 245–251.
 - [18] Wainer, H. & Robinson, D.H. (2003). Shaping up the practice of null hypothesis significance testing, *Educational Researcher* **32**, 22–30.

RICHARD J. HARRIS

Classical Test Models

JOSÉ MUÑIZ

Volume 1, pp. 278–282

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Classical Test Models

Classical test theory (CTT) covers a whole set of psychometric models and procedures whose main objective is to solve the basic problems arising in the measurement of variables in the social sciences. Thus, for example, when psychologists measure a variable such as intelligence or extraversion, to name but two examples, and obtain an empirical score, their interest lies not so much in the score itself as in the inferences and interpretations that can be made from it, with a view to gaining more knowledge about some aspect of a person's behavior. Of course, for these interpretations and inferences to have the appropriate foundation, we need accurate knowledge of the different psychometric properties of the instrument used. CTT permits the detailed description of the metric characteristics of the measurement instruments commonly used by social scientists and other social science professionals. This set of procedures is referred to as 'classical' because it has been firmly established for some 100 years, and is widely used by those working in the social sciences, especially in the fields of psychology and education, in which the majority of these procedures originated and were developed. In an alternative to this classical approach, there began to emerge in the 1960s other psychometric models in response to some problems that had not been dealt with adequately within the classical framework, and which are referred to under the general name of **Item Response Theory** (IRT). The two approaches are complementary, each one being employed according to the type of problem faced.

Brief Historical Note

The first developments in what we know today under the generic name of classical test theory are to be found in the pioneering work of **Spearman** [29–31]. Later, Guilford [14] published a text in which he tried to collate and organize all the work done so far. But it would be Gulliksen [15] who produced the key work in the field, in which he set down in a systematic way the basic contributions of CTT. It can indeed be claimed that by the time his book was published the essential basis of classical test theory was complete. The year 1954 saw the publication

of the first *Technical recommendations for psychological tests and diagnostic techniques*; since then, these guidelines, developed jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, and which constitute an update of the psychometric consensus, have undergone several revisions, the most recent in 1999 [2]. Another essential text, and one which bridges the classical approach and the new psychometric models of IRT, is that of Lord and Novick [20]. This important work served both to reanalyze the classical approach and to provide a springboard for new item response theory models, which, as already pointed out, provided the solution to some of the problems for which the classical framework was inadequate. Other texts in which the reader can find clear and well-documented presentations of CTT are [21, 1, 35, 8, 23, 25, 26].

Below is a brief chronology showing some of the landmarks of CTT (Table 1), adapted from [25, 26]. It does not include the more recent developments within the paradigm of item response theory.

Classical Linear Model

The psychometric knowledge and techniques included within the generic term of 'classical test theory' derive from developments based on the linear model, which has its origins in the pioneering work of Spearman [29–31]. In this model, a person's empirical score from a test (X) is assumed to be made up of two additive components, the true score (T) that corresponds to the expected score on the test for the person (over multiple parallel tests), and a certain measurement error (e). Formally, the model can be expressed as:

$$X = T + e, \quad (1)$$

where X is the empirical score obtained, T is the true score and e is the measurement error.

In order to be able to derive the formulas necessary for calculating the reliability, the model requires three assumptions and a definition. It is assumed (a) that a person's true score in a test would be that obtained as an average if the test were applied an infinite number of times [$E(X) = T$], (b) that there is no correlation between person true score and measurement error ($\rho_{te} = 0$), and (c) that the

2 Classical Test Models

Table 1 Classical Psychometric Chronology

1904	Spearman publishes his two-factor theory of intelligence and the attenuation formulas E. L. Thorndike publishes the book <i>An Introduction to the Theory of Mental and Social Measurements</i> [34]
1905	Binet and Simon publish the first intelligence scale
1910	Spearman–Brown formula that relates reliability to the length of tests is published
1916	Terman publishes Stanford’s revision of the Binet–Simon scale
1918	Creation of the Army Alpha tests
1931	Thurstone publishes <i>The Reliability and Validity of Tests</i> [36]
1935	The Psychometric Society is founded Buros publishes his first review of tests (<i>Mental Measurements Yearbook</i>)
1936	Guilford publishes <i>Psychometric Methods</i> First issue of the journal <i>Psychometrika</i>
1937	Kuder and Richardson publish their formulas including KR_{20} and KR_{21}
1938	Bender, Raven, and PMA tests published
1939	Wechsler proposes his scale for measuring intelligence
1940	Appearance of the personality questionnaire <i>Minnesota Multiphasic Personality Inventory</i> (MMPI)
1946	Stevens proposes his four measurement scales: Nominal, ordinal, interval and ratio [33]
1948	Educational Testing Service (ETS) in the United States is established
1950	Gulliksen publishes <i>Theory of Mental Tests</i>
1951	Cronbach introduces coefficient Alpha [9] First edition of <i>Educational Measurement</i> is edited by Lindquist
1954	First edition of <i>Technical Recommendations for Psychological Tests and Diagnostic Techniques</i> is published
1955	Construct validity is introduced by Cronbach and Meehl [11]
1958	Torgerson publishes <i>Theory and Methods of Scaling</i> [37]
1959	Discriminant convergent validity is introduced by Campbell and Fiske [6]
1960	Rasch proposes the one-parameter logistic model better known as the Rasch model
1963	Criterion-referenced testing is introduced by Robert Glaser [13] Generalizability theory is introduced by Lee Cronbach
1966	Second edition of <i>Standards for Educational and Psychological Tests</i> is published
1968	Lord and Novick publish <i>Statistical Theories of Mental Tests Scores</i>
1971	Second edition of <i>Educational Measurement</i> is published, edited by Robert Thorndike
1974	Third edition of <i>Standards for Educational and Psychological Tests</i> is published
1980	Lord publishes <i>Applications of Item Response Theory to Practical Testing Problems</i>
1985	Fourth edition of <i>Standards for Educational and Psychological Tests</i> is published Hambleton and Swaminathan’s book, <i>Item Response Theory: Principles and Applications</i> , is published
1989	The third edition of <i>Educational Measurement</i> , edited by Robert Linn, is published
1997	Seventh edition of Anastasi’s <i>Psychological Testing</i> is published [3] <i>Handbook of Modern IRT Models</i> by van der Linden and Hambleton [38] is published
1999	Fifth edition of <i>Standards for Educational and Psychological Tests</i> is published

measurement errors from parallel forms are not correlated. Moreover, parallel tests are defined as those that measure the same construct, and in which a person has the same true score on each one, and the standard error of measurement (the standard deviation of the error scores) is also identical across parallel forms.

From this model, by means of the corresponding developments, it will be possible to arrive at operational formulas for the estimation of the errors (e) and true scores (T) of persons. These deductions constitute the essence of CTT, whose formulation is described in the classic texts already mentioned.

Reliability

Through the corresponding developments, we obtain the formula of the *reliability coefficient* ($\rho_{xx'}$) (see **Reliability: Definitions and Estimation**). Its formula expresses the amount of variance of true measurement (σ_T^2) there is in the empirical measurement (σ_x^2). This formula, which is purely conceptual, cannot be used for the empirical calculation of the reliability coefficient value. This calculation is carried out using three main designs: (a) the correlation between two parallel forms of the test, (b) the correlation between two

random halves of the test corrected using the Spearman–Brown formula, and (c) the correlation between two applications of the same test to a sample of persons. Each one of these procedures has its pros and cons, and suits some situations better than others. In all cases, the value obtained (reliability coefficient) is a numerical value between 0 and 1, indicating, as it approaches 1, that the test is measuring consistently. In the psychometric literature, there are numerous classic formulas for obtaining the empirical value of the reliability coefficient, some of the most important being those of Rulon [28], Guttman [16], Flanagan [12], the KR_{20} and KR_{21} [19], or the popular coefficient alpha [9], which express the reliability of the test according to its internal consistency.

Regardless of the formula used for calculating the reliability coefficient, what is most important is that all measurements have an associated degree of accuracy that is empirically calculable. The commonest sources of error in psychological measurement have been widely studied by specialists, who have made detailed classifications of them [32]. In general, it can be said that the three most important sources of error in psychological measurement are: (a) *the assessed persons themselves*, who come to the test in a certain mood, with certain attitudes and fears, and levels of anxiety in relation to the test, or affected by any kind of event prior to the assessment, all of which can influence the measurement errors, (b) *the measurement instrument* used, and whose specific characteristics can differentially affect those assessed, and (c) *the application, scoring, and interpretation* by the professionals involved [24].

Validity

From persons' scores, a variety of inferences can be drawn, and validating the test consists in checking empirically that the inferences made based on the test are correct (*see Validity Theory and Applications*). It could therefore be said that, strictly speaking, it is not the test that is validated, but rather the inferences made on the basis of the test. The procedure followed for validating these inferences is the one commonly used by scientists, that is, defining working hypotheses and testing them empirically. Thus, from a methodological point of view, the validation process for a test does not differ in essence from customary scientific methodology. Nevertheless, in this specific

context of test validation, there have been highly effective and fruitful forms of collecting empirical evidence for validating tests, and which, classically, are referred to as: *Content validity*, *predictive validity*, and *construct validity*. These are not three forms of validity – there is only one – but rather three common forms of obtaining data in the validation process. Content validity refers to the need for the content of the test to adequately represent the construct assessed. Predictive validity indicates the extent to which the scores in the test predict a criterion of interest; it is operationalized by means of the correlation between the test and the criterion, which is called the validity coefficient (ρ_{xy}). Construct validity [11] refers to the need to ensure that the assessed construct has entity and consistency, and is not merely spurious. There are diverse strategies for evaluating construct validity. Thus, for example, when we use the technique of **factor analysis** (or, more generally, **structural equation modeling**), we refer to factorial validity. If, on the other hand, we use the data of a multitrait-multimethod matrix (*see Multitrait–Multimethod Analyses*), we talk of convergent-discriminant validity. Currently, the concept of validity has become more comprehensive and unitary, with some authors even proposing that the consequences of test use be included in the validation process [2, 22].

Extensions of Classical Test Theory

As pointed out above, the classical linear model permits estimation of the measurement errors, but not their source; this is presumed unknown, and the errors randomly distributed. Some models within the classical framework have undertaken to break down the error, and, thus, offer not only the global reliability but also its quantity as a function of the sources of error. The most well known model is that of *generalizability theory*, proposed by Cronbach and his collaborators [10]. This model allows us to make estimations about the size of the different error sources. The reliability coefficient obtained is referred to as the generalizability coefficient, and indicates the extent to which a measurement is generalizable to the population of measurements involved in the measurement (*see Generalizability Theory: Basics*). A detailed explanation of generalizability theory can be found in [5].

The tests mentioned up to now are those most commonly used in the field of psychology for assessing constructs such as intelligence, extraversion or neuroticism. They are generally referred to as *normative tests*, since the scores of the persons assessed are expressed according to the norms developed in a normative group. A person's score is expressed according to the position he or she occupies in the group, for example, by means of centiles or standard scores. However, in educational and professional contexts, it is often of more interest to know the degree to which people have mastery in a particular field than their relative position in a group of examinees. In this case, we talk about criterion-referenced tests [13, 17] for referring to tests whose central objective is to assess a person's ability in a field, domain, or criterion (*see Criterion-Referenced Assessment*). In these circumstances, the score is expressed not according to the group, but rather as an indicator of the extent of the person's ability in the area of interest. However, the classical reliability coefficients of normative tests are not particularly appropriate for this type of test, for which we need to estimate other indicators based on the reliability of classifications [4]. Another specific technical problem with criterion-referenced tests is that of setting cut-off points for discriminating between those with mastery in the field and those without. A good description of the techniques available for setting cut-off points can be found in [7].

Limitations of the Classical Test Theory Approach

The classical approach is still today commonly used in constructing and analyzing psychological and educational tests [27]. The reasons for this widespread use are basically its relative simplicity, which makes it easy to understand for the majority of users, and the fact that it works well and can be adapted to the majority of everyday situations faced by professionals and researchers. These are precisely its strong points. Nevertheless, in certain assessment situations, the new psychometric models derived from item response theory have many advantages over the classical approach [18]. More about the limitations of CTT are found in (*see Item Response Theory (IRT) Models for Dichotomous Data*). It is fair to point out that

workers using the classical approach have developed diverse statistical strategies for the appropriate solutions of many of the problems that surface in practice, but the more elegant and technically satisfactory solutions are provided by item response models.

References

- [1] Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*, Brooks/Cole, Monterrey.
- [2] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.
- [3] Anastasi, A. & Urbina, S. (1997). *Psychological Testing*, 7th Edition, Prentice Hall, Upper Saddle River.
- [4] Berk, R.A., ed. (1984). *A Guide to Criterion-Referenced Test Construction*, The Johns Hopkins University Press, Baltimore.
- [5] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.
- [6] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
- [7] Cizek, G.J. ed. (2001). *Setting Performance Standards. Concepts, Methods, and Perspectives*, LEA, London.
- [8] Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*, Holt, Rinehart and Winston, New York.
- [9] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.
- [10] Cronbach, L.J., Glesser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurement: Theory of Generalizability for Scores and Profiles*, Wiley, New York.
- [11] Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin* **52**, 281–302.
- [12] Flanagan, J.L. (1937). A note on calculating the standard error of measurement and reliability coefficients with the test score machine, *Journal of Applied Psychology* **23**, 529.
- [13] Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions, *American Psychologist* **18**, 519–521.
- [14] Guilford, J.P. (1936, 1954). *Psychometric Methods*, McGraw-Hill, New York.
- [15] Gulliksen, H. (1950). *Theory of Mental Tests*, Wiley, New York.
- [16] Guttman, L. (1945). A basis for analyzing test-retest reliability, *Psychometrika* **10**, 255–282.
- [17] Hambleton, R.K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice* **13**(4), 21–26.

-
- [18] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory. Principles and Applications*, Kluwer, Boston.
- [19] Kuder, G.F. & Richardson, M.W. (1937). The theory of estimation of test reliability, *Psychometrika* **2**, 151–160.
- [20] Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Tests Scores*, Addison-Wesley, Reading.
- [21] Magnuson, D. (1967). *Test Theory*, Addison-Wesley, Reading.
- [22] Messick, S. (1989). Validity, in *Educational Measurement*, R. Linn, ed., American Council on Education, Washington, pp. 13–103.
- [23] Muñoz, J. ed. (1996). *Psicometría [Psychometrics]*, Universitas, Madrid.
- [24] Muñoz, J. (1998). La medición de lo psicológico [Psychological measurement], *Psicothema* **10**, 1–21.
- [25] Muñoz, J. (2003a). *Teoría Clásica De Los Tests [Classical Test Theory]*, Pirámide, Madrid.
- [26] Muñoz, J. (2003b). Classical test theory, in *Encyclopedia of Psychological Assessment*, R. Fernández-Ballesteros, ed., Sage Publications, London, pp. 192–198.
- [27] Muñoz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R. & Zaal, J. (2001). Testing practices in European countries, *European Journal of Psychological Assessment* **17**(3), 201–211.
- [28] Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves, *Harvard Educational Review* **9**, 99–103.
- [29] Spearman, C. (1904). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.
- [30] Spearman, C. (1907). Demonstration of formulae for true measurement of correlation, *American Journal of Psychology* **18**, 161–169.
- [31] Spearman, C. (1913). Correlations of sums and differences, *British Journal of Psychology* **5**, 417–126.
- [32] Stanley, J.C. (1971). Reliability, in *Educational Measurement*, R.L. Thorndike, ed., American Council on Education, Washington, pp. 356–442.
- [33] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.
- [34] Thorndike, E.L. (1904). *An Introduction to the Theory of Mental and Social Measurements*, Science Press, New York.
- [35] Thorndike, R.L. (1982). *Applied Psychometrics*, Houghton-Mifflin, Boston.
- [36] Thurstone, L.L. (1931). *The Reliability and Validity of Tests*, Edward Brothers, Ann Arbor.
- [37] Torgerson, W.S. (1958). *Theory and Methods of Scaling*, Wiley, New York.
- [38] Van der Linden, W.J., & Hambleton, R.K., eds (1997). *Handbook of Modern Item Response Theory*, Springer-Verlag, New York.

JOSÉ MUÑOZ

Classical Test Score Equating

MICHAEL J. KOLEN AND YE TONG

Volume 1, pp. 282–287

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Classical Test Score Equating

Introduction

Educational and psychological tests often are developed with *alternate forms* that contain different sets of test questions. The alternate forms are administered on different occasions. To enhance the security of the tests and to allow examinees to be tested more than once, alternate forms are administered on different occasions. *Test content specifications* detail the number of questions on a test from each of a number of content areas. *Test statistical specifications* detail the statistical properties (e.g., difficulty) of the test questions. Alternate forms of tests are built to the same content and statistical specifications, which is intended to lead to alternate forms that are very similar in content and statistical properties.

Although alternate forms are built to be similar, they typically differ somewhat in difficulty. *Test equating* methods are statistical methods used to adjust test scores for the differences in test difficulty among the forms. To appropriately apply test equating methodology, the alternate forms must be built to the same content and statistical specifications. Equating methods adjust for small differences in test difficulty among the forms. As emphasized by Kolen and Brennan [5], p. 3, '*equating adjusts for differences in difficulty, not for differences in content*'. The goal of equating is to enable scores on the alternate test forms to be used interchangeably. Test equating is used when alternate forms of a test exist, and examinees who are administered the different test forms are considered for the same decision.

The implementation of test equating requires a process for collecting data, referred to as an *equating design*. A variety of equating designs exist. Some of the more popular ones are considered here. Statistical *equating methods* are also a component of the equating process. Traditional and **item response theory (IRT)** statistical methods exist. Only the traditional methods are considered in this entry.

Test equating has been conducted since the early twentieth century. The first comprehensive

treatment of equating was presented by Flanagan [3]. Subsequent treatments by Angoff [2], Holland and Rubin [4], Petersen, Kolen, and Hoover [7], Kolen and Brennan [5, 6], and von Davier, Holland, and Thayer [8] trace many of the developments in the field. AERA/APA/NCME [1] provides standards that are to be met when equating tests in practice.

The Scaling and Equating Process

Raw scores on tests often are computed as the number of test questions that a person answers correctly. Other types of raw scores exist, such as scores that are corrected for guessing, and scores on written essays that are scored by judges. Raw scores typically are transformed to *scale scores*. Scale scores are used to facilitate score interpretation ([7], [5]). Often, properties of score scales are set with reference to a particular population. The score scale can be established using an initial alternate form of a test. Raw scores on a subsequent alternate form are equated to raw scores on this initial form. The raw-to-scale score transformation for the initial form is then applied to the equated scores on the subsequent form. Later, raw scores on new forms are equated to previously equated forms and then transformed to scale scores. The scaling and equating process leads to scores from all forms being reported on a common scale. The intent of this process is to be able to say, for example, that 'a scale score of 26 indicates the same level of proficiency whether it is earned on Form X, Form Y, or Form Z'.

Equating Designs

Equating requires that data be collected and analyzed. Various data collection designs are used to conduct equating. Some of the most common designs are discussed in this section. In discussing each of these designs, assume that Form X is a new form. Also assume that Form Y is an old form, for which a transformation of raw-to-scale scores already has been developed. The equating process is intended to relate raw scores on Form X to raw scores on Form Y and to scale scores.

2 Classical Test Score Equating

Random Groups

In the *random groups design*, alternate test forms are randomly assigned to examinees. One way to implement the random groups design is to package the test booklets so that the forms alternate. For example, if two test forms, Form X and Form Y, are to be included in an equating study, Form X and Form Y test booklets would be alternated in the packages. When the forms are distributed to examinees, the first examinee would receive a Form X test booklet, the second examinee a Form Y booklet, and so on. This assignment process leads to comparable, *randomly equivalent groups*, being administered Form X and Form Y.

Assuming that the random groups are fairly large, differences between raw score means on Form X and Form Y can be attributed to differences in difficulty of the two forms. Suppose, for example, following a random group data collection, the mean raw score for Form X is 70 and the mean raw score for Form Y is 75. These results suggest that Form X is 5 raw score points more difficult than Form Y. This conclusion is justified because the group of examinees taking Form X is randomly equivalent to the group of examinees taking Form Y.

Single Group Design

In the *single group design*, the same examinees are administered two alternate forms. The forms are separately timed. The order of the test forms is usually counterbalanced. One random half of the examinees is administered Form X followed by Form Y. The other random half is administered Form Y followed by Form X. Counterbalancing is used to control for context effects, such as practice or fatigue. Counterbalancing requires the assumption that the effect of taking Form X prior to Form Y has the same effect as taking Form Y prior to Form X. If this assumption does not hold, then *differential order effects* are said to be present, and the data on the form taken second are discarded, resulting in a considerable loss of data.

Common-item Nonequivalent Groups

In the *common-item nonequivalent groups design*, Form X and Form Y are administered to different (nonequivalent) groups of examinees. The two forms

have items in common. Two variants of this design exist. When using an *internal set of common items*, the common items contribute to the examinee's score on the form. With an internal set, typically, the common items are interspersed with the other items on the test. When using an *external set of common items*, the common items do not contribute to the examinee's score on the form taken. With an external set, the common items typically appear in a separately timed section.

When using the common-item nonequivalent groups design, the common items are used to indicate how different the group of examinees administered Form X is from the group of examinees administered Form Y. Strong statistical assumptions are used to translate the differences between the two groups of examinees on the common items to differences between the two groups on the complete forms.

Because scores on the common items are used to indicate differences between the examinee groups, it is important that the common items fully represent the content of the test forms. Otherwise, a misleading picture of group differences is provided. In addition, it is important that the common items behave in the same manner when they are administered with Form X as with Form Y. So, the common items should be administered in similar positions in the test booklets in the two forms, and the text of the common items should be identical.

Comparison of Equating Designs

The benefits and limitations of the three designs can be compared in terms of ease of test development, ease of administration, security of test administration, strength of statistical assumptions, and sample size requirements. Of the designs considered, the common-item nonequivalent groups design requires the most complex test development process. Common-item sections must be developed that mirror the content of the total test so that the score on the common-item sections can be used to give an accurate reflection of the difference between the group of examinees administered the old form and the group of examinees administered the new form. Test development is less complex for the random groups and single group designs, because there is no need to construct common-item sections.

However, the common-item nonequivalent groups design is the easiest of the three designs to administer.

Only one test form needs to be administered on each test date. For the random groups design, multiple forms must be administered on a test date. For the single group design, each examinee must take two forms, which typically cannot be done in a regular test administration.

The common-item nonequivalent design tends to lead to greater test security than the other designs, because only one form needs to be administered at a given test date. With the random groups and single group designs, multiple forms are administered at a particular test date to conduct equating. However, security issues can be of concern with the common-item nonequivalent groups design, because the common items must be repeatedly administered.

The common-item nonequivalent groups design requires the strongest statistical assumptions. The random groups design requires only weak assumptions, mainly that the random assignment process was successful. The single group design requires stronger assumptions than the random groups design, in that it assumes no differential order effects.

The random groups design requires the largest sample sizes of the three designs. Assuming no differential order effects, the single group design has the smallest sample size requirements of the three designs because, effectively, each examinee serves as his or her own control.

As is evident from the preceding discussion, each of the designs has strengths and weaknesses. The choice of design depends on weighing the strengths and weaknesses with regard to the testing program under consideration. Each of these designs has been used to conduct equating in a variety of testing programs.

Statistical Methods

Equating requires that a relationship between alternate forms be estimated. Equating methods result in a transformation of scores on the alternate forms so that the scores possess specified properties. For traditional equating methods, transformations of scores are found such that for the alternate forms, after equating, the distributions, or central moments of the distributions, are the same in a population of examinees for the forms to be equated.

Traditional observed score equating methods define score correspondence on alternate forms by

setting certain characteristics of score distributions equal for a specified population of examinees. In traditional *equipercentile equating*, a transformation is found such that, after equating, scores on alternate forms have the same distribution in a specified population of examinees. Assume that scores on Form X are to be equated to the raw score scale of Form Y. Define X as the random variable score on Form X, Y as the random variable score on Form Y, F as the cumulative distribution function of X in the population, and G as the cumulative distribution function of Y in the population. Let e_Y be a function that is used to transform scores on Form X to the Form Y raw score scale, and let G^* be the cumulative distribution function of e_Y in the same population. The function e_Y is defined to be the equipercentile equating function in the population if

$$G^* = G. \quad (1)$$

Scores on Form X can be transformed to the Form Y scale using equipercentile equating by taking,

$$e_Y(x) = G^{-1}[F(x)], \quad (2)$$

where x is a particular value of X , and G^{-1} is the inverse of the cumulative distribution function G .

Finding equipercentile equivalents would be straightforward if the distributions of scores were continuous. However, test scores typically are discrete (e.g., number of items correctly answered). To conduct equipercentile equating with discrete scores, the percentile rank of a score on Form X is found for a population of examinees. The equipercentile equivalent of this score is defined as the score on Form Y that has the same percentile rank in the population. Owing to the discreteness of scores, the resulting equated score distributions are only approximately equal.

Because many parameters need to be estimated in equipercentile equating (percentile ranks at each Form X and Form Y score), equipercentile equating is subject to much sampling error. For this reason, smoothing methods are often used to reduce sampling error. In *presmoothing methods*, the score distributions are smoothed. In *posts smoothing methods*, the equipercentile function is smoothed. Kolen and Brennan [5] discussed a variety of smoothing methods. von Davier, Holland, and Thayer [8] presented a comprehensive set of procedures, referred to as **kernel smoothing**, that incorporates procedures for

presmoothing score distributions, handling the discreteness of test score distributions, and estimating standard errors of equating.

Other traditional methods are sometimes used that can be viewed as special cases of the equipercentile method. In *linear equating*, a transformation is found that results in scores on Form X having the same mean and standard deviation as scores on Form Y. Defining $\mu(X)$ as the mean score on Form X, $\sigma(X)$ as the standard deviation of Form X scores, $\mu(Y)$ as the mean score on Form Y, $\sigma(Y)$ as the standard deviation of Form Y scores, and l_Y as the linear equating function,

$$l_Y(x) = \sigma(Y) \left[\frac{x - \mu(X)}{\sigma(X)} \right] + \mu(Y). \quad (3)$$

Unless the shapes of the score distributions for Form X and Form Y are identical, linear and equipercentile methods produce different results. However, even when the shapes of the distributions differ, equipercentile and linear methods produce similar results near the mean. When interest is in scores near the mean, linear equating often is sufficient. However, when interest is in scores all along the score scale and the sample size is large, then equipercentile equating is often preferable to linear equating.

For the random groups and single group designs, the sample data typically are viewed as representative of the population of interest, and the estimation of the traditional equating functions proceeds without needing to make strong statistical assumptions. However, estimation in the common-item nonequivalent groups design requires strong statistical assumptions. First, a population must be specified in order to define the equipercentile or linear equating relationship. Since Form X is administered to examinees from a different population than is Form Y, the population used to define the equating relationship typically is viewed as a combination of these two populations. The combined population is referred to as the *synthetic population*. Three common ways to define the synthetic population are to equally weight the population from which the examinees are sampled to take Form X and Form Y, weight the two populations by their respective sample sizes, or define the synthetic population as the population from which examinees are sampled to take Form X. It turns out that the definition of the synthetic population typically has little effect on the final equating results. Still, it is necessary to define a synthetic population

in order to proceed with traditional equating with this design.

Kolen and Brennan [5] described a few different equating methods for the common-item nonequivalent groups design. The methods differ in terms of their statistical assumptions. Define V as score on the common items. In the Tucker linear method, the linear regression of X on V is assumed to be the same for examinees taking Form X and the examinees taking Form Y. A similar assumption is made about the linear regression of Y on V . In the Levine linear observed score method, similar assumptions are made about true scores, rather than observed scores. No method exists to directly test all of the assumptions that are made using data that are collected for equating. Methods also exist for equipercentile equating under this design that make somewhat different regression assumptions.

Equating Error

Minimizing equating error is a major goal when developing tests that are to be equated, designing equating studies, and conducting equating. *Random equating error* is present whenever samples from populations of examinees are used to estimate equating relationships. Random error depends on the design used for data collection, the score point of interest, the method used to estimate equivalents, and the sample size. Standard errors of equating are used to index random error. Standard error equations have been developed to estimate standard errors for most designs and methods, and resampling methods like the **bootstrap** can also be used. In general, standard errors diminish as sample size increases. Standard errors of equating can be used to estimate required sample sizes for equating, for comparing the precision of various designs and methods, and for documenting the amount of random error in equating.

Systematic equating error results from violations of assumptions of the particular equating method used. For example, in the common-item nonequivalent groups design, systematic error will result if the Tucker method is applied and the regression-based assumptions that are made are not satisfied. Systematic error typically cannot be quantified in operational equating situations.

Equating error of both types needs to be controlled because it can propagate over equatings and result

in scores on later test forms not being comparable to scores on earlier forms. Choosing a large enough sample size given the design is the best way to control random error. To control systematic error, the test must be constructed and the equating implemented so as to minimize systematic error. For example, the assumptions for any of the methods for the common-item nonequivalent groups design tend to hold better when the groups being administered the old and the new form do not differ too much from each other. The assumptions also tend to hold better when the forms to be equated are very similar to one another, and when the content and statistical characteristics of the common items closely represent the content and statistical characteristics of the total test forms. One other way to help control error is to use what is often referred to as *double-linking*. In double-linking, a new form is equated to two previously equated forms. The results for the two equatings often are averaged to produce a more stable equating than if only one previously equated form had been used. Double-linking also provides for a built-in check on the adequacy of the equating.

Selected Practical Issues

Owing to practical constraints, equating cannot be used in some situations where its use might be desirable. Use of any of the equating methods requires test security. In the single group and random groups design, two or more test forms must be administered in a single test administration. If these forms become known to future examinees, then the equating and the entire testing program could be jeopardized. With the common-item nonequivalent groups design, the common items are administered on multiple test dates. If the common items become known to examinees, the equating also is jeopardized. In addition, equating requires that detailed content and statistical test specifications be used to develop the alternate forms. Such specifications are a prerequisite to conducting adequate equating.

Although the focus of this entry has been on equating multiple-choice tests that are scored number-correct, equating often can be used with tests that are scored in other ways such as essay tests scored by human raters. The major problem with equating

such tests is that, frequently, very few essay questions can be administered in a reasonable time frame, which can lead to concern about the comparability of the content from one test form to another. It also might be difficult, or impossible, when the common-item nonequivalent groups design is used to construct common-item sections that represent the content of the complete tests.

Concluding Comments

Test form equating has as its goal to use scores from alternate test forms interchangeably. Test development procedures that have detailed content and statistical specifications allow for the development of alternate test forms that are similar to one another. These test specifications are a necessary prerequisite to the application of equating methods.

References

- [1] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.
- [2] Angoff, W.H. (1971). Scales, norms, and equivalent scores, in *Educational Measurement*, 2nd Edition, R.L. Thorndike, ed., American Council on Education, Washington, pp. 508–600.
- [3] Flanagan, J.C. (1951). Units, scores, and norms, in *Educational Measurement*, E.F. Lindquist, ed., American Council on Education, Washington, pp. 695–793.
- [4] Holland, P.W. & Rubin, D.B. (1982). *Test Equating*, Academic Press, New York.
- [5] Kolen, M.J. & Brennan, R.L. (1995). *Test Equating: Methods and Practices*, Springer-Verlag, New York.
- [6] Kolen, M.J. & Brennan, R.L. (2004). *Test Equating, Scaling and Linking: Methods and Practices*, 2nd Edition, Springer-Verlag, New York.
- [7] Petersen, N.S., Kolen, M.J. & Hoover, H.D. (1989). Scaling, norming, and equating, in *Educational Measurement*, 3rd Edition, R.L. Linn, ed., Macmillan Publishers, New York, pp. 221–262.
- [8] von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004). *The Kernel Method of Test Equating*, Springer-Verlag, New York.

MICHAEL J. KOLEN AND YE TONG

Classification and Regression Trees

BRIAN S. EVERITT

Volume 1, pp. 287–290

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Classification and Regression Trees

There are a variety of regression type models, for example, **multiple linear regression**, **generalized linear models**, **generalized linear mixed effects models**, **generalized additive models** and **nonlinear models**. These models are widely used, but they may not give faithful data descriptions when the assumptions on which they are based are not met, or in the presence of higher-order interactions among some of the explanatory variables. An alternative approach, classification and regression trees (CART), has evolved to overcome some of these potential problems with the more usual types of regression models. The central idea of the CART procedure is the formation of subgroups of individuals within which the response variable is relatively homogeneous. Interpretation in terms of prognostic group identification is frequently possible.

Tree-based Models

Developing a tree model involves the construction of a collection of rules that eventually lead to the terminal nodes. An example of a rule for data consisting of a response variable y and a set of explanatory variables, $x_1, \dots, x_p$, might be if $(x_2 < 410)$ and $(x_4 \in \{C, D, E\})$ and $(x_5 > 10)$ then the predicted value of y is 4.75 (if y is continuous), or, the probability that $y = 1$ is 0.7, if y is binary.

The complete collection of rules that defines the tree is arrived at by a process known as *recursive partitioning*. The essential steps behind this procedure are as follows:

- A series of binary splits is made based on the answers to questions of the type, ‘Is observation or case $i \in A$ ’, where A is a region of the covariate space.
- Answering such a question induces a partition, or split, of the covariate space; cases for which the answer is yes are assigned to one group, and those for which the answer is no to an alternative group.
- Most implementations of tree-modelling proceed by imposing the following constraints.

1. Each split depends on the value of only a single covariate.
 2. For ordered (continuous or categorical) covariates x_j , only splits resulting from questions of the form ‘Is $x_j < C$?’ are considered. Thus, ordering is preserved.
 3. For categorical explanatory variables, all possible splits into disjoint subsets of the categories are allowed.
- A tree is grown as follows:
 1. Examine every allowable split on each explanatory variable.
 2. Select and execute (i.e., create left and right ‘daughter’ nodes) from the best of these splits.
 - The initial or root node of the tree comprises the whole sample. Steps (1) and (2) above are then reapplied to each of the daughter nodes. Various procedures are used to control tree size, as we shall describe later.
 - To determine the best node to split into left and right daughter nodes at any stage in the construction of the tree, involves the use of a numerical *split function* $\phi(s, g)$; this can be evaluated for any split s of node g . The form of $\phi(s, g)$ depends on whether the response variable is continuous or categorical. For a continuous response variable the normal split function is based on the within-node sum of squares, that is, for a node g with N_g cases, the term

$$SS(g) = \sum_{i \in g} [y_i - \bar{y}(g)]^2 \quad (1)$$

where y_i denotes the response variable value for the i th individual and $\bar{y}(g)$ is the mean of the responses of the N_g cases in node g . If a particular split, s , of node g is into left and right daughter nodes, g_L and g_R then the least squares split function is

$$\phi(s, g) = SS(g) - SS(g_L) - SS(g_R) \quad (2)$$

and the best split of node g is determined as the one that corresponds to the maximum of (2) amongst all allowable splits.

For a categorical response variable (in particular, binary variables) split functions are based on trying to make the probability of a particular category of the

2 Classification and Regression Trees

variable close to one or zero in each node. Most commonly used is a log-likelihood function (see **Maximum Likelihood Estimation**) defined for node g as

$$LL(g) = -2 \sum_{i \in g} \sum_{k=1}^K y_{ik} \log(p_{gk}) \quad (3)$$

where K is the number of categories of the response variable, y_{ik} is an indicator variable taking the value 1 if individual i is in category k of the response and zero otherwise, and p_{gk} is the probability of being in the k th category of the response in node g , estimated as n_{gk}/N_g where n_{gk} is the number of individuals in category k in node g . The corresponding split function $\phi(s, g)$ is then simply

$$\phi(s, g) = LL(g) - LL(g_L) - LL(g_k) \quad (4)$$

and again the chosen split is that maximizing $\phi(s, g)$. (The split function $\phi(s, g)$ is often referred to simply as *deviance*.)

Trees are grown by recursively splitting nodes to maximize ϕ , leading to smaller and smaller nodes of progressively increased homogeneity. A critical question is ‘when should tree construction end and terminal nodes be declared?’ Two simple ‘stopping’ rules are as follows:

- *Node size-stop* when this drops below a threshold value, for example, when $N_g < 10$.
- *Node homogeneity* – stop when a node is homogeneous enough, for example, when its deviance is less than 1% of the deviance of the root node.

Neither of these is particularly attractive because they have to be judged relative to preset thresholds, misspecification of which can result in overfitting or underfitting. An alternative more complex approach is to use what is known as a *pruning algorithm*. This involves growing a very large initial tree to capture all potentially important splits, and then collapsing this backup using what is known as *cost complexity pruning* to create a nested sequence of trees.

Cost complexity pruning is a procedure which snips off the least important splits in the initial tree, where importance is judged by a measure of within-node homogeneity or *cost*. For a continuous variable, for example, cost would simply be the sum of squares term defined in (5.1). The cost of the entire tree, G , is then defined as

$$\text{Cost}(G) = \sum_{g \in \tilde{G}} SS(g) \quad (5)$$

where $\tilde{G}$ is the collection of terminal nodes of G . Next we define the complexity of G as the number of its terminal nodes, say $N_{\tilde{G}}$, and finally, we can define the cost-complexity of G as

$$CC_{\alpha}(G) = \text{Cost}(G) + \alpha N_{\tilde{G}} \quad (6)$$

where $\alpha \geq 0$ is called the *complexity parameter*. The aim is to minimize simultaneously both cost and complexity; large trees will have small cost but high complexity with the reverse being the case for small trees. Solely minimizing cost will err on the side of overfitting; for example, with $SS(g)$ we can achieve zero cost by splitting to a point where each terminal node contains only a single observation. In practice we use (6) by considering a range of values of α and for each find the subtree $G(\alpha)$ of our initial tree that minimizes $CC_{\alpha}(G)$. If α is small $G(\alpha)$ will be large, and as α increases, $N_{\tilde{G}}$ decreases. For a sufficiently large α , $N_{\tilde{G}} = 1$.

In this way, we are led to a sequence of possible trees and need to consider how to select the best. There are two possibilities:

- If a separate validation sample is available, we can predict on that set of observations and calculate the deviance versus α for the pruned trees. This will often have a minimum, and so the smallest tree whose sum of squares is close to the minimum can be chosen.
- If no validation set is available, one can be constructed from the observations used in constructing the tree, by splitting the observations into a number of (roughly) equally sized subsets. If n subsets are formed this way, $n - 1$ can be used to grow the tree and it can be tested on the remaining subset. This can be done n ways, and the results averaged.

Full details are available in Breiman et al. [1]

An Example of the Application of the CART Procedure

As an example of the application of tree-based models in a particular area we shall use data on the birthweight of babies given in Hosmer [2]. Birthweight of babies is often a useful indicator of how they will thrive in the first few months of their life. Low birthweight, say below 2.5 kg, is often a cause of concern for their welfare. The part of the data

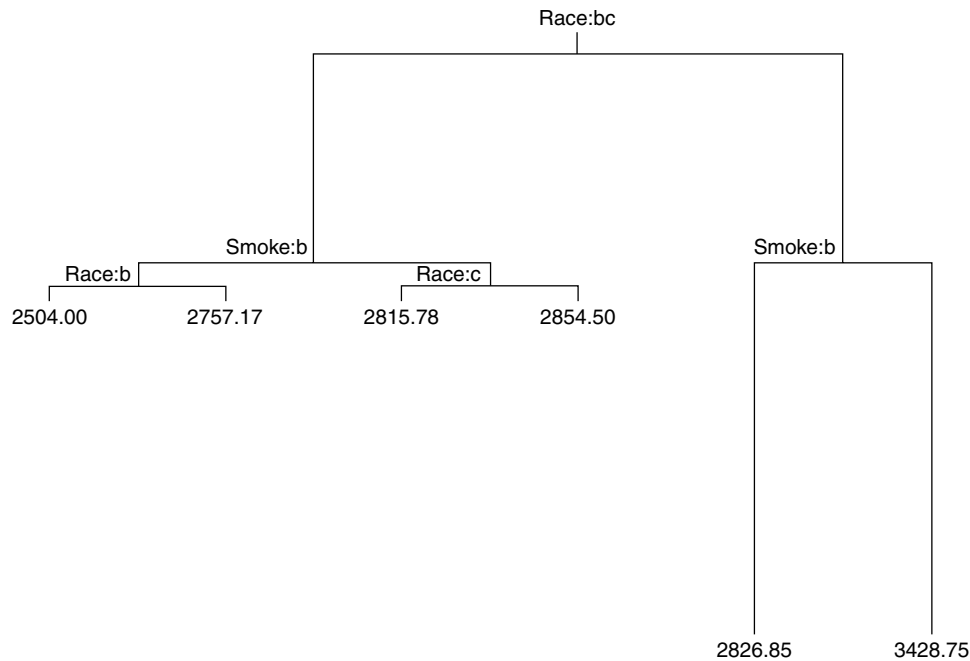


Figure 1 Regression tree to predict birthweight from race and smoking status

with which we will be concerned is that involving the actual birthweight and two explanatory variables, *race* (white/black/other) and *smoke*, a binary variable indicating whether or not the mother was a smoker during pregnancy.

The regression tree for the data can be constructed using suitable software (for example, the `tree` function in S-PLUS (see **Software for Statistical Analyses**) and the tree is displayed graphically in Figure 1. Here, the first split is on race into white and black/other. Each of the new nodes is then further split on the smoke variable into smokers and nonsmokers, and then, in the left-hand side of the tree, further nodes are introduced by splitting race into black and other. The six terminal nodes and their average birthweights are as follows:

1. black, smokers: 2504, $n = 10$;
2. other, smokers: 2757, $n = 12$;
3. other, nonsmokers: 2816, $n = 55$;

4. black, nonsmokers: 2854, $n = 16$;
5. white, smokers: 2827, $n = 52$;
6. white, nonsmokers: 3429, $n = 44$.

Here, there is evidence of a race $\times$ smoke interaction, at least for black and other women. Among smokers, black women produce babies with lower average birthweight than do 'other' women. But for nonsmokers the reverse is the case.

References

- [1] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*, Chapman and Hall/CRC, New York.
- [2] Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*, Wiley, New York.

BRIAN S. EVERITT

Clinical Psychology

TERESA A. TREAT AND V. ROBIN WEERSING

Volume 1, pp. 290–301

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Clinical Psychology

Quantitative sophistication is increasingly central to research in clinical psychology. Both our theories and the statistical techniques available to test our hypotheses have grown in complexity over the last few decades, such that the novice clinical researcher now faces a bewildering array of analytic options. The purpose of this article is to provide a conceptual overview of the use of statistics in clinical science. The first portion of this article describes five major research questions that clinical psychology researchers commonly address and provides a brief overview of the statistical methods that frequently are employed to address each class of questions. These questions are neither exhaustive nor mutually exclusive, but rather are intended to serve as a heuristic for organizing and thinking about classes of research questions in clinical psychology and the techniques most closely associated with them. The second portion of the article articulates guiding principles that underlie the responsible use of statistics in clinical psychology.

Five Classes of Research Questions in Clinical Psychology

Defining and Measuring Constructs

Careful attention to the definition and measurement of constructs is the bread and butter of clinical research. Constructs refer to abstract psychological entities and phenomena such as depression, marital violence, genetic influences, attention to negative information, acculturation, and cognitive-behavioral therapy (CBT). We specify these unobserved variables (*see* **Latent Variable**), as well as their interrelationships, in a theoretical model (e.g., CBT might be assumed to decrease depression in one's partner, which then decreases the likelihood of marital violence in the relationship). Our measurement model (*see* **Measurement: Overview**) specifies the way in which we operationally define the constructs of interest (e.g., our 'measurement variable', or 'indicator variable', for the construct of depression might be patient scores on the Beck Depression Inventory (BDI) [4]). Finally, our analytical model refers to the

way in which we statistically evaluate the hypothesized relationships between our measured variables (e.g., we might use **structural-equation modeling** (SEM), **analysis of variance** (ANOVA), or **logistic regression**). Later in this article, we discuss the importance of the consistency between these three models for making valid inferences about a theoretical model, as well as the importance of 'starting at the top' (i.e., the importance of theory for the rapid advancement of clinical research). Readers are urged to consult McFall and Townsend [36] for a more comprehensive overview of the specification and evaluation of the multiple layers of scientific models in clinical research.

Deciding how best to measure our constructs – that is, specifying the measurement model for the theoretical model of interest – is a critical first step in every clinical research project. Sometimes this step entails a challenging process of thinking logically and theoretically about how best to assess a particular construct. Consider, for example, the difficulty in defining what 'counts' as a suicide attempt. Is any dangerous personal action 'suicidal' (e.g., driving recklessly, jumping from high places, mixing barbiturates and alcohol)? Does the person have to report intending to kill herself, or are others' perceptions of her intention enough? How should intention be assessed in the very young or the developmentally delayed? Does the exhibited behavior have to be immediately life-threatening? What about life-threatening parasuicidal behaviors? Similar difficulties arise in attempting to decide how to assess physical child abuse, cognitive therapy, or an episode of overeating. These examples are intended to highlight the importance of recognizing that all phenomena of interest to clinical researchers are constructs. As a result, theoretical models of a construct and the chosen measurement models always should be distinguished – not collapsed and treated as one and the same thing – and the fit between theoretical and measurement models should be maximized.

More commonly, defining and measuring constructs entails scale development, in which researchers (a) create a set of items that are believed to assess the phenomenon or construct; (b) obtain many participants' responses to these items; and (c) use factor-analytic techniques (*see* **History of Factor Analysis: A Psychological Perspective**) to reduce the complexity of the numerous items to a much

smaller subset of theoretically interpretable constructs, which commonly are referred to as ‘factors’ or ‘latent variables’. For example, Walden, Harris, and Catron [53] used factor analysis when developing ‘How I Feel’, a measure on which children report the frequency and intensity of five emotions (happy, sad, mad, excited, and scared), as well as how well they can control these emotions. The authors generated 30 relevant items (e.g., the extent to which children were ‘scared almost all the time’ during the past three months) and then asked a large number of children to respond to them. **Exploratory factor analyses** of the data indicated that three underlying factors, or constructs, accounted for much of the variability in children’s responses: Positive Emotion, Negative Emotion, and Control. For example, the unobserved Negative Emotion factor accounted particularly well for variability in children’s responses to the sample item above (i.e., this item showed a large factor loading on the Negative Emotion factor, and small factor loadings on the remaining two factors). One particularly useful upshot of conducting a factor analysis is that it produces **factor scores**, which index a participant’s score on each of the underlying latent variables (e.g., a child who experiences chronic sadness over which she feels little control presumably would obtain a high score on the Negative Emotion factor and a lot score on the Control factor). Quantifying factor scores remains a controversial enterprise, however, and researchers who use this technique should understand the relevant issues [20]. Both Reise, Waller, and Comrey [44] and Fabrigar, Wegener, MacCallum, and Strahan [19] provide excellent overviews of the major decisions that clinical researchers must make when using exploratory factor-analytic techniques.

Increasingly, clinical researchers are making use of **confirmatory factor-analytic** techniques when defining and measuring constructs. Confirmatory approaches require researchers to specify both the number of factors and which items load on which factors *prior* to inspection and analysis of the data. Exploratory factor-analytic techniques, on the other hand, allow researchers to base these decisions in large part on what the data indicate are the best answers. Although it may seem preferable to let the data speak for themselves, the exploratory approach capitalizes on sampling variability in the data, and the resulting factor structures may be less likely to cross-validate (i.e., to hold up well in new samples

of data). Thus, when your theoretical expectations are sufficiently strong to place *a priori* constraints on the analysis, it typically is preferable to use the confirmatory approach to evaluate the fit of your theoretical model to the data. Walden et al. [53] followed up the exploratory factor analysis described above by using confirmatory factor analysis to demonstrate the validity and temporal stability of the factor structure for ‘How I Feel’.

Clinical researchers also use **item response theory**, often in conjunction with factor-analytic approaches, to assist in the definition and measurement of constructs [17]. A detailed description of this approach is beyond the scope of this article, but it is helpful to note that this technique highlights the importance of inspecting item-specific measurement properties, such as their difficulty level and their differential functioning as indicators of the construct of interest. For clinical examples of the application of this technique, see [27] and [30].

Cluster analysis is an approach to construct definition and measurement that is closely allied to factor analysis but exhibits one key difference. Whereas factor analysis uncovers unobserved ‘factors’ on the basis of the similarity of variables, cluster analysis uncovers unobserved ‘typologies’ on the basis of the similarity of people. Cluster analysis entails (a) selecting a set of variables that are assumed to be relevant for distinguishing members of the different typologies; (b) obtaining many participants’ responses to these variables; and (c) using cluster-analytic techniques to reduce the complexity among the numerous participants to a much smaller subset of theoretically interpretable typologies, which commonly are referred to as ‘clusters’. Representative recent examples of the use of this technique can be found in [21] and [24]. Increasingly, clinical researchers also are using **latent class analysis** and taxometric approaches to define typologies of clinical interest, because these methods are less descriptive and more model-based than most cluster-analytic techniques. See [40] and [6], respectively, for application of these techniques to defining and measuring clinical typologies.

Evaluating Differences between Either Experimentally Created or Naturally Occurring Groups

After establishing a valid measurement model for the particular theoretical constructs of interest, clinical

researchers frequently evaluate hypothesized group differences in dependent variables (DVs) using one of many analytical models. For this class of questions, group serves as a discrete independent or quasi-independent variable (IV or QIV). In **experimental research**, group status serves as an IV, because participants are assigned randomly to groups, as in randomized controlled trials. In **quasi-experimental research**, in contrast, group status serves as a QIV, because group differences are naturally occurring, as in psychopathology research, which examines the effect of diagnostic membership on various measures. Thus, when conducting quasi-experimental research, it often is unclear whether the QIV (a) ‘causes’ any of the observed group differences; (b) results from the observed group differences; or (c) has an illusory relationship with the DV (e.g., a third variable has produced the correlation between the QIV and the DV). Campbell and Stanley [9] provide an excellent overview of the theoretical and methodological issues surrounding the distinction between quasi-experimental and experimental research and describe the limits of causality inferences imposed by the use of quasi-experimental research designs.

In contrast to the IV or QIV, the DVs can be continuous or discrete and are presumed to reflect the influence of the IV or QIV. Thus, we might be interested in (a) evaluating differences in perfectionism (the DV) for patients who are diagnosed with anorexia versus bulimia (a QIV, because patients are not assigned randomly to disorder type); (b) examining whether the frequency of rehospitalization (never, once, two or more times) over a two-year period (the DV) varies for patients whose psychosis was or was not treated with effective antipsychotic medication during the initial hospitalization (an IV, if drug assignment is random); (c) investigating whether the rate of reduction in hyperactivity (the DV) over the course of psychopharmacological treatment with stimulants is greater for children whose parents are assigned randomly to implement behavioral-modification programs in their homes (an IV); (d) assessing whether the time to a second suicide attempt (the DV) is shorter for patients who exhibit marked, rather than minimal, impulsivity (a QIV); or (e) evaluating whether a 10-day behavioral intervention versus no intervention (an IV) reduces the overall level of a single child’s disruptive behavior (the DV).

What sets apart this class of questions about the influence of an IV or QIV on a DV is the discreteness of the predictor; the DVs can be practically any statistic, whether means, proportions, frequencies, slopes, correlations, time until a particular event occurs, and so on. Thus, many statistical techniques aim to address the same meta-level research question about group differences but they make different assumptions about the nature of the DV. For example, clinical researchers commonly use ANOVA techniques to examine group differences in means (perhaps to answer question 1 above); chi-square or **log-linear** approaches to evaluate group differences in frequencies (question 2; see [52]); **growth-curve** or multilevel modeling (MLM) (see **Hierarchical Models**) techniques to assess group differences in the intercept, slope, or acceleration parameters of a regression line (question 3; see [48] for an example); survival analysis to investigate group differences in the time to event occurrence, or ‘survival time’ (question 4; see [7] and [8]); and **interrupted time-series analysis** to evaluate the effect of an intervention on the level or slope of a single participant’s behavior within a multiple-baseline design (question 5; see [42] for an excellent example of the application of this approach). Thus, these five very different analytical models all aim to evaluate very similar theoretical models about group differences. A common extension of these analytical models provides simultaneous analysis of two or more DVs (e.g., **Multivariate Analysis of Variance (MANOVA)** evaluates mean group differences in two or more DVs).

Many analyses of group differences necessitate inclusion of one or more covariates, or variables other than the IV or QIV that also are assumed to influence the DV and may correlate with the predictor. For example, a researcher might be interested in evaluating the influence of medication compliance (a QIV) on symptoms (the DV), apart from the influence of social support (the covariate). In this circumstance, researchers commonly use **Analysis of Covariance (ANCOVA)** to ‘control for’ the influence of the covariate on the DV. If participants are assigned randomly to levels of the IV, then ANCOVA can be useful for increasing the **power** of the evaluation of the effect of the IV on the DV (i.e., a true effect is more likely to be detected). If, however, participants are not assigned randomly to IV levels and the groups differ on the covariate – a common circumstance in clinical research and a likely

characteristic of the example above – then ANCOVA rarely is appropriate (i.e., this analytical model likely provides an invalid assessment of the researcher's theoretical model). This is an underappreciated matter of serious concern in psychopathology research, and readers are urged to consult [39] for an excellent overview of the relevant substantive issues.

Predicting Group Membership

Clinical researchers are interested not only in examining the effect of group differences on variables of interest (as detailed in the previous section) but also in predicting group differences. In this third class of research questions, group differences become the DV, rather than the IV or QIV. We might be interested in predicting membership in diagnostic categories (e.g., schizophrenic or not) or in predicting important discrete clinical outcomes (e.g., whether a person commits suicide, drops out of treatment, exhibits partner violence, reoffends sexually after mandated treatment, or holds down a job while receiving intensive case-management services). In both cases, the predictors might be continuous, discrete, or a mix of both. **Discriminant function analysis (DFA)** and **logistic regression** techniques commonly are used to answer these kinds of questions. Note that researchers use these methods for a purpose different than that of researchers who use the typology-definition methods discussed in the first section (e.g., cluster analysis, latent class analysis); the focus in this section is on the *prediction* of group membership (which already is known before the analysis), rather than the *discovery* of group membership (which is unknown at the beginning of the analysis).

DFA uses one or more weighted linear combinations of the predictor variables to predict group membership. For example, Hinshaw, Carte, Sami, Treuting, and Zupan [22] used DFA to evaluate how well a class of 10 neuropsychiatric variables could predict the presence or absence of attention-deficit/hyperactivity disorder (ADHD) among adolescent girls. Prior to conducting the DFA, Hinshaw and colleagues took the common first step of using MANOVA to examine whether the groups differed on a linear combination of the class of 10 variables (i.e., they first asked the group-differences question that was addressed in the previous section). Having determined that the groups differed on the class of

variables, as well as on each of the 10 variables in isolation, the authors then used DFA to predict whether each girl did or did not have ADHD. DFA estimated a score for each girl on the weighted linear combination (or discriminant function) of the predictor variables, and the girl's predicted classification was based on whether her score cleared a particular cutoff value that also was estimated in the analysis. The resulting discriminant function, or prediction equation, then could be used in other samples or studies to predict the diagnosis of girls for whom ADHD status was unknown. DFA produces a two-by-two classification table, in which the two dimensions of the table are 'true' and 'predicted' states (e.g., the presence or absence of ADHD). Clinical researchers use the information in this table to summarize the predictive power of the collection of variables, commonly using a percent-correct index, a combination of sensitivity and specificity indices, or a combination of positive and negative predictive power indices. The values of these indices frequently vary as a function of the relative frequency of the two states of interest, as well as the cutoff value used for classification purposes, however. Thus, researchers increasingly are turning to alternative indices without these limitations, such as those drawn from **signal-detection theory** [37].

Logistic regression also examines the prediction of group membership from a class of predictor variables but relaxes a number of the restrictive assumptions that are necessary for the valid use of DFA (e.g., multivariate normality, linearity of relationships between predictors and DV, and homogeneity of variances within each group). Whereas DFA estimates a score for each case on a weighted linear combination of the predictors, logistic regression estimates the probability of one of the outcomes for each case on the basis of a nonlinear (logistic) transformation of a weighted linear combination of the predictors. The predicted classification for a case is based on whether the estimated probability clears an estimated cutoff. Danielson, Youngstrom, Findling, and Calabrese [16] used logistic regression in conjunction with signal-detection theory techniques to quantify how well a behavior inventory discriminated between various diagnostic groups. At this time, logistic regression techniques are preferred over DFA methods, given their less-restrictive assumptions.

Evaluating Theoretical Models That Specify a Network of Interrelated Constructs

As researchers' theoretical models for a particular clinical phenomenon become increasingly sophisticated and complex, the corresponding analytical models also increase in complexity (e.g., evaluating a researcher's theoretical models might require the simultaneous estimation of multiple equations that specify the relationships between a network of variables). At this point, researchers often turn to either multiple-regression models (MRM) (see **Multiple Linear Regression**) or SEM to formalize their analytical models. In these models, constructs with a single measured indicator are referred to as measured (or manifest) variables; this representation of a construct makes the strong assumption that the measured variable is a perfect, error-free indicator of the underlying construct. In contrast, constructs with multiple measured indicators are referred to as latent variables; the assumption in this case is that each measured variable is an imperfect indicator of the underlying construct and the inclusion of multiple indicators helps to reduce error.

MRM is a special case of SEM in which all constructs are treated as measured variables and includes single-equation multiple-regression approaches, **path-analytic methods**, and **linear multilevel models** techniques. Suppose, for example, that you wanted to test the hypothesis that the frequency of negative life events influences the severity of depression, which in turn influences physical health status. MRM would be sufficient to evaluate this theoretical model if the measurement model for each of these three constructs included only a single variable. SEM likely would become necessary if your measurement model for even one of the three constructs included more than one measured variable (e.g., if you chose to measure physical health status with scores on self-report scale as well as by medical record review, because you thought that neither measure in isolation reliably and validly captured the theoretical construct of interest). Estimating SEMs requires the use of specialized software, such as LISREL, AMOS, M-PLUS, Mx, or EQS (see **Structural Equation Modeling: Software**).

Two types of multivariate models that are particularly central to the evaluation and advancement of theory in clinical science are those that specify either **mediation** or **moderation** relationships

between three or more variables [3]. Mediation hypotheses specify a mechanism (B) through which one variable (A) influences another (C). Thus, the example in the previous paragraph proposes that severity of depression (B) mediates the relationship between the frequency of negative life events (A) and physical health (C); in other words, the magnitude of the association between negative life events and physical health should be greatly reduced once depression enters the mix. The strong version of the mediation model states that the A-B-C path is causal and complete – in our example, that negative life events cause depression, which in turn causes a deterioration in physical health – and that the relationship between A and C is completely accounted for by the action of the mediator. Complete mediation is rare in social science research, however. Instead, the weaker version of the mediation model is typically more plausible, in which the association between A and C is reduced significantly (but not eliminated) once the mediator is introduced to the model.

In contrast, moderation hypotheses propose that the magnitude of the influence of one variable (A) on another variable (C) depends on the value of a third variable (B) (i.e., moderation hypotheses specify an interaction between A and B on C). For example, we might investigate whether socioeconomic status (SES) (B) moderates the relationship between negative life events (A) and physical health (C). Conceptually, finding a significant moderating relationship indicates that the A–C relationship holds only for certain subgroups in the population, at least when the moderator is discrete. Such subgroup findings are useful in defining the boundaries of theoretical models and guiding the search for alternative theoretical models in different segments of the population.

Although clinical researchers commonly specify mediation and moderation theoretical models, they rarely design their studies in such a way as to be able to draw strong inferences about the hypothesized theoretical models (e.g., many purported mediation models are evaluated for data collected in **cross-sectional designs** [54], which raises serious concerns from both a logical and data-analytic perspective [14]). Moreover, researchers rarely take all the steps necessary to evaluate the corresponding analytical models. Greater attention to the relevant literature on appropriate statistical evaluation of mediation and moderation

hypotheses should enhance the validity of our inferences about the corresponding theoretical models [3, 23, 28, 29].

In addition to specifying mediating or moderating relationships, clinical researchers are interested in networks of variables that are organized in a nested or hierarchical fashion. Two of the most common **hierarchical**, or multilevel, data structures are (a) nesting of individuals within social groups or organizations (e.g., youths nested within classrooms) or (b) nesting of observations within individuals (e.g., multiple symptoms scores over time nested within patients). Prior to the 1990s, options for analyzing these nested data structures were limited. Clinical researchers frequently collapsed multilevel data into a flat structure (e.g., by disaggregating classroom data to the level of the child or by using difference scores to measure change within individuals). This strategy resulted in the loss of valuable information contained within the nested data structure and, in some cases, violated assumptions of the analytic methods (e.g., if multiple youths are drawn from the same classroom, their scores will likely be correlated and violate independence assumptions). In the 1990s, however, advances in statistical theory and computer power led to the development of MLM techniques. Conceptually, MLM can be thought of as hierarchical multiple regression, in which regression equations are estimated for the smallest (or most nested) unit of analysis and then the parameters of these regression equations are used in second-order analyses. For example, a researcher might be interested in both individual-specific and peer-group influences on youth aggression. In an MLM analysis, two levels of regression equations would be specified: (a) a first-level equation would specify the relationship of individual-level variables to youth aggression (e.g., gender, attention problems, prior history of aggression in a different setting, etc.); and (b) a second-level equation would predict variation in these individual regression parameters as a function of peer-group variables (e.g., the effect of average peer socioeconomic status (SES) on the relationship between gender and aggression). In practice, these two levels are estimated simultaneously. However, given the complexity of the models that can be evaluated using MLM techniques, it is frequently useful to map out each level of the MLM model separately. For a through overview of MLM techniques and available

statistical packages, see the recent text by Raudenbush and Byrk [43], and for recent applications of MLM techniques in the clinical literature, see [41] and [18].

Researchers should be forewarned that numerous theoretical, methodological, and statistical complexities arise when specifying, estimating, and evaluating an analytical model to evaluate a hypothesized network of interrelated constructs, particularly when using SEM methods. Space constraints preclude description of these topics, but researchers who wish to test more complex theoretical models are urged to familiarize themselves with the following particularly important issues: (a) Evaluation and treatment of missing-data patterns; (b) assessment of power for both the overall model and for individual parameters of particular interest; (c) the role of capitalization on chance and the value of cross-validation when respecifying poorly fitting models; (d) the importance of considering different models for the network of variables that make predictions identical to those of the proposed theoretical model; (e) the selection and interpretation of appropriate fit indices; and (f) model-comparison and model-selection procedures (e.g., [2, 14, 25, 32, 33, 34, 51]). Finally, researchers are urged to keep in mind the basic maxim that the strength of causal inferences is affected strongly by research design, and the experimental method applied well is our best strategy for drawing such inferences. MRM and SEM analytical techniques often are referred to as causal models, but we deliberately avoid that language here. These techniques may be used to analyze data from a variety of experimental or quasi-experimental research designs, which may or may not allow you to draw strong causal inferences.

Synthesizing and Evaluating Findings Across Studies or Data Sets

The final class of research questions that we consider is research synthesis or **meta-analysis**. In meta-analyses, researchers describe and analyze empirical findings across studies or datasets. As in any other research enterprise, conducting a meta-analysis (a) begins with a research question and statement of hypotheses; (b) proceeds to data collection, coding, and transformation; and (c) concludes with analysis and interpretation of findings. Meta-analytic investigations differ from other studies in that the unit of

data collection is the *study* rather than the participant. Accordingly, 'data collection' in meta-analysis is typically an exhaustive, well-documented literature search, with predetermined criteria for study inclusion and exclusion (e.g., requiring a minimum sample size or the use of random assignment). Following initial data collection, researchers develop a coding scheme to capture the critical substantive and methodological characteristics of each study, establish the reliability of the system, and code the findings from each investigation. The empirical results of each investigation are transformed into a common metric of effect sizes (see [5] for issues about such transformations). Effect sizes then form the unit of analysis for subsequent statistical tests. These statistical analyses may range from a simple estimate of a population effect size in a set of homogenous studies to a complex multivariate model designed to explain variability in effect sizes across a large, diverse literature.

Meta-analytic inquiry has become a substantial research enterprise within clinical psychology, and results of meta-analyses have fueled some of the most active debates in the field. For example, in the 1980s and 1990s, Weisz and colleagues conducted several reviews of the youth therapy treatment literature, estimated population effect sizes for the efficacy of treatment versus control conditions, and sought to explain variability in these effect sizes in this large and diverse treatment literature (e.g., [56]). Studies included in the meta-analyses were coded for theoretically meaningful variables such as treatment type, target problem, and youth characteristics. In addition, studies were classified comprehensively in terms of their methodological characteristics – from the level of the study (e.g., sample size, type of control group) down to the level of each individual outcome measure, within each treatment group, within each study (e.g., whether a measure was an unnecessarily reactive index of the target problem). This comprehensive coding system allowed the investigators to test the effects of the theoretical variables of primary interest as well as to examine the influence of methodological quality on their findings. Results of these meta-analyses indicated that (a) structured, behavioral treatments outperformed unstructured, nonbehavioral therapies across the child therapy literature; and (b) psychotherapy in everyday community clinic settings was more likely to entail use of nonbehavioral treatments and to have lower effect sizes than those seen in research studies of

behavioral therapies (e.g., [55]). The debate provoked by these meta-analytic findings continues, and the results have spurred research on the moderators of therapy effects and the dissemination of evidence-based therapy protocols to community settings.

As our example demonstrates, meta-analysis can be a powerful technique to describe and explain variability in findings across an entire field of inquiry. However, meta-analysis is subject to the same limitations as other analytic techniques. For example, the effects of a meta-analysis can be skewed by biased sampling (e.g., an inadequate literature review), use of a poor measurement model (e.g., an unreliable scheme for coding study characteristics), low power (e.g., an insufficiently large literature to support testing cross-study hypotheses), and data-quality problems (e.g., a substantial portion of the original studies omit data necessary to evaluate meta-analytic hypotheses, such as a description of the ethnicity of the study sample). Furthermore, most published meta-analyses do not explicitly model the nested nature of their data (e.g., effect sizes on multiple symptom measures are nested within treatment groups, which are nested within studies). Readers are referred to the excellent handbook by Cooper and Hedges [15] for a discussion of these and other key issues involved in conducting a meta-analysis and interpreting meta-analytic data.

Overarching Principles That Underlie the Use of Statistics in Clinical Psychology

Having provided an overview of the major research questions and associated analytical techniques in clinical psychology, we turn to a brief explication of four principles and associated corollaries that characterize the responsible use of statistics in clinical psychology. The intellectual history of these principles draws heavily from the work and insight of such luminaries as **Jacob Cohen**, Alan Kazdin, Robert McCallum, and Paul Meehl. Throughout this section, we refer readers to more lengthy articles and texts that expound on these principles.

Principle 1: The specification and evaluation of theoretical models is critical to the rapid advancement of clinical research.

Corollary 1: Take specification of theoretical, measurement, and analytical models seriously. As theoretical models specify unobserved constructs and

their interrelationships (see earlier section on defining and measuring constructs), clinical researchers must draw inferences about the validity of their theoretical models from the fit of their analytical models. Thus, the strength of researchers' theoretical inferences depends critically on the consistency of the measurement and analytical models with the theoretical models [38]. Tightening the fit between these three models may preclude the use of 'off-the-shelf' measures or analyses, when existing methods do not adequately capture the constructs or their hypothesized interrelationships. For example, although more than 25 years of research document the outstanding psychometric properties of the BDI, the BDI emphasizes the cognitive and affective aspects of the construct of depression more than the vegetative and behavioral aspects. This measurement model may be more than sufficient for many investigations, but it would not work well for others (e.g., a study targeting sleep disturbance). Neither measurement nor analytical models are 'assumption-free', so we must attend to the psychometrics of measures (e.g., their reliability and validity), as well as to the assumptions of analytical models. Additionally, we must be careful to maintain the distinctions among the three models. For example, clinical researchers tend to collapse the theoretical and measurement models as work progresses in a particular area (e.g., we reify the construct of depression as the score on the BDI). McFall and Townsend [36] provide an eloquent statement of this and related issues.

Corollary 2: Pursue theory-driven, deductive approaches to addressing research questions whenever possible, and know the limitations of relying on more inductive strategies. Ad hoc storytelling about the results of innumerable exploratory data analyses is a rampant research strategy in clinical psychology. Exploratory research and data analysis often facilitate the generation of novel theoretical perspectives, but it is critical to replicate the findings and examine the validity of a new theoretical model further before taking it too seriously.

Principle 2: The heart of the clinical research enterprise lies in model (re-)specification, evaluation, and comparison.

Corollary 1: Identify the best model from a set of plausible alternatives, rather than evaluating the adequacy of a single model. Clinical researchers often

evaluate a hypothesized model only by comparing it to models of little intrinsic interest, such as a null model that assumes that there is no relationship between the variables or a saturated model that accounts perfectly for the observed data. Serious concerns still may arise in regard to a model that fits significantly better than the null model and nonsignificantly worse than the saturated model, however, (see [51] for an excellent overview of the issues that this model-fitting strategy raises). For example, a number of equivalent models may exist that make predictions identical to those of the model of interest [34]. Alternatively, nonequivalent alternative models may account as well or better for the observed data. Thus, methodologists now routinely recommend that researchers specify and contrast competing theoretical models (both equivalent and nonequivalent) because this forces the researcher to specify and evaluate a variety of theoretically based explanations for the anticipated findings [34, 51].

Corollary 2: Model modifications may increase the validity of researchers' theoretical inferences, but they also may capitalize on sampling variability. When the fit of a model is less than ideal, clinical researchers often make post hoc modifications to the model that improve its fit to the observed data set. For example, clinical researchers who use SEM techniques often delete predictor variables, modify the links between variables, or alter the pattern of relationships between error terms. Other analytic techniques also frequently suffer from similar overfitting problems (e.g., stepwise regression (see **Regression Models**), DFA). These data-driven modifications improve the fit of the model significantly and frequently can be cast as theoretically motivated. However, these changes may do little more than capitalize on systematic but idiosyncratic aspects of the sample data, in which case the new model may not generalize well to the population as a whole [33, 51]. Thus, it is critical to cross-validate respecified models by evaluating their adequacy with data from a new sample; alternatively, researchers might develop a model on a randomly selected subset of the sample and then cross-validate the resulting model on the remaining participants. Moreover, to be more certain that the theoretical assumptions about the need for the modifications are on target, it is

important to evaluate the novel theoretical implications of the modified model with additional data sets.

Principle 3: Mastery of research design and the mechanics of statistical techniques is critical to the validity of researchers' statistical inferences.

Corollary 1: Know your data. Screening data is a critical first step in the evaluation of any analytical model. Inspect and address patterns of missing data (e.g., pair-wise deletion, list-wise deletion, estimation of missing data). Evaluate the assumptions of statistical techniques (e.g., normality of distributions of errors, absence of outliers, linearity, homogeneity of variances) and resolve any problems (e.g., make appropriate data transformations, select alternative statistical approaches). Tabachnick and Fidell [50] provide an outstanding overview of the screening process in the fourth chapter of their multivariate text.

Corollary 2: Know the power of your tests. Jacob Cohen [10] demonstrated more than four decades ago that the **power** to detect hypothesized effects was dangerously low in clinical research, and more recent evaluations have come to shockingly similar conclusions [47, 49]. Every clinical researcher should understand how sample size, effect size, and α affect power; how low power increases the likelihood of erroneously rejecting our theoretical models; and how exceedingly high power may lead us to retain uninteresting theoretical models. Cohen's [12] power primer is an excellent starting place for the faint of heart.

Corollary 3: Statistics can never take you beyond your methods. First, remember GIGO (garbage in–garbage out): Running statistical analyses on garbage measures invariably produces garbage results. Know and care deeply about the psychometric properties of your measures (e.g., various forms of reliability, validity, and **generalizability**; see [26] for a comprehensive overview). Second, note that statistical techniques rarely can eliminate confounds in your research design (e.g., it is extremely difficult to draw compelling causal inferences from quasi-experimental research designs). If your research questions demand quasi-experimental methods, familiarize yourself with designs that minimize threats to

the internal and external validity of your conclusions [9, 26].

Principle 4: Know the limitations of Null-Hypothesis Statistical Testing (NHST).

Corollary 1: The alternative or research hypotheses tested within the NHST framework are very imprecise and almost always true at a population level. With enough power, almost any two means will differ significantly, and almost any two variables will show a statistically significant correlation. This weak approach to the specification and evaluation of theoretical models makes it very difficult to reject or falsify a theoretical model, or to distinguish between two theoretical explanations for the same phenomena. Thus, clinical researchers should strive to develop and evaluate more precise and risky predictions about clinical phenomena than those traditionally examined with the NHST framework [11, 13, 31, 38]. When the theoretical models in a particular research area are not advanced enough to allow more precise predictions, researchers are encouraged to supplement NHST results by presenting **confidence intervals** around sample statistics [31, 35].

Corollary 2: P values do not tell you the likelihood that either the null or alternative hypothesis is true. *P* values specify the likelihood of observing your findings if the null hypothesis is true – *not* the likelihood that the null hypothesis is true, given your findings. Similarly, $(1.0 - p)$ is not equivalent to the likelihood that the alternative hypothesis is true, and larger values of $(1.0 - p)$ do not mean that the alternative hypothesis is more likely to be true [11, 13]. Thus, as Abelson [1] says, 'Statistical techniques are aids to (hopefully wise) judgment, not two-valued logical declarations of truth or falsity' (p. 9–10).

Corollary 3: Evaluate practical significance as well as statistical significance. The number of 'tabular asterisks' in your output (i.e., the level of significance of your findings) is influenced strongly by your sample size and indicates more about reliability than about the practical importance of your findings [11, 13, 38]. Thus, clinical researchers should report information on the practical significance, or magnitude, of their effects, typically by presenting effect-size indices and the confidence intervals around them [13,

45, 46]. Researchers also should evaluate the adequacy of an effect's magnitude by considering the domain of application (e.g., a small but reliable effect size on mortality indices is nothing to scoff at!).

Conclusions

Rapid advancement in the understanding of complex clinical phenomena places heavy demands on clinical researchers for thoughtful articulation of theoretical models, methodological expertise, and statistical rigor. Thus, the next generation of clinical psychologists likely will be recognizable in part by their quantitative sophistication. In this article, we have provided an overview of the use of statistics in clinical psychology that we hope will be particularly helpful for students and early career researchers engaged in advanced statistical and methodological training. To facilitate use for teaching and training purposes, we organized the descriptive portion of the article around core research questions addressed in clinical psychology, rather than adopting alternate organizational schemes (e.g., grouping statistical techniques on the basis of mathematical similarity). In the second portion of the article, we synthesized the collective wisdom of statisticians and methodologists who have been critical in shaping our own use of statistics in clinical psychological research. Readers are urged to consult the source papers of this section for thoughtful commentary relevant to all of the issues raised in this article.

References

- [1] Abelson, R.P. (1995). *Statistics as Principled Argument*, Lawrence Erlbaum Associates, Hillsdale.
- [2] Allison, P.D. (2003). Missing data techniques for structural equation modeling, *Journal of Abnormal Psychology* **112**, 545–557.
- [3] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [4] Beck, A.T., Steer, R.A. & Brown, G.K. (1996). *Manual for the Beck Depression Inventory*, 2nd Edition, The Psychological Corporation, San Antonio.
- [5] Becker, B.J., ed. (2003). Special section: metric in meta-analysis, *Psychological Methods* **8**, 403–467.
- [6] Blanchard, J.J., Gangestad, S.W., Brown, S.A. & Horan, W.P. (2000). Hedonic capacity and schizotypy revisited: a taxometric analysis of social anhedonia, *Journal of Abnormal Psychology* **109**, 87–95.
- [7] Brent, D.A., Holder, D., Kolko, D., Birmaher, B., Baugher, M., Roth, C., Iyengar, S. & Johnson, B.A. (1997). A clinical psychotherapy trial for adolescent depression comparing cognitive, family, and supportive therapy, *Archives of General Psychiatry* **54**, 877–885.
- [8] Brown, G.K., Beck, A.T., Steer, R.A. & Grisham, J.R. (2000). Risk factors for suicide in psychiatric outpatients: a 20-year prospective study, *Journal of Consulting & Clinical Psychology* **68**, 371–377.
- [9] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-experimental Designs for Research*, Rand McNally, Chicago.
- [10] Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review, *Journal of Abnormal and Social Psychology* **65**, 145–153.
- [11] Cohen, J. (1990). Things I have learned (so far), *American Psychologist* **45**, 1304–1312.
- [12] Cohen, J. (1992). A power primer, *Psychological Bulletin* **112**, 155–159.
- [13] Cohen, J. (1994). The earth is round, *American Psychologist* **49**, 997–1003.
- [14] Cole, D.A. & Maxwell, S.E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling, *Journal of Abnormal Psychology* **112**, 558–577.
- [15] Cooper, H. & Hedges, L.V., eds (1994). *The Handbook of Research Synthesis*, Sage, New York.
- [16] Danielson, C.K., Youngstrom, E.A., Findling, R.L. & Calabrese, J.R. (2003). Discriminative validity of the general behavior inventory using youth report, *Journal of Abnormal Child Psychology* **31**, 29–39.
- [17] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Hillsdale.
- [18] Espelage, D.L., Holt, M.K. & Henkel, R.R. (2003). Examination of peer-group contextual effects on aggression during early adolescence, *Child Development* **74**, 205–220.
- [19] Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research, *Psychological Methods* **4**, 272–299.
- [20] Grice, J.W. (2001). Computing and evaluating factor scores, *Psychological Methods* **6**, 430–450.
- [21] Grilo, C.M., Masheb, R.M. & Wilson, G.T. (2001). Subtyping binge eating disorder, *Journal of Consulting and Clinical Psychology* **69**, 1066–1072.
- [22] Hinshaw, S.P., Carte, E.T., Sami, N., Treuting, J.J. & Zupan, B.A. (2002). Preadolescent girls with attention-deficit/hyperactivity disorder: II. Neuropsychological performance in relation to subtypes and individual classification, *Journal of Consulting and Clinical Psychology* **70**, 1099–1111.
- [23] Holmbeck, G.N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and

- moderators: examples from the child-clinical and pediatric psychology literature, *Journal of Consulting and Clinical Psychology* **65**, 599–610.
- [24] Holtzworth-Munroe, A., Meehan, J.C., Herron, K., Rehman, U. & Stuart, G.L. (2000). Testing the Holtzworth-Munroe and Stuart (1994) Batterer Typology, *Journal of Consulting and Clinical Psychology* **68**, 1000–1019.
- [25] Hu, L. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification, *Psychological Methods* **3**, 424–452.
- [26] Kazdin, A.E. (2003). *Research Design in Clinical Psychology*, Allyn and Bacon, Boston.
- [27] Kim, Y., Pilkonis, P.A., Frank, E., Thase, M.E. & Reynolds, C.F. (2002). Differential functioning of the beck depression inventory in late-life patients: use of item response theory, *Psychology & Aging* **17**, 379–391.
- [28] Kraemer, H.C., Stice, E., Kazdin, A., Offord, D. & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors, *American Journal of Psychiatry* **158**, 848–856.
- [29] Kraemer, H.C., Wilson, T., Fairburn, C.G. & Agras, W.S. (2002). Mediators and moderators of treatment effects in randomized clinical trials, *Archives of General Psychiatry* **59**, 877–883.
- [30] Lambert, M.C., Schmitt, N., Samms-Vaughan, M.E., An, J.S., Fairclough, M. & Nutter, C.A. (2003). Is it prudent to administer all items for each child behavior checklist cross-informant syndrome? Evaluating the psychometric properties of the youth self-report dimensions with confirmatory factor analysis and item response theory, *Psychological Assessment* **15**, 550–568.
- [31] Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data, *Current Directions in Psychological Science* **5**, 161–171.
- [32] MacCallum, R.C. & Austin, J.T. (2000). Applications of structural equation modeling in psychological research, *Annual Review of Psychology* **51**, 201–226.
- [33] MacCallum, R.C., Roznowski, M. & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance, *Psychological Bulletin* **111**, 490–504.
- [34] MacCallum, R.C., Wegener, D.T., Uchino, B.N. & Fabrigar, L.R. (1993). The problem of equivalent models in applications of covariance structure analysis, *Psychological Bulletin* **114**, 185–199.
- [35] Masson, M.E.J. & Loftus, G.R. (2003). Using confidence intervals for graphically based data interpretation, *Canadian Journal of Experimental Psychology* **57**, 203–220.
- [36] McFall, R.M. & Townsend, J.T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science, *Psychological Assessment* **10**, 316–330.
- [37] McFall, R.M. & Treat, T.A. (1999). Quantifying the information value of clinical assessments with signal detection theory, *Annual Review of Psychology* **50**, 215–241.
- [38] Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology* **46**, 806–834.
- [39] Miller, G.A. & Chapman, J.P. (2001). Misunderstanding analysis of covariance, *Journal of Abnormal Psychology* **110**, 40–48.
- [40] Nelson, C.B., Heath, A.C. & Kessler, R.C. (1998). Temporal progression of alcohol dependence symptoms in the U.S. household population: results from the national comorbidity survey, *Journal of Consulting and Clinical Psychology* **66**, 474–483.
- [41] Peeters, F., Nicolson, N.A., Berkhof, J., Delespaul, P. & deVries, M. (2003). Effects of daily events on mood states in major depressive disorder, *Journal of Abnormal Psychology* **112**, 203–211.
- [42] Quesnel, C., Savard, J., Simard, S., Ivers, H. & Morin, C.M. (2003). Efficacy of cognitive-behavioral therapy for insomnia in women treated for nonmetastatic breast cancer, *Journal of Consulting and Clinical Psychology* **71**, 189–200.
- [43] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, Thousand Oaks.
- [44] Reise, S.P., Waller, N.G. & Comrey, A.L. (2000). Factor analysis and scale revision, *Psychological Assessment* **12**, 287–297.
- [45] Rosenthal, R., Rosnow, R.L. & Rubin, D.B. (2000). *Contrasts and Effect Sizes in Behavioral Research*, Cambridge University Press, Cambridge.
- [46] Rosnow, R.L. & Rosenthal, R. (2003). Effect sizes for experimenting psychologists, *Canadian Journal of Experimental Psychology* **57**, 221–237.
- [47] Rossi, J.S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology* **58**, 646–656.
- [48] Scott, K.L. & Wolfe, D.A. (2003). Readiness to change as a predictor of outcome in batterer treatment, *Journal of Consulting & Clinical Psychology* **71**, 879–889.
- [49] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* **105**, 309–316.
- [50] Tabachnick, B.G. & Fidell, L.S. (2001). *Using Multivariate Statistics*, Allyn and Bacon, Boston.
- [51] Tomarken, A.J. & Waller, N.G. (2003). Potential problems with “well fitting” models, *Journal of Abnormal Psychology* **112**, 578–598.
- [52] von Eye, A. & Schuster, C. (2002). Log-linear models for change in manifest categorical variables, *Applied Developmental Science* **6**, 12–23.
- [53] Walden, T.A., Harris, V.S. & Catron, T.F. (2003). How I feel: a self-report measure of emotional arousal and regulation for children, *Psychological Assessment* **15**, 399–412.
- [54] Weersing, V. & Weisz, J.R. (2002). Mechanisms of action in youth psychotherapy, *Journal of Child Psychology & Psychiatry & Allied Disciplines* **43**, 3–29.

12 Clinical Psychology

- [55] Weisz, J.R., Donenberg, G.R., Han, S.S. & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy, *Journal of Consulting and Clinical Psychology* **63**, 688–701.
- [56] Weisz, J.R., Weiss, B., Han, S.S., Granger, D.A. & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies, *Psychological Bulletin* **117**, 450–468.

TERESA A. TREAT AND V. ROBIN WEERSING

Clinical Trials and Intervention Studies

EMMANUEL LESAFFRE AND GEERT VERBEKE

Volume 1, pp. 301–305

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Clinical Trials and Intervention Studies

The Intervention Study

In a controlled observational **cohort study**, two groups of subjects are selected from two populations that (hopefully) differ in only one characteristic at the start. The groups of subjects are studied for a specific period and contrasted at the end of the study period. For instance, smokers and nonsmokers are studied for a period of 10 years, and at the end the proportions of smokers and nonsmokers that died in that period are compared. On the other hand, in an intervention study, the subjects are selected from one population with a particular characteristic present; then, immediately after baseline, the total study group is split up into a group that receives the intervention and a group that does not receive that intervention (control group). The comparison of the outcomes of the two groups at the end of the study period is an evaluation of the intervention. For instance, smokers can be divided into those who will be subject to a smoking-cessation program and those who will not be motivated to stop smoking.

Interventions have the intention to improve the condition of an individual or a group of individuals. Some examples of intervention studies in public health research are studies that evaluate the impact of a program: (a) to promote a healthier lifestyle (avoiding smoking, reducing alcohol drinking, increasing physical activity, etc.), (b) to prevent HIV-transmission, (c) to start brushing teeth early in babies, and so on. Ample intervention studies can also be found in other disciplines; two examples illustrate this. First, Palmer, Brown, and Barrera [4] report on an intervention study that tests a short-term group program for abusive husbands against a control program. The two groups are compared with respect to the recidivism rates of the men regarding abuse of their female partners. Second, Moens et al. [1] evaluated in a controlled intervention study the effect of teaching of how to lift and transfer patients to nursing students in a nursing school. After two years of follow-up, the incidence risk of one or more episodes of back pain was compared between the two groups of nursing students.

Controlled clinical trials constitute a separate but important class of intervention studies. There, the aim is to compare the effectiveness and safety of two (or more) medical treatments or surgical operations or combinations thereof. Clearly, now the target population constitutes patients with a specific disease or symptom. More aspects of clinical trials will be highlighted in section 'Typical Aspects of Clinical Trials'.

Intervention studies are often applied on an individual level but they can also be applied on a group level. For instance, promoting better brushing habits for children could be done on an individual basis, for example, by means of a personal advice to the parents of the child, or on a group basis, for example, by introducing special courses on good brushing habits in school. Intervention studies operating on a group level need dedicated statistical methods. We will start with the intervention studies on individual level but come back to intervention studies on group level in section 'Intervention Studies on Group Level'.

Basic Aspects of an Intervention Study

The first step in any intervention study is to specify the target population, which is the population to which the findings should be extrapolated. This requires a specific definition of the subjects in the study prior to selection. In a clinical trial, this is achieved by specifying inclusion and exclusion criteria. In general, the inclusion criteria specify the type of patients who need the treatment under examination and the exclusion criteria exclude patients for which there will be most likely safety concerns or for which the treatment effect might not be clear, for example, because they are already on another, competing, treatment.

To obtain a clear idea about the effect of the intervention, the two groups (intervention and control) should be comparable at the start. More specifically, at baseline, the two groups should be selected from the same population – only in that case a difference between the two groups at the end of the study is a sign of an effect of the intervention. Comparability or balance at baseline is achieved by randomly allocating subjects to the two groups; this is known as randomization. Simple **randomization** corresponds to tossing a coin and when (say) heads, the subject will receive the intervention and in the other

case (s)he will be in the control group. But other randomization schemes exist, like block- and stratified randomization (see **Block Random Assignment; Stratification**). It is important to realize that randomization can only guarantee balance for large studies and that random imbalance can often occur in small studies.

For several types of intervention studies, balance at baseline is a sufficient condition for an interpretable result at the end. However, in a clinical trial we need to be more careful. Indeed, while most interventions aim to achieve a change in attitude (a psychological effect), medical treatments need to show their effectiveness apart from their psychological impact, which is also called the *placebo effect*. The **placebo effect** is the pure psychological effect that a medical treatment can have on a patient. This effect can be measured by administering placebo (inactive medication with the same taste, texture, etc. as the active medication) to patients who are blinded for the fact that they haven't received active treatment. Placebo-controlled trials, that is, trials with a placebo group as control, are quite common. When only the patient is unaware of the administered treatment, the study is called *single-blinded*. Sometimes, also the treating physician needs to be blinded, if possible, in order to avoid bias in scoring the effect and safety of the medication. When patients as well as physician(s) are blinded, we call it a double-blinded clinical trial. Such a trial allows distinguishing the biological effect of a drug from its psychological effect.

The advantage of randomization (plus blinding in a clinical trial) is that the analysis of the results can often be done with simple statistical techniques such as an unpaired *t* Test for continuous measurements or a chi-squared test for categorical variables. This is in contrast to the analysis of controlled observational cohort studies where **regression models** are needed to take care of the imbalance at baseline since subjects are often self-selected in the two groups.

To evaluate the effect of the intervention, a specific outcome needs to be chosen. In the context of clinical trials, this outcome is called the *endpoint*. It is advisable to choose one endpoint, the primary endpoint, to avoid multiple-testing issues. If this is not possible, then a correction for multiple testing such as a Bonferroni adjustment (see **Multiple Comparison Procedures**) is needed. The choice of the primary endpoint has a large impact on the design of the study, as will be exemplified in the section 'Typical Aspects

of Clinical Trials'. Further, it is important that the intervention study is able to detect the anticipated effect of the intervention with a high probability. To this end, the necessary sample size needs to be determined such that the **power** is high enough (in clinical trials, the minimal value nowadays equals 0.80).

Although not a statistical issue, it is clear that any intervention study should be ethically sound. For instance, an intervention study is being set up in South Africa where on the one hand adolescents are given guidelines of how to avoid HIV-transmission and on the other hand, for ethical reasons, adolescents are given general guidelines to live a healthier life (like no smoking, etc.). In clinical trials, ethical considerations are even more of an issue. Therefore, patients are supposed to sign an informed consent document.

Typical Aspects of Clinical Trials

The majority of clinical trials are drug trials. It is important to realize that it takes many years of clinical research and often billions of dollars to develop and register a new drug. In this context, clinical trials are essential, partly because regulatory bodies like the Food and Drug Administration (FDA) in the United States and the European Medicine Agency (EMA) in Europe have imposed stringent criteria on the pharmaceutical industry before a new drug can be registered. Further, the development of a new drug involves different steps such that drug trials are typically subdivided into phases. Four phases are often distinguished. Phase I trials are small, often involve volunteers, and are designed to learn about the drug, like establishing a safe dose of the drug, establishing the schedule of administration, and so on. Phase II trials build on the results of phase I trials and study the characteristics of the medication with the purpose to examine if the treatment should be used in large-scale randomized studies. Phase II designs usually involve patients, are sometimes double blind and randomized, but most often not placebo-controlled. When a drug shows a reasonable effect, it is time to compare it to a placebo or standard treatment; this is done in a phase III trial. This phase is the most rigorous and extensive part of the investigation of the drug. Most often, phase III studies are double-blind, controlled, randomized, and involve many centers (often hospitals); it is the typical controlled clinical trial as

introduced above. The size of a phase III trial will depend on the anticipated effect of the drug. Such studies are the basis for registration of the medication. After approval of the drug, large-scale studies are needed to monitor for (rare) adverse effects; they belong to the phase IV development stage.

The typical clinical trial design varies with the phase of the drug development. For instance, in phase I studies, an **Analysis of variance** design comparing the different doses is often encountered. In phase II studies, **crossover designs**, whereby patients are randomly assigned to treatment sequences, are common. In phase III studies, the most common design is the simple parallel-group design where two groups of patients are studied over time after drug administration. Occasionally, three or more groups are compared; when two (or more) types of treatments are combined, a factorial design is popular allowing the estimation of the effects of each type of treatment.

Many phase III trials need a lot of patients and take a long time to give a definite answer about the efficacy of the new drug. For economic as well as ethical reasons, one might be interested in having an idea of the effect of the new drug before the planned number of patients is recruited and/or is studied over time. For this reason, one might want to have interim looks at the data, called *interim analyses*. A clinical trial with planned interim analyses has a so-called group-sequential design indicating that specific statistical (correction for multiple testing) and practical (interim meetings and reports) actions are planned. Usually, this is taken care of by an independent committee, called the *Data and Safety Monitoring Board* (DSMB). The DSMB consists of clinicians and statisticians overlooking the efficacy but especially the safety of the new drug.

Most of the clinical trials are superiority trials with the aim to show a better performance of the new drug compared to the control drug. When the control drug is not placebo but a standard active drug, and it is conceived to be difficult to improve upon the efficacy of that standard drug, one might consider showing that the new drug has comparable efficacy. When the new drug is believed to have comparable efficacy and has other advantages, for example, a much cheaper cost, a noninferiority trial is an option. For a noninferiority trial, the aim is to show that the new medication is not (much) worse than the standard treatment (*see Equivalence Trials*).

Currently, noninferiority trials are becoming quite frequent due to the difficulty to improve upon existing therapies.

The choice of the primary endpoint can have a large impact on the design of the study. For instance, changing from a binary outcome evaluating short-term survival (say at 30 days) to survival time as endpoint not only changes the statistical test from a chi-square test to, say, a logrank test but can also have a major practical impact on the trial. For instance, with long-term survival as endpoint, a group-sequential design might become a necessity.

Despite the fact that most clinical trials are carefully planned, many problems can occur during the conduct of the study. Some examples are as follows: (a) patients who do not satisfy the inclusion and/or exclusion criteria are included in the trial; (b) a patient is randomized to treatment A but has been treated with B; (c) some patients drop out from the study; (d) some patients are not compliant, that is, do not take their medication as instructed, and so on. Because of these problems, one might be tempted to restrict the comparison of the treatments to the ideal patients, that is, those who adhered perfectly to the clinical trial instructions as stipulated in the protocol. This population is classically called the *per-protocol population* and the analysis is called the *per-protocol analysis*. A per-protocol analysis envisages determining the biological effect of the new drug. However, by restricting the analysis to a selected patient population, it does not show the practical value of the new drug. Therefore, regulatory bodies push the **intention-to-treat** (ITT) analysis forward. In the ITT population, none of the patients is excluded and patients are analyzed according to the randomization scheme. Although medical investigators have often difficulties in accepting the ITT analysis, it is the pivotal analysis for FDA and EMEA.

Although the statistical techniques employed in clinical trials are often quite simple, recent statistical research tackled specific and difficult clinical trial issues, like dropouts, compliance, noninferiority studies, and so on. Probably the most important problem is the occurrence of dropout in a clinical trial. For instance, when patients drop out before a response can be obtained, they cannot be included in the analysis, even not in an ITT analysis. When patients are examined on a regular basis, a series of measurements is obtained. In that case, the measurements obtained before the patient dropped out can be used

to establish the unknown measurement at the end of the study. FDA has been recommending for a long time the Last-Observation-Carried-Forward (LOCF) method. Recent research shows that this method gives a biased estimate of the treatment effect and underestimates the variability of the estimated result [6]. More sophisticated methods are reviewed in [7] (*see Missing Data*).

Intervention Studies on Group Level

Many intervention studies act on the group level; they are called *group-randomized studies*. For instance, Murray et al. [2] describe an intervention study to evaluate four interventions to reduce the tobacco use among adolescents. Forty-eight schools were randomized to the four interventions. After two years of follow-up, the proportion of children using 'smokeless tobacco' was compared. The proportions found in two of the four treatment groups were $58/1341 = 0.043$ and $91/1479 = 0.062$. A simple chi-square test gives a P value of 0.03. However, this test assumes independence of the subjects. When adolescents are motivated in a school context, there will be a high interaction among adolescents of the same class/school, that is, the outcome of one adolescent will depend on the outcome of another adolescent. Hence, the chi-square test is not appropriate. An adjusted chi-square test taking the correlation among the adolescents into account (see [2]) gives a P value of 0.18.

In general, the appropriate statistical techniques for group-randomized studies need to take the correlation among subjects in the same group into account (*see Intraclass Correlation*). This implies the use of techniques like **Generalized Estimating Equations** (GEE) and random effects models; see, for

example, [7] and **Linear Multilevel Models; Generalized Linear Models (GLM)**.

Further Reading

An excellent source for clinical trial methodology can be found in [5]. Intervention studies operating on group level gained in importance the last decade; for an overview of these designs, we refer to [3].

References

- [1] Moens, G., Johannik, K., Dohogne, T. & Vandepoele, G. (2002). The effectiveness of teaching appropriate lifting and transfer techniques to nursing students: results after two years of follow-up, *Archives of Public Health* **60**, 115–223.
- [2] Murray, D.M., Hannan, P.J. & Zucker, D. (1998). Analysis issues in school-based health promotion studies, *Health Education Quarterly* **16**(2), 315–330.
- [3] Murray, D.M. (1998). *Design and Analysis of Group-randomized Trials*, Monographs in Epidemiology and Biostatistics, 27, Oxford University Press, New York.
- [4] Palmer, S.E., Brown, R.A. & Barrera, M.E. (1992). Group treatment program for abusive husbands: long-term evaluation, *American Journal of Orthopsychiatry* **62**(2), 276–283.
- [5] Piantadosi, S. (1997). *Clinical Trials: A Methodological Perspective*, John Wiley & Sons, New York.
- [6] Verbeke, G. & Molenberghs, G. *Linear Mixed Models in Practice: A SAS-oriented Approach*, Lecture Notes in Statistics 126, Springer-Verlag, New York, 1997.
- [7] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, Springer-Verlag, New York.

EMMANUEL LESAFFRE AND GEERT VERBEKE

Cluster Analysis: Overview

KENNETH G. MANTON, GENE LOWRIMORE, ANATOLI YASHIN AND MIKHAIL KOVTUN

Volume 1, pp. 305–315

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cluster Analysis: Overview

Research in various fields gives rise to large data sets of high-dimension where the structure of the underlying processes is not well-enough understood to construct informative models to estimate parameters. *Cluster analysis* is an exploratory method designed to elicit information from such data sets. It is exploratory in that the analyst is more interested in generating hypotheses than in testing them. Researchers generally use cluster analysis as a heuristic tool to gain insight into a set of data. Cluster analysis seeks to reduce the dimensionality of data by grouping objects into a small number of groups, or clusters, whose members are more similar to each other than they are to objects in other clusters. Because it is heuristic, there is wide variation in cluster analysis techniques and algorithms (*see Hierarchical Clustering; k-means Analysis*). Cluster analysis is widely used because it provides useful information in different areas, ranging from identifying genes that have similar function [3] to identifying market segments that have similar consumers.

Many researchers are unwilling to make the assumption of ‘crisp’ group membership, that is, that a person is in only one group. This has led to the extension of cluster analysis to admit the concept of partial group membership and to ‘fuzzy’ clusters ([12], *see Fuzzy Cluster Analysis*).

Cluster analysis works with results of multiple measurements made on a sample of individuals. In this case, initial data may be represented in the form of matrix

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \cdots & x_p^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_p^n \end{pmatrix}, \quad (1)$$

where x_j^i is a result of j th measurement on i th individual. In the terminology of cluster analysis, rows of the matrix $\mathbf{X}$ are referred as *cases*, and columns as *variables*.

Given the data array, (1) (*see Multivariate Analysis: Overview*), the objective is to group objects into clusters of similar objects based on attributes that objects possess. The ‘objects’ are the cases (or rows) of (1), each case having values for each of p variables, $\mathbf{x}^i = (x_1^i, \dots, x_p^i)$. For example, if rows

in (1) represent genes, and attributes are gene expression values (phenotypes) measured under experimental conditions, one might be interested to see which genes exhibit similar phenotypes and functions.

One might find it appropriate to consider variables (columns) as objects of interest, each variable having an attribute value on each of n cases $\mathbf{x}_j = (x_j^1, \dots, x_j^n)$. One important application for which clustering variables is appropriate is in questionnaire design to identify sets of questions that measure the same phenomenon. In other applications, it may make sense to simultaneously cluster cases and variables (*see Two-mode Clustering*).

When cases or variables may be considered as realizations of random vectors, and the probability density function of this random vector is multimodal (with relatively small number of modes), natural clustering will create clusters around the modes. Multimodal probability density functions often arise when the observed distribution is a **finite mixture** [4]. For categorical data, equivalents of finite mixture models are latent class models [18, 17, 11]. Again, natural clustering in this case will produce clusters corresponding to latent classes. In this sense, cluster analysis relates to latent class models in the case of categorical data, and to finite mixture models in the case of continuous data (*see Model Based Cluster Analysis*).

This article identifies common cluster analysis variants and discusses their uses. We will illustrate cluster analysis variants using two data sets, (1) individual disability measures from the National Long Term Care Survey (NLTCs) and (2) multiple software metrics on a number of modules developed as part of a worldwide customer service system for a lodging business.

Distance Measures

A notion of ‘similarity’ is the cornerstone of cluster analysis. Quantitatively, similarity is expressed as a *distance* between cases (variables): the less the distance, the more similar are the cases (variables). Cluster algorithms are based on two distances (a) $d(\mathbf{x}_j, \mathbf{x}_k)$, the distance, measured in a suitable metric, between the two vectors, $\mathbf{x}_j$ and $\mathbf{x}_k$ (in examples of distance below, we define distance between variables; distance between cases is defined in exactly the same way), and (b) the distance between two clusters, $D(c_1, c_2)$.

2 Cluster Analysis: Overview

There are many ways to define $d(\mathbf{x}_j, \mathbf{x}_k)$ (see **Proximity Measures**). New ones can be readily constructed. The suitability of any given measure is largely in the hands of the researcher, and is determined by the data, objectives of the analysis, and assumptions. Although, in many cases, distance is a metric on the space of cases or variables (i.e., it satisfies (a) $d(\mathbf{x}_j, \mathbf{x}_k) \geq 0$; $d(\mathbf{x}_j, \mathbf{x}_k) = 0$ if, and only if, $\mathbf{x}_j = \mathbf{x}_k$; (b) $d(\mathbf{x}_j, \mathbf{x}_k) = d(\mathbf{x}_k, \mathbf{x}_j)$; (c) $d(\mathbf{x}_j, \mathbf{x}_l) \leq d(\mathbf{x}_k, \mathbf{x}_j) + d(\mathbf{x}_k, \mathbf{x}_l)$), the last condition, the *triangle inequality*, is not necessary for the purpose of cluster analysis, and, often, is not used. In the examples below, (2) and (4) are metrics, while (3) is not; (5) is a metric if, and only if, $p = r$.

For measured variables, the standard *Euclidean distance* is often used where the distance between two objects, x_j and x_k , is,

$$d(\mathbf{x}_j, \mathbf{x}_k) = \left(\sum_{i=1}^n (x_k^i - x_j^i)^2 \right)^{1/2}. \quad (2)$$

It is sometimes appropriate to emphasize large distances; in this case, the square of the Euclidean distance could be appropriate

$$d(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^n (x_k^i - x_j^i)^2. \quad (3)$$

If one suspects the data contains outliers, and desires to reduce their effect, then the distance could be defined as the so-called *Manhattan distance* (named to reflect its similarity with the path a taxi travels between two points in an urban area)

$$d(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^n |x_k^i - x_j^i|. \quad (4)$$

The distances defined above are special cases of the general formula

$$d(\mathbf{x}_j, \mathbf{x}_k) = \left(\sum_{i=1}^n |x_j^i - x_k^i|^p \right)^{1/p}, \quad (5)$$

where p and r are parameters used to control the behavior of the distance measure in the presence of outliers or small values.

For categorical attributes (i.e., for attributes that take values from an unordered finite state), the

Hamming distance is used:

$$d(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^n (1 - \delta(x_k^i, x_j^i)),$$

$$\text{where } \delta(x_k^i, x_j^i) = \begin{cases} 1, & \text{if } x_k^i = x_j^i \\ 0, & \text{if } x_k^i \neq x_j^i \end{cases}. \quad (6)$$

The Hamming distance is the number of places in which $\mathbf{x}_j$ and $\mathbf{x}_k$ differ.

It is important that attributes be measured in the same units: otherwise, the unit of the distance measure is not defined. If attributes are measured principally in different units (say, length and weight), one possibility to overcome this obstacle is to replace dimensional values by dimensionless ratios to mean values, that is, replace x_j^i by $x_j^i / \sum_{k=1}^n x_j^k$.

If a statistical package does not support a desired distance measure, it will usually accept an externally computed distance matrix in lieu of raw data.

Just as there is a plethora of inter-object distance measures, there are also many ways to define intercluster distances. **Hierarchical clustering** algorithms are defined by the measure of intercluster distance used.

A *single-linkage* distance between two clusters is a distance between nearest members of these clusters,

$$D_s(c_1, c_2) = \min_{\mathbf{x}_j \in c_1, \mathbf{x}_k \in c_2} d(\mathbf{x}_j, \mathbf{x}_k). \quad (7)$$

A *complete linkage* distance between two clusters is a distance between most distant members of these clusters,

$$D_c(c_1, c_2) = \max_{\mathbf{x}_j \in c_1, \mathbf{x}_k \in c_2} d(\mathbf{x}_j, \mathbf{x}_k). \quad (8)$$

An *average linkage* or UPGA (unweighted pair group average) distance is just an average distance between pairs of objects taken one from each of two clusters,

$$D_a(c_1, c_2) = \frac{1}{M} \sum_{\mathbf{x}_j \in c_1, \mathbf{x}_k \in c_2} d(\mathbf{x}_j, \mathbf{x}_k), \quad (9)$$

where M is a total number of all possible pairs (number of summands in the sum). In the case when objects may be considered as a vector in Euclidean space (i.e., when all measurements x_j^i are real numbers and the Euclidean distance (2) is used), the average linkage distance between two

clusters is the Euclidean distance between their centers of gravity.

All the above distances between clusters possess one property, crucial for efficiency of hierarchical clustering algorithms: if a cluster c_2 is a union of clusters c'_2 and c''_2 , then the distance from any cluster c_1 to the cluster c_2 is readily expressed via distances $D(c_1, c'_2)$ and $D(c_1, c''_2)$. Namely,

$$\begin{aligned} D_s(c_1, c_2) &= \min(D_s(c_1, c'_2), D_s(c_1, c''_2)) \\ D_c(c_1, c_2) &= \max(D_c(c_1, c'_2), D_c(c_1, c''_2)) \\ D_a(c_1, c_2) &= \frac{1}{M' + M''} (M' \times D_a(c_1, c'_2) \\ &\quad + M'' \times D_a(c_1, c''_2)), \end{aligned} \quad (10)$$

Again, if a statistical package does not support a desired distance measure, it will usually accept an external procedure to compute $D(c_1, c_2)$ from $D(c_1, c'_2)$ and $D(c_1, c''_2)$.

The choice of distance measure has a significant impact on the result of a clustering procedure. This choice is usually dictated by the subject domain, and all reasonable possibilities have to be carefully investigated. One important case, which leads to a unique (and, in some sense, ideal) clusterization, is the *ultrametricity* of the distance (see **Ultrametric Inequality**). The distance $d(\mathbf{x}, \mathbf{y})$ is called *ultrametric*, if it satisfies the requirement $d(\mathbf{x}, \mathbf{z}) \leq \max(d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z}))$. This requirement is stronger than the triangle inequality $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$, and implies a number of good properties. Namely, the clusters constructed by the hierarchical clustering algorithm (described below) have the properties: (a) the distance between two members of two clusters does not depend on the choice of these members, that is, if $\mathbf{x}$ and $\mathbf{x}'$ are vectors corresponding to members of a cluster c_1 , and $\mathbf{y}$ and $\mathbf{y}'$ correspond to members of a cluster c_2 , then $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}', \mathbf{y}')$; moreover, all the distances between clusters defined above coincide, and are equal to the distance between any pair of their members, $D_s(c_1, c_2) = D_c(c_1, c_2) = D_a(c_1, c_2) = d(\mathbf{x}, \mathbf{y})$; (b) the distance between any two members of one cluster is smaller than the distance between any member of this cluster and any member of another cluster, $d(\mathbf{x}, \mathbf{x}') \leq d(\mathbf{x}, \mathbf{y})$.

Hierarchical Clustering Algorithm

A main objective of cluster analysis is to define a cluster so that it is, in some sense, as 'far' from other clusters as possible. A hierarchical clustering algorithm starts by assigning each object to its own cluster. Then, at each step, the pair of closest clusters is combined into one cluster. This idea may be implemented very efficiently by first writing distances between objects as a matrix,

$$\begin{pmatrix} d(x_1, x_1) & d(x_1, x_2) & \cdots & d(x_1, x_p) \\ d(x_2, x_1) & d(x_2, x_2) & \cdots & d(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_p, x_1) & d(x_p, x_2) & \cdots & d(x_p, x_p) \end{pmatrix}. \quad (11)$$

This is a symmetric matrix with the diagonal elements equal to 0; thus, it can be stored as an upper triangle matrix. A step consists of the following:

1. Find the smallest element of this matrix; let it be in row j and column j' . As the matrix is upper triangular, we have always $j < j'$. The clusters j and j' are to be combined on this step. Combining may be considered as removal of the cluster j' and replacing cluster j by the union of clusters j and j' .
2. Remove row j' and column j' from the matrix and recalculate values in row j and column j'' using (10).

Here, one can see the importance of the property of distance between clusters given by (10): the distances in the matrix at the beginning of a step are sufficient to calculate new distances; one does not need to know distances between all members of the clusters.

The results of a hierarchical cluster analysis are presented as a *dendrogram*. One axis of a dendrogram is the intercluster distance; the identity of objects is displayed on the other.

Horizontal lines in Figure 1 connect a cluster with its parent cluster; the lengths of the lines indicate distances between the clusters. Where does one slice the dendrogram? It can be at a prespecified measure of dissimilarity, or at a point that yields a certain number of clusters. Wallace [24] suggests stopping at a point on the dendrogram where 'limbs are long and there are not many branches' (see **Number of Clusters**).

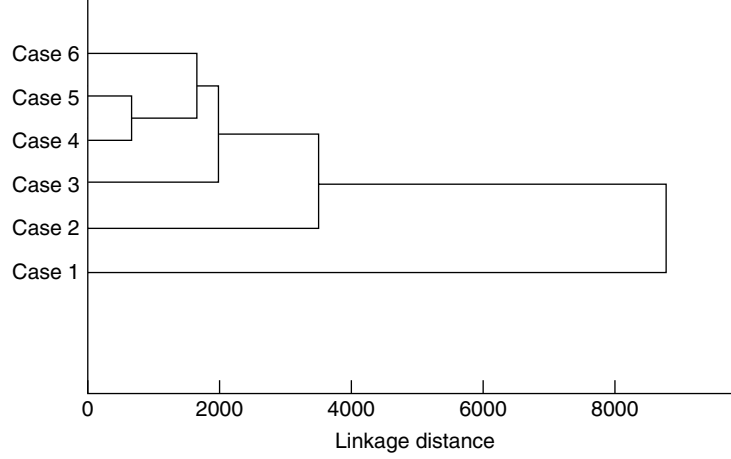


Figure 1 Dendrogram

Depending on the choice of the distance measure between clusters, the resulting clusterization possesses different properties. The single-linkage clustering tends to produce an elongated chain of clusters. Because of ‘chaining’, single-linkage clustering has fallen out of favor [16], though it has attractive theoretical properties. In one dimension, this distance metric seems the obvious choice. It is also related to the **minimum spanning tree (MST)** [7]. The MST is the graph of minimum length connecting all data points. Single-linkage clusters can be arrived at by successively deleting links in the MST [10]. Single-linkage is consistent in one dimension [10]. Complete linkage and average linkage algorithms work best when the data has a strong clustering tendency.

***k*-means Clustering**

In hierarchical clustering, the number of clusters is not known *a priori*. In *k*-means clustering, suitable only for quantitative data, the number of clusters, *k*, is assumed known (see ***k*-means Analysis**).

Every cluster c_k is characterized by its *center*, which is the point $\mathbf{v}_k$ that minimizes the sum

$$\sum_{\mathbf{x}_j \in c_k} d(\mathbf{x}_j, \mathbf{v}_k). \quad (12)$$

Again, it is possible to use different notions of distance; however, *k*-means clustering is significantly more sensitive to the choice of distance, as the

minimization of (12) may be a very difficult problem. Usually, the squared Euclidean distance is used; in this case, the center of a cluster is just its center of gravity, which is easy to calculate.

Objects are allocated to clusters so that the sum of distances from objects to the centers of clusters to which they belong, taken over all clusters, is minimal. Mathematically, this means minimization of

$$\sum_{k=1}^K \sum_{\mathbf{x}_j \in c_k} d(\mathbf{x}_j, \mathbf{v}_k), \quad (13)$$

which depends on continuous vector parameters $\mathbf{v}_k$ and discrete parameters representing membership in clusters. This is a nonlinear constrained optimization problem, and has no obvious analytical solution. Therefore, heuristic methods are adopted; the one most widely used is described below.

First, the centers of clusters $\mathbf{v}_k$ are randomly chosen (or, equivalently, objects are randomly assigned to clusters and the centers of clusters are calculated). Second, each object $\mathbf{x}_j$ is assigned to a cluster c_k whose center is the nearest to the object. Third, centers of clusters are recalculated (based on the new membership) and the second step repeated. The algorithm terminates when the next iteration does not change membership.

Unfortunately, the result to which the above algorithm converges depends on the first random choice of clusters. To obtain a better result, it is recommended to perform several runs of the algorithm and then select the best result.

The advantage of the k -means clustering algorithm is its low computational complexity: on every step, only $n \times k$ (where n is the number of objects) distances have to be calculated, while hierarchical clustering requires computation of (approximately) $n^2/2$ distances. Thus, k -means clustering works significantly faster than hierarchical clustering when k is much smaller than n , which is the case in many practical situations. However, it is hard to determine the right number of clusters in advance. Therefore, some researchers recommend performing a hierarchical clustering algorithm on a subsample to estimate k , and then performing a k -means algorithm on the complete data set.

Wong [25] proposed a hybrid clustering method that does the reverse, that is, a k -means algorithm is used to compute an empirical distribution on n observations from which k clusters are defined. Then, a single-linkage hierarchical algorithm is used to cluster the clusters found by k -means algorithm. The second clustering starts with a $k \times k$ distance matrix *between clusters*. The intercluster distance between neighboring clusters is defined by the distance between closest neighboring objects. The distance between non-neighboring clusters is defined to be infinite. The method is ‘set-consistent’ for one dimension [25]. Consistency for higher dimensions remains a conjecture. Using the k -means algorithm for a first step makes for a computationally efficient algorithm suitable for large data sets. Wong [26] developed a single-linkage procedure that does not use the k -means first step and is set-consistent in high dimensions, but it is suitable only for small samples.

Disabilities Example

The first example uses the NLTCs, a nationally representative longitudinal survey of persons aged 65 and older, and conducted by the Center for Demographic Studies at Duke University. The survey is a probability sample from US Medicare Eligibility Files. Survey waves are conducted every five years, and, for each wave, a nationally representative sample of persons who have become 65 since the prior wave is added to the total survey sample. Disability is assessed by the presence of ADLs or IADLs. An Activity of Daily Living (ADL) is an essential

Table 1 Disability measures

1	Help Eating
2	Help getting in and out of bed
3	Help getting around inside
4	Help dressing
5	Help bathing
6	Help using toilet
7	Is bedfast
8	No inside activity
9	Is wheelchairfast
10	Help with heavy work
11	Needs help with light work
12	Help with laundry
13	Help with cooking
14	Help with grocery shopping
15	Help getting about outside
16	Help traveling
17	Help managing money
18	Help taking medicine
19	Help telephoning
20	Climbing one flight of stairs
21	Bend for socks
22	Hold 10 lb. package
23	Reach over head
24	Difficulty combing hair
25	Washing hair
26	Grasping small objects
27	Reading newspaper

activity of daily living such as eating or toileting for which one requires help. An Instrumental Activity of Daily Living (IADL) is an activity such as grocery shopping or traveling for which one requires help. At each wave, the survey screens individuals with a short survey instrument, and those that are determined to be disabled (one or more ADLs or IADLs lasting for than 90 days), residing either in the community or in an institution, are given a more detailed set of questions. The NLTCs data to be used here is for one survey wave and consists of 5089 observations with 19 questions about disabilities having a binary response, and eight questions about the degree of difficulty in performing certain activities having four responses: no difficulty, some difficulty, much difficulty, or cannot perform the activity at all. The questions (variables) are listed in Table 1. We will use this data to illustrate three algorithms (single-linkage, complete linkage, and UPGA) applied to the first 50 cases in the data set. We restricted the example because more than 50 cases resulted in dendograms too difficult to read. We will also use this data to

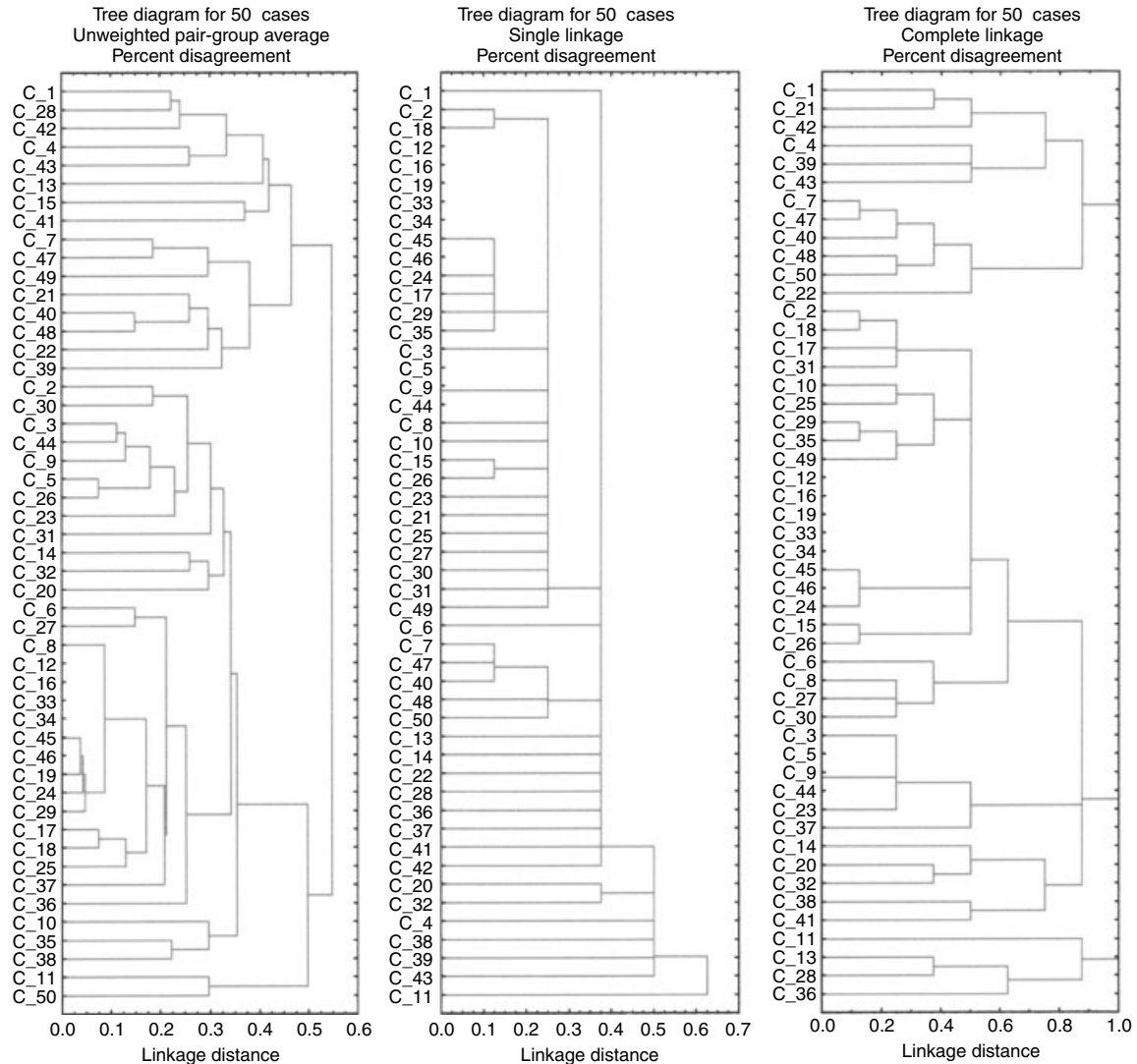


Figure 2 Fifty cases From NLTCs Data set clustered by three linkage algorithms

do a cluster analysis of variables. Since all questions have categorical responses, we use the percent difference measure of dissimilarity as the distance function (6).

Figure 2 presents the results from the dendrograms from the three hierarchical algorithms run on the first 50 NLTCs cases. Clusters for the single-linkage algorithm tend to form a chain-like structure. The two other algorithms give more compact results. The dendrograms have the 50 cases identified on the vertical axis. The horizontal axis scale is the

distance measure. To read a dendrogram, select a distance on the horizontal axis and draw a vertical line at that point. Every line in the dendrogram that is cut defines a cluster *at that distance*. The length of the horizontal lines represents distances between clusters. While the three dendrograms appear to have the same distance metric, distances are calculated differently.

Since the distance measures are defined differently, we cannot cut each dendrogram at a fixed distance, say 0.5, and expect to get comparable results.

Instead, we start at the top of the tree where every object is a member of a single cluster and cut immediate level two. In the UPGA dendrogram, that cut will be at approximately 0.42; in the single-linkage dendrogram, that cut will be at 0.45, and in the complete-linkage dendrogram, at about 0.8. These cuts define 4, 7, and 7 clusters, respectively. The UPGA cluster sizes were {2, 32, 8, 8}; the single-linkage cluster sizes were {1, 1, 1, 1, 2, 44}, and the complete-linkage cluster sizes were {3, 1, 5, 6, 33, 6, 7}. In practice, one would examine the cases that make up each cluster.

These algorithms can be used to cluster variables by reversing the roles of variables and cases in distance calculations. Variables in the data set were split into two sets according to outcome sets. The complete-linkage clustering was used.

The results of the first 19 variables, which had binary response sets, are displayed in Figure 3. At a distance of 0.3, the variables divide into two clusters. Moving about inside or outside, bathing, grocery shopping, traveling, and heavy work make up one cluster. The remaining 13 variables make up the second cluster. The second cluster could be seen as being two subclusters: (1) in/out-bed, dressing, toileting, light work, cooking, and laundry; and (2) activities that indicate a greater degree of disability (help eating, being bedfast, wheelchair-fast, etc.).

The remaining eight variables that each have four responses were also analyzed. The results are

in Figure 4. At a distance of 0.5, the variables fit into three sets. Climbing a single flight of stairs (one-flight-stairs) comprises its own cluster; socks (bending for) and holding a 10 pound package comprise another; and combing hair, washing hair, grasping an object and reading a newspaper constitute a third.

Software Metrics Example

The second data set consists of 26 software metrics measured on the source code of 180 individual software modules that were part of an international customer service system for a large lodging business. The modules were all designed and developed according to Michael Jackson methods [13, 14]. The measurements included Halstead measures [8] of vocabulary, volume, length, McCabe complexity [5], as well as other measures such as comments, number of processes. The original analysis determined if there were modules that should be rewritten. This second set will be analyzed by *k*-means clustering. The original analysis was done using **Principal Component Analysis** [13, 14], and that analysis is presented here for comparison.

In the original Principal Component Analysis, since the variables were on such different scales, and to keep one variable from dominating the analysis, variables were scaled to have equal variances. The resulting covariance matrix (26 variables) was

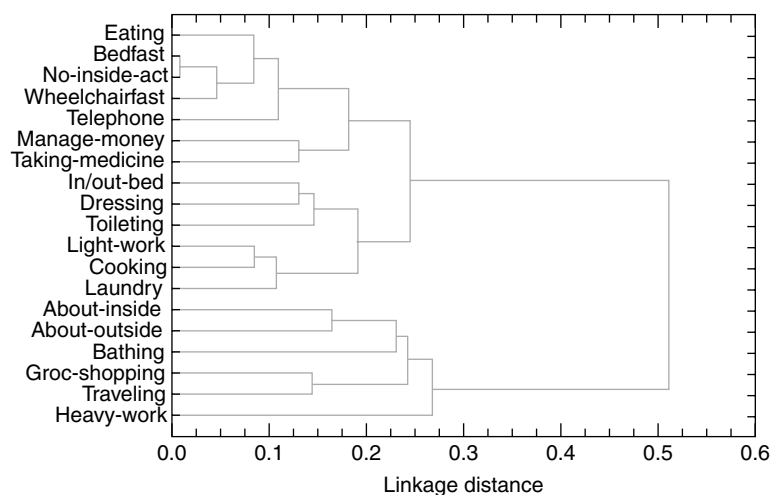
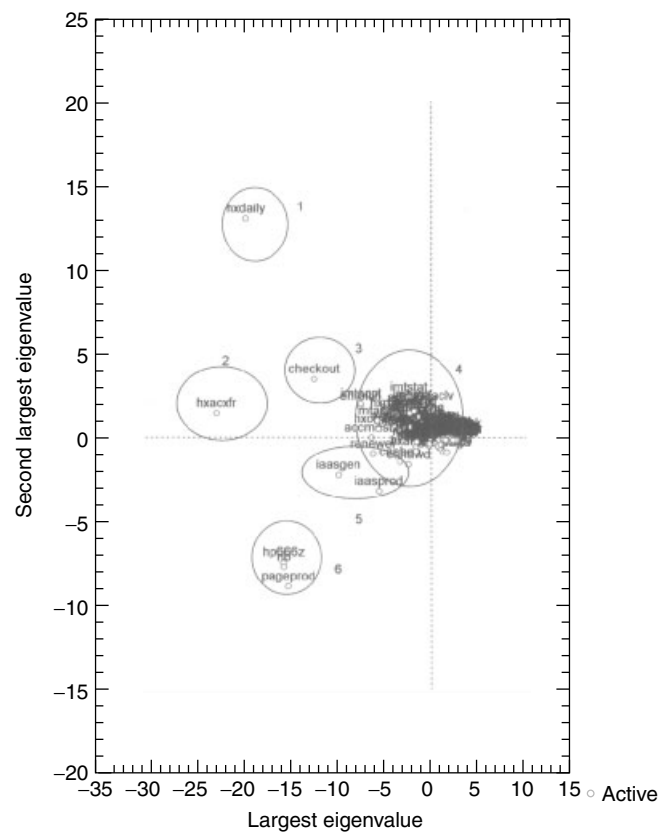
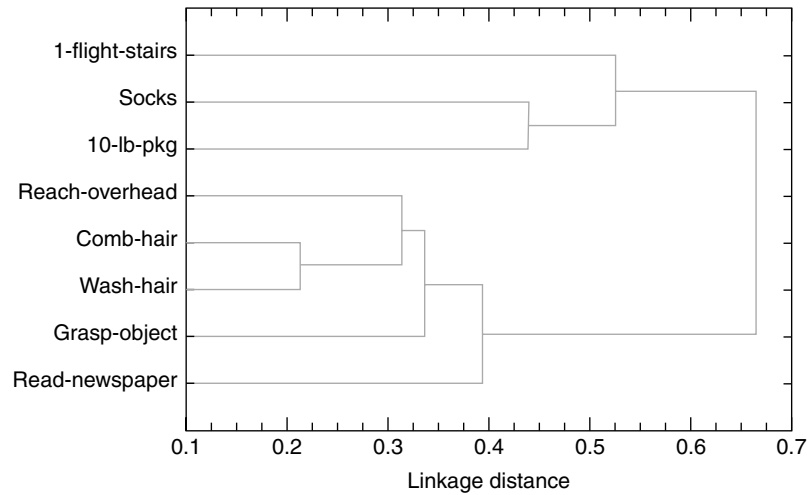


Figure 3 Tree diagram for 19 disability measures complete-linkage dissimilarity measure = percent disagreement



singular, so analyses used the generalized inverse. The two largest eigenvalues accounted for a little over 76% of the variance. The data was transformed by the eigenvectors associated with the two largest eigenvalues and plotted in Figure 5. The circles indicate the five clusters identified by the eye.

The analysis identified clusters, but does not tell us which programs are bad. Without other information, the analysis is silent. One could argue that we know we have a good process and, therefore, any deviant results must be bad. In this case, three of the programs in the smaller clusters had repeatedly missed deadlines and, therefore, could be considered bad. Other clusters were similar. Two clusters differed from the main body of programs on one eigenvector and not on the other. The other two small clusters differed from the main body on both eigenvectors.

We analyzed the same data by k -means clustering with $k = 4$. The four clusters contained 5, 9, 29, and 145 objects. Cluster 1 consisted of the five objects in the three most remote clusters in Figure 5. Cluster 3 was made up of the programs in Figure 5, whose names are far enough away from the main body to have visible names. Clusters 2 and 4 were the largest and break up the main body of Figure 4.

Conclusions

The concept of cluster analysis first appeared in the literature in the 1950s and was popularized by [22, 21, 1, 9], and others. It has recently enjoyed increased interest as a data-mining tool for understanding large volumes of data such as gene expression experiments and transaction data such as online book sales or occupancy records in the lodging industry.

Cluster analysis methods are often simple to use in practice. Procedures are available in commonly used statistical packages (e.g., SAS, STATA, STATISTICA, and SPSS) as well as in programs devoted exclusively to cluster analysis (e.g., CLUSTAN). Algorithms tend to be computationally intensive. The analyses presented here were done using STATISTICA 6.1 software (see **Software for Statistical Analyses**).

Cluster analyses provide insights into a wide variety of applications without many statistical assumptions. Further examples of cluster analysis applications are (a) identifying of market segments

from transaction data; (b) estimating of readiness of accession countries to join the European Monetary Union [2]; (c) understanding benign prostate hyper trophy (BHP) [6]; (d) studying antisocial behavior [15]; (e) to study southern senatorial voting [19]; (f) studying multiple planetary flow regimes [20]; and (g) reconstructing of fossil organisms [23]. The insights they provide, because of a lack of a formal statistical theory of inference, require validating in other venues where formal hypothesis tests can be used.

References

- [1] Andenberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press, New York.
- [2] Boreiko, D. & Oesterreichische Nationalbank. (2002). *EMU and Accession Countries: Fuzzy Cluster Analysis of Membership*, Oesterreichische Nationalbank, Wien.
- [3] Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1999). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences* **95**(25), 14863–14868.
- [4] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman & Hall.
- [5] Gilb, T. (1977). *Software Metrics*, Wintrop Publishers, Cambridge.
- [6] Girman, C.J. (1994). Cluster analysis and classification tree methodology as an aid to improve understanding of benign prostatic hyperplasia, Dissertation, Institute of Statistics, the University of North Carolina, Chapel Hill.
- [7] Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single-linkage cluster analysis *Applied Statistics* **18**(18), 54–65.
- [8] Halstead, M.H. (1977). *Elements of Software Science*, Elsevier Science, New York.
- [9] Hartigan, J. (1975). *Clustering Algorithms*, Wiley, New York.
- [10] Hartigan, J.A. (1981). Consistency of single linkage for high-density clusters, *Journal of the American Statistical Association* **76**(374), 388–394.
- [11] Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models*, SAGE, Thousand Oaks.
- [12] Höppner, F. (1999). *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition*, John Wiley, Chichester; New York.
- [13] Jackson, M.A. (1975). *Principles of Program Design*, Academic Press, New York.
- [14] Jackson, M.A. (1983). *System Development*, Addison-Wesley, New York.
- [15] Jordan, B.K. (1986). A fuzzy cluster analysis of antisocial behaviour: implications for deviance theory, Dissertation, Duke University, Durham.
- [16] Lance, G.N. & William, W.T. (1967). A general theory of classificatory sorting strategies: 1. hierarchical systems, *Computer Journal* **9**, 373–380.

- [17] Langeheine, R. & Rost, J. (1988). *Latent Trait and Latent Class Models*, Plenum Press, New York.
- [18] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton Mifflin, Boston.
- [19] Kammer, W.N. (1965). A Cluster-blot analysis of southern senatorial voting behavior, 1947–1963, Dissertation, Duke University, Durham.
- [20] Mo, K. & Ghil, M. (1987). Cluster analysis of multiple planetary flow regimes, NASA contractor report, National Aeronautics and Space Administration.
- [21] Sneath, P.H.A. & Sokal, R.R. (1973). *Numerical Taxonomy*, Freeman, San Francisco.
- [22] Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*, W.H. Freeman, San Francisco.
- [23] Von Bitter, P.H., Merrill G.K. (1990). *The Reconstruction of Fossil Organisms Using Cluster Analysis: A Case Study from Late Palaeozoic Conodonts*, Toronto, Royal Ontario Museum.
- [24] Wallace, C. (1978). Notes on the distribution of sex and shell characters in some Australian populations of *Potamopyrgus* (Gastropoda: Hydrobiidae), *Journal of the Malacological Society of Australia* **4**, 71–76.
- [25] Wong, A.M. (1982). A hybrid clustering method for identifying high density clusters, *Journal of the American Statistical Association* **77**(380), 841–847.
- [26] Wong, M.A. & Lanr, T. (1983). A kth nearest neighbor clustering procedure, *Journal of the Royal Statistical Society. Series B (Methodological)* **45**(3), 362–368.

Further Reading

- Avetisov, V.A., Bikulov, A.H., Kozyrev, S.V. & Osipov, V.A. (2002). p -adic models of ultrametric diffusion constrained by hierarchical energy landscapes, *Journal of Physics A-Mathematical and General* **35**, 177–189.
- Ling, R.F. (1973). A probability theory of cluster analysis, *Journal of the American Statistical Association* **68**(341), 159–164.
- Manton, K., Woodbury, M. & Tolley, H. (1994). *Statistical Applications Using Fuzzy Sets*, Wiley Interscience, New York.
- Woodbury, M.A. & Clive, J. (1974). Clinical pure types as a fuzzy partition, *Journal of Cybernetics* **4**, 111–121.

(See also **Overlapping Clusters**)

KENNETH G. MANTON, GENE LOWRIMORE,
ANATOLI YASHIN AND MIKHAIL KOVTUN

Clustered Data

GARRETT M. FITZMAURICE

Volume 1, pp. 315–315

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Clustered Data

There are many studies in the behavioral sciences that give rise to data that are clustered. Clustered data commonly arise when intact groups are assigned to interventions or when naturally occurring groups in the population are sampled. An example of the former is cluster-randomized trials. In a cluster-randomized trial, groups of individuals, rather than the individuals themselves, are randomized to different interventions and data on the outcomes of interest are obtained on all individuals within a group. Alternatively, clustered data can arise from random sampling of naturally occurring groups in the population. Families, households, neighborhoods, and schools are all instances of naturally occurring clusters that might be the primary sampling units in a study. Clustered data also arise when data on the outcome of interest are simultaneously obtained either from multiple raters or from different measurement instruments. Finally, longitudinal and repeated measures studies give rise to clustered data (*see* **Longitudinal Data Analysis; Repeated Measures Analysis of Variance**). In these studies, the cluster is composed of the repeated measurements obtained from a single individual at different occasions or under different conditions.

Although we have distinguished between clustering that occurs naturally and clustering due to study design, the consequence of clustering is the same: units that are grouped in a cluster are likely to respond more similarly. Intuitively, we might reasonably expect that measurements on units within a cluster are more alike than the measurements on units in different clusters. For example, two children

selected at random from the same family are expected to respond more similarly than two children randomly selected from different families. In general, the degree of clustering can be represented by positive correlation among the measurements on units within the same cluster (*see* **Intraclass Correlation**). This within-cluster correlation can be thought of as a measure of the homogeneity of responses within a cluster relative to the variability of such responses between clusters. The intraclass correlation invalidates the crucial assumption of independence that is the cornerstone of so many standard statistical techniques. As a result, the straightforward application of standard regression models (e.g., **multiple linear regression** for a continuous response or **logistic regression** for a binary response) to clustered data is no longer valid unless some allowance for the clustering is made.

From a statistical standpoint, the main feature of clustered data that needs to be accounted for in any analysis is the fact that units from the same cluster are likely to respond more similarly. As a consequence, clustered data may not contain quite as much information as investigators might otherwise like to believe it does. In general, neglecting to account for clusters in the data will lead to incorrect estimates of precision. For example, when estimating averages or comparing averages for groups of clusters, failure to account for clustering will result in estimated standard errors that are too small, **confidence intervals** that are too narrow, and *P* values that are too small. In summary, failure to make some adjustment to the nominal standard errors to correct for clusters in the data can result in quite misleading scientific inferences.

GARRETT M. FITZMAURICE

Cochran, William Gemmell

BRIAN S. EVERITT

Volume 1, pp. 315–316

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cochran, William Gemmell

Born: July 15, 1909, in Rutherglen, Scotland.

Died: March 29, 1980, in Orleans, Massachusetts.

William Cochran was born into a family of modest means in a suburb of Glasgow. In 1927, he was awarded a bursary that allowed him to take his first degree at Glasgow University; he obtained an M.A. in mathematics and physics in 1931, and a scholarship to St John's College, Cambridge. At Cambridge, he took the only course in mathematical statistics then available and produced a paper on the distribution of quadratic forms, now known as *Cochran's theorem* [1]. But Cochran did not finish his Ph.D. studies in Cambridge because in 1934, **Frank Yates**, who had just become head at Rothamsted Experimental Station on Fisher's departure to the Galton Chair at University College, offered him a job, which he accepted.

During his six years at Rothamsted, Cochran worked on experimental design and sample survey techniques, publishing 29 papers. He worked closely with Yates and also spent time with **Fisher**, who was a frequent visitor to Rothamsted. A visit to the Iowa Statistical Laboratory in 1938 led to the offer of a

statistics post from **George Snedecor**, and in 1939, Cochran emigrated to the United States.

In the United States, Cochran continued to apply sound experimental techniques in agriculture and biology, and after World War II, he moved to the newly formed Institute of Statistics in North Carolina to work with **Gertrude Cox**. Their collaboration led, in 1950, to the publication of what quickly became the standard textbook on experimental design [2]. In 1957, Cochran was appointed head of the Department of Statistics at Harvard, where he remained until his retirement in 1976.

Cochran was made president of the Biometric Society from 1954 to 1955 and vice president of the American Association for the Advancement of Science in 1966. In 1974, he was elected to the US National Academy of Science. Cochran was an outstanding teacher and among his 40 or so doctoral students are many who have become internationally famous leaders in a variety of areas of statistics.

References

- [1] Cochran, W.G. (1934). The distribution of quadratic forms in a normal system, *Proceedings of the Cambridge Philosophical Society* **30**, 178–189.
- [2] Cochran, W.G. & Cox, G. (1950). *Experimental Design*, Wiley, New York.

BRIAN S. EVERITT

Cochran's C Test

PATRICK MAIR AND ALEXANDER VON EYE

Volume 1, pp. 316–317

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cochran's C Test

Cochran's [1] C test is used to examine model assumptions made when applying **analysis of variance (ANOVA)**. The simple model equation in a one-factorial ANOVA is $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, for $i = 1, \dots, n$ and $j = 1, \dots, p$. In this equation, Y_{ij} is the observation for case i in treatment category j , μ is the grand mean, α_j is the j th treatment, and ε_{ij} is the residual of Y_{ij} . According to the model equation, there are three core assumptions for ANOVA:

1. The model contains all sources of the variation that affect Y_{ij} ;
2. The experiment involves all of the treatment effects of interest; and
3. The residuals are independent of each other, are normally distributed within each treatment population, and have a mean of 0 and constant variance, σ^2 .

The last part of the third assumption is known as the assumption of *homogeneity of residual variances*, or *homoscedasticity*. Written in the form of a null hypothesis, this assumption is

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_p^2. \quad (1)$$

The alternative hypothesis is $\sigma_j^2 \neq \sigma_{j'}^2$, for some j and j' and $j \neq j'$. In a fashion similar to tests of distributional assumptions, the ANOVA model requires that this null hypotheses be retained. If this hypothesis can be retained, standard ANOVA can be performed. Otherwise, that is, in the case of *heteroscedasticity*, nonparametric approaches such as the **Kruskal–Wallis test** may be more suitable.

There exists a number of tests of homogeneity of residual variances. If the design is orthogonal (sample sizes are the same under each treatment), the tests from Hartley, Leslie and Brown, and Levene, or Cochran's test are suitable. If the design is nonorthogonal (samples sizes differ across treatment conditions), Bartlett's test is suitable (for a more detailed treatment see [3, 4]).

Because of its simple mathematical definition, Cochran's C test, together with Hartley's test, is a so-called *easy-to-use* test. Hartley [2] proposes calculating an F-ratio by dividing the maximum of the empirical variances by their minimum (over the

various treatments). In most cases, Hartley's and Cochran's tests lead to the same results. Cochran's test, however, tends to be more powerful (*see Power*), because it uses more of the information in the data. A second reason why Cochran's test may be preferable is that it performs better when one treatment variance is substantially larger than all other treatment variances. Cochran's test statistic is

$$C = \frac{\sigma_{\max}^2}{\sum_{j=1}^p \sigma_j^2}, \quad (2)$$

where $\sigma_{\max}^2$ is the largest of the p sample variances, and $\sum_{j=1}^p \sigma_j^2$ is the sum of all treatment variances. Cochran [1] calculated the theoretical distribution of this ratio. Thus, it is possible to test the null hypothesis of homoscedasticity. Degrees of freedom are equal to p and $n - 1$, and the critical values are tabulated [1].

Consider the following hypothetical data example [5]. Five populations have been studied with the following sample variances: $s_1^2 = 26$; $s_2^2 = 51$; $s_3^2 = 40$; $s_4^2 = 24$; $s_5^2 = 28$. It is obvious that the second variance is substantially larger than the other variances. Therefore, this variance is placed in the numerator of the formula for C . Suppose, each of these variances is based on $df = 9$, and $\alpha = 0.05$. The application of the above formula yields $C = 0.302$. The critical value is 0.4241. The calculated value is smaller. Therefore, we retain the null hypothesis of variance homogeneity. This core assumption of an ANOVA application is met.

Table 1 summarizes Sachs' [5] recommendations for tests of homoscedasticity under various conditions.

Table 1 Testing homoscedasticity under various conditions

Population	Recommended test
Skewed	Cochran
Normal $N(\mu, \sigma)$	For $p < 10$: Hartley, Cochran; for $p \geq 10$: Bartlett
Platykurtic, that is, flatter than $N(\mu, \sigma)$	Levene
Leptokurtic, that is, taller than $N(\mu, \sigma)$	For $p < 10$: Cochran; for $p \geq 10$: Levene

2 Cochran's C Test

References

- [1] Cochran, W.G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total, *Annals of Eugenics* **11**, 47–52.
- [2] Hartley, H.O. (1940). Testing the homogeneity of a set of variances, *Biometrika* **31**, 249–255.
- [3] Kirk, R.E. (1995). *Experimental Design*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [4] Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996). *Applied Linear Statistical Models*, Irwin, Chicago.
- [5] Sachs, L. (2002). *Angewandte Statistik [Applied Statistics]*, 10th Edition, Springer-Verlag, Berlin.

PATRICK MAIR AND ALEXANDER VON EYE

Coefficient of Variation

PAT LOVIE

Volume 1, pp. 317–318

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Coefficient of Variation

Although the coefficient of variation (CV) is defined for both distributions and for samples, it is in the latter context, as a descriptive measure for data, that it is generally encountered.

The CV of a sample of data is defined as the sample standard deviation (SD) divided by the sample mean (m), that is,

$$CV = \frac{SD}{m}. \quad (1)$$

The value is sometimes expressed as a percentage. Two important characteristics of the CV are that it is independent of both the units of measurement and the magnitude of the data.

Suppose that we measure the times taken by right-handed subjects to complete a tracking task using a joystick held in either the right hand or the left hand. The mean and SD of times for the right-hand joystick (RH) group are 5 sec and 1 sec, and those for the left-hand (LH) group are greater at 12 sec and 1.5 sec. Then, $CV_{RH} = 1/5 = 0.20$, whereas $CV_{LH} = 1.5/12 = 0.125$.

Here, we see that the *relative* variability is greater for the RH group even though the SD is only two-thirds of that for the LH group.

Equally well, if we had counted also the number of errors (e.g., deviations of more than some fixed

amount from the required route), the CV would allow us to compare the relative variability of errors to that of times because it does not depend on the units of measurement.

The notion of the CV is generally attributed to **Karl Pearson** [2]. In an early article from 1896, which attempts to provide a theoretical framework for **Galton's** rather informal ideas on correlation and regression, Pearson used the CV for assessing the relative variability of data on variables ranging from stature of family members to skull dimensions of ancient and modern races. Pearson pointed specifically to the fact that differences in relative variability indicate 'inequality of mutual correlations' (or, similarly, of mutual regressions).

A more recent application of the CV, especially relevant to behavioral researchers working in experimental fields, is as a means of assessing within subject variability [1].

References

- [1] Bland, J.M. & Altman, D.G. (1996). Statistics notes: measurement error proportional to the mean, *British Medical Journal* **313**, 106.
- [2] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society of London, A* **187**, 253–318.

PAT LOVIE

Cohen, Jacob

PATRICIA COHEN

Volume 1, pp. 318–319

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cohen, Jacob

Jacob Cohen's contributions to statistical analysis in the behavioral sciences reflected his lifelong perspective as a data-analyst, insisting that the role of statistics was to help scientists answer the questions motivating any given study, and that there was no substitute for human judgment in this process [3]. Cohen's best-known contributions began when he called attention to the poor statistical **power** of much psychological research [1] and developed readily applied methods of estimating such power in planned research [2, 4]. Such estimates were necessarily based on the probability of rejecting the null hypothesis with an acceptably low level of statistical significance (*alpha level*). Ironically, among his last contributions to methodological developments in the field was a rejection of a statistical criterion as the central basis for interpreting study findings.

Cohen entered City College of New York (CCNY) at the age of 15, following graduation from Townsend Harris High School in New York. After two years of dismal performance (except in ping pong), he worked in war-related occupations and then enlisted in the Army Intelligence Service in time to participate in the final year of World War II in France. Upon returning to the US, he completed his undergraduate education at CCNY (1947) and his doctoral degree in clinical psychology at New York University (1950), writing a dissertation based on **factor analyses** of the Wechsler IQ test in samples of patients and comparison groups. In the beginning, during his studies, Cohen carried out research in the Veteran's Administration and continued thereafter as staff psychologist and director of research while also teaching on a part-time basis. During those years, he developed Cohen's *kappa* statistic (see **Rater Agreement – Kappa**) [1], a measure of chance-corrected agreement later further elaborated to take partial agreement into account. In 1959, he was appointed to full-time faculty status in the psychology department at New York University, where he remained as head of quantitative psychology until his retirement in 1993. Throughout these years and until his death, Cohen consulted on research design and data analysis at nearly every behavioral research and university site in the city, especially at New York

State Psychiatric Institute. He was president of the *Society for Multivariate Experimental Psychology* in 1969 and honored with a lifetime achievement award by Division 7 of the *American Psychological Association* in 1997.

Cohen's first widely influential work followed an analysis of the *statistical power* of studies published in the 1960 *Journal of Abnormal and Social Psychology* [2]. His development of rough norms for small, medium, and large **effect sizes** and easily used methods for estimating statistical power for a planned study made his book *Statistical Power Analysis for the Behavioral Sciences* [4] the classic in its field, with widely used subsequent editions and eventually computer programs. His second major contribution to the field was the adoption of *multiple regression analysis as a general data analytic system* [3, 7]. In these books and articles, Cohen employed the accessible language and conversational style that also made his work, particularly his 1990 'Things I have learned (so far)' and 1994 'The earth is round ($p < .05$)', so widely appreciated [5, 6]. His name also appeared in the top rung of citation indices in the behavioral sciences over an extended period.

References

- [1] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.
- [2] Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review, *Journal of Abnormal and Social Psychology* **65**(3), 145–153.
- [3] Cohen, J. (1968). Multiple regression as a general data-analytic system, *Psychological Bulletin* **70**(6), 426–443.
- [4] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, 1st Edition, Lawrence Erlbaum Associates, Hillsdale (2nd Edition, 1988).
- [5] Cohen, J. (1990). Things I have learned (so far), *American Psychologist* **45**(12), 1304–1312.
- [6] Cohen, J. (1994). The earth is round ($p < .05$), *American Psychologist* **49**(12), 997–1003.
- [7] Cohen, J. & Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 1st Edition, Lawrence Erlbaum Associates, Mahwah (2nd Edition, 1983; 3rd Edition, with West, S.J. & Aiken, L.S., 2003).

PATRICIA COHEN

Cohort Sequential Design

PETER PRINZIE AND PATRICK ONGHENA

Volume 1, pp. 319–322

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cohort Sequential Design

Two broad approaches are very common for the investigation of human development in time: individuals of different ages are studied at only one point in time (i.e., cross-sectional) (*see* **Cross-sectional Design**), or the same subjects are assessed at more than one – and preferable many – points in time (i.e., longitudinal) (*see* **Longitudinal Data Analysis**). The limitations of cross-sectional designs are well known [12, 16]. Differences in different cohorts of children at 3 years of age, 5 years, and 7 years can be due to age effects of the different individuals in each cohort as well as due to the experiences of the different individuals in each cohort. Cross-sectional studies fail to provide insight into the underlying life course mechanisms of development, whether they are societal, institutional, or personal because they confound age and cohort effects.

Longitudinal studies, in contrast, focus on the change within individuals over time and, hence, address directly the limitations of the cross-sectional approach. Longitudinal designs do not confound cohort and age, because the same individuals are assessed at each age. Although problems of true longitudinal designs are frequently reported in the methodological literature, they are often overlooked [8, 10, 22]. The most important substantial disadvantage of longitudinal designs is that age trends may reflect historical (time of measurement) effects during the study rather than true developmental change. In methodological terms, age and period effects are confounded (*see* **Confounding Variable; Quasi-experimental Designs**). This makes it hard to know the degree to which results specific to one cohort can be generalized to other cohorts. Another important practical disadvantage is that the same individuals must be followed repeatedly and tested at many different times. The amount of time it takes to complete a longitudinal study makes it very expensive and increases the chance on attrition (loss of subjects from refusal, tracing difficulties, etc.). In addition, there is a real risk that findings, theories, methods, instrumentation, or policy concerns may be out-of-date by the time data collection and analysis end. Finally, participants can be affected by repeated measurements.

Therefore, researchers have long argued that alternative approaches that maintain the advantages of the

longitudinal design but minimize its disadvantages were needed. The accelerated longitudinal design was first proposed by Bell [3, 4]. He advocated the method of ‘convergence’ as a means for meeting research needs not satisfied by either longitudinal or cross-sectional methods. This method consists of limited repeated measurements of independent age cohorts, resulting in temporally overlapping measurements of the various age groups. This technique, also termed as *cross-sequential design* [20], *the cohort-sequential design* [16], or *mixed longitudinal design* [5], provides a means by which adjacent segments consisting of limited longitudinal data on a specific age cohort can be linked together with small, overlapping segments from other temporally related age cohorts to determine the existence of a single developmental curve. The researcher, therefore, approximates a long-term longitudinal study by conducting several short-term longitudinal studies of different age cohorts simultaneously. In addition, this technique allows the researcher to determine whether those trends observed in the repeated observations are corroborated within short time periods for each age cohort. By assessing individuals of different ages on multiple occasions, cohort-sequential designs permit researchers to consider cohort and period effects, in addition to age changes [7, 13, 21].

Recently, Prinzie, Onghena, and Hellinckx [17] used an accelerated design to investigate developmental trajectories of externalizing problem behavior in children from 4 to 9 years of age. Table 1 represents their research design. Four cohorts (4, 5, 6, and 7 years of age at the initial assessment) have been assessed annually for three years. These age cohorts were ‘approximately staggered’ [14], meaning that the average age of the first cohort at the second measurement period was about the same as the average age of the second cohort at the initial measurement, and so forth. One row represents one longitudinal design and one column represents one cross-sectional design. Table 1 demonstrates that, for one row, age and period are completely confounded, and that, for one column, age and cohort are completely confounded. The more cross sections are added, the more the entanglement of age and cohort is rectified. As illustrated by Table 1, three annual measurements (1999–2001) of four cohorts span an age range of 6 years (from 4 to 9 years of age). In this study, the same developmental model was assumed in each age cohort, allowing for tests of hypotheses

2 Cohort Sequential Design

Table 1 Accelerated longitudinal design with four cohorts and three annual assessments with an age range from 4 to 9 years of age

Cohort	Period 1999 (Years)	2000 (Years)	2001 (Years)
Cohort 1	4	5	6
Cohort 2	5	6	7
Cohort 3	6	7	8
Cohort 4	7	8	9

concerning convergence across separate groups and the feasibility of specifying a common growth trajectory over the 6 years represented by the latent variable cohort-sequential design (see below).

Advantages and Disadvantages

A noticeable advantage of the cohort-sequential over the single-cohort longitudinal design is the possibility to study age effects independent of period and cohorts effects, but only if different cohorts are followed up between the same ages in different periods. Another advantage is the shorter follow-up period. This reduces the problems of cumulative testing effects and attrition, and produces quicker results. Finally, tracking several cohorts, rather than one, allows the researcher to determine whether those trends observed in the repeated observations are corroborated within short time periods for each age cohort. Two basic principles should be considered when designing cohort-sequential studies: efficiency of data collection and sufficiency of overlap. According to Anderson [1, p. 147], a proper balance between these two principles can be achieved by ensuring that (a) at least three data points overlap between adjacent groups, and (b) the youngest age group is followed until they reach the age of the oldest group at the first measurement.

The main disadvantage of the cohort-sequential design in comparison with the single-cohort longitudinal design is that within-individual developmental sequences are tracked over shorter periods. As a result, some researchers have questioned the efficacy of the cohort-sequential approach in adequately recovering information concerning the full longitudinal curve from different cohort segments when the criterion of convergence is not met. The cohort-sequential design is not the most appropriate design

for the investigation of long-term causal effects that occur without intermediate effects or sequences (e.g., between child abuse and adult violence). In addition, questions remain concerning the ability of the cohort-sequential approach to assess the impact of important events and intervening variables on the course of development [19].

Statistical Analysis

Several data-analytical strategies are developed to analyze data from a cohort-sequential design. The most well known are matching cross-cohorts based on statistical tests of significance, the use of **structural equation modeling** and **Linear multilevel models**. Bell [4] linked cohorts by matching characteristics of the subjects using a method he described as '*ad hoc*'. Traditional **analysis of variance** and regression methods (see **Multiple Linear Regression**) have been employed for cohort linkage and are criticized by Nesselroade and Baltes [16]. More recently, two statistical approaches have been proposed to depict change or growth adequately: the *hierarchical linear model* [6, 18] and the latent curve analysis (see **Structural Equation Modeling: Latent Growth Curve Analysis**) [11, 14, 15]. Both approaches have in common that growth profiles are represented by the parameters of initial status and the rate of change (see **Growth Curve Modeling**). The hierarchical linear model is easier for model specification, is more efficient computationally in yielding results and provide a flexible approach that allows for missing data, unequal spacing of time points, and the inclusion of time-varying and between-subject covariates measured either continuously or discretely. The latent curve analysis has the advantage of providing model evaluation, that is, an overall test of goodness of fit, and is more flexible in modeling and hypothesis testing. The separate cohorts' developmental paths are said to converge to a single developmental path if a model that assumes unequal paths produces results that are not statistically distinguishable from results produced by a simpler model that specifies a single path. Chou, Bentler, and Pentz [9] and Wendorf [23] compared both techniques and concluded that both approaches yielded very compatible results. In fact, both approaches might have more in common than once thought [2].

References

- [1] Anderson, E.R. (1995). Accelerating and maximizing information in short-term longitudinal research, in *The Analysis of Change*, J. Gottman, ed., Erlbaum, Hillsdale, pp. 139–164.
- [2] Bauer, D.J. (2003). Estimating multilevel linear models as structural equation models, *Journal of Educational and Behavioral Statistics* **28**, 135–167.
- [3] Bell, R.Q. (1953). Convergence: an accelerated longitudinal approach, *Child Development* **24**, 145–152.
- [4] Bell, R.Q. (1954). An experimental test of the accelerated longitudinal approach, *Child Development* **25**, 281–286.
- [5] Berger, M.P.F. (1986). A comparison of efficiencies of longitudinal, mixed longitudinal, and cross-sectional designs, *Journal of Educational Statistics* **2**, 171–181.
- [6] Bryk, A.S. & Raudenbush, S.W. (1987). Application of hierarchical linear models to assessing change, *Psychological Bulletin* **101**, 147–158.
- [7] Buss, A.R. (1973). An extension of developmental models that separate ontogenetic changes and cohort differences, *Psychological Bulletin* **80**, 466–479.
- [8] Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching, in *Handbook for Research on Teaching*, N.L. Gage, ed., Rand MacNally, Chicago, pp. 171–246.
- [9] Chou, C.P., Bentler, P.M. & Pentz, M.A. (1998). Comparisons of two statistical approaches to study growth curves: the multilevel model and the latent curve analysis, *Structural Equation Modeling* **5**, 247–266.
- [10] Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Rand MacNally, Chicago.
- [11] Duncan, T.E., Duncan, S.C., Strycker, L.A., Li, F. & Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications*, Erlbaum, Mahwah.
- [12] Farrington, D.P. (1991). Longitudinal research strategies: advantages, problems, and prospects, *The Journal of the American Academy of Child and Adolescent Psychiatry* **30**, 369–374.
- [13] Labouvie, E.W. & Nesselroade, J.R. (1985). Age, period, and cohort analysis and the study of individual development and social change, in *Developmental and Social Change: Explanatory Analysis*, J.R. Nesselroade & A. von Eye, eds, Academic Press, New York, pp. 189–212.
- [14] McArdle, J.J. & Anderson, E. (1990). Latent variable growth models for research on aging, in *Handbook of the Psychology of Aging*, J.E. Birren & K.W. Schaie, eds, Academic Press, New York, pp. 21–43.
- [15] Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.
- [16] Nesselroade, J.R. & Baltes, P.B., eds (1979). *Longitudinal Research in the Study of Behavior and Development*, Academic Press, New York.
- [17] Prinzie, P., Onghena, P. & Hellinckx, W. (in press). Parent and child personality traits and children's externalizing problem behavior from age 4 to 9 years: a cohort-sequential latent growth curve analysis, *Merrill-Palmer Quarterly*.
- [18] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, Newbury Park.
- [19] Raudenbush, S.W. & Chan, W.S. (1992). Growth curve analysis in accelerated longitudinal designs with application to the National Youth Survey, *Journal of Research on Crime and Delinquency* **29**, 387–411.
- [20] Schaie, K.W. (1965). A general model for the study of developmental problems, *Psychological Bulletin* **64**, 92–107.
- [21] Schaie, K.W. & Baltes, P.B. (1975). On sequential strategies in developmental research: Description or explanation? *Human Development* **18**, 384–390.
- [22] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
- [23] Wendorf, C.A. (2002). Comparisons of structural equation modeling and hierarchical linear modeling approaches to couples' data, *Structural Equation Modeling* **9**, 126–140.

PETER PRINZIE AND PATRICK ONGHENA

Cohort Studies

ROSS L. PRENTICE

Volume 1, pp. 322–326

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cohort Studies

Background

Cohort studies constitute a central epidemiologic approach to the study of relationships between personal characteristics or exposures and the occurrence of health-related events, and hence to the identification of disease prevention hypotheses and strategies.

Consider a conceptually infinite population of individuals moving forward in time. A cohort study involves sampling a subset of such individuals and observing the occurrence of events of interest, generically referred to as *disease events*, over some follow-up period. Such a study may be conducted to estimate the rates of occurrence of the diseases to be ascertained, but most often, estimation of relationships between such rates and individual characteristics or exposures is the more fundamental study goal. If cohort-study identification precedes the follow-up period, the study is termed *prospective*, while a retrospective or historical cohort study involves cohort identification after a conceptual follow-up period (see **Prospective and Retrospective Studies**). This presentation assumes a prospective design.

Other research strategies for studying exposure-disease associations, and for identifying disease prevention strategies, include **case-control studies** and randomized controlled disease prevention trials. Compared to case-control studies, cohort studies have the advantages that a wide range of health events can be studied in relation to exposures or characteristics of interest and that prospectively ascertained exposure data are often of better quality than the retrospectively obtained data that characterize case-control studies. On the other hand, a cohort study of a particular association would typically require much greater cost and longer duration than would a corresponding case-control study, particularly if the study disease is rare. Compared to randomized controlled trials, cohort studies have the advantage of allowing study of a broad range of exposures or characteristics in relation to health outcomes of interest, and typically of much simplified study logistics and reduced cost. Randomized intervention trials can also examine a broad range of exposures and disease associations in an observational manner, but the randomized assessments are necessarily restricted to

examining the health consequences of a small number of treatments or interventions. On the other hand, disease prevention trials have the major advantage that the intervention comparisons are not confounded by prerandomization disease risk factors, whether or not these are even recognized. The choice among research strategies may depend on the distribution of the exposures in the study population and on the ability to reliably measure such exposures, on the knowledge and measurement of confounding factors, on the reliability of outcome ascertainment, and on study costs in relation to the public health potential of study results.

There are many examples of associations that have been identified or confirmed using cohort study techniques, including that between cigarette smoking and lung cancer; between blood pressure, blood cholesterol, cigarette smoking, and coronary heart disease; between current use of the original combined oral contraceptives and the risk of various vascular diseases; and between atomic bomb radiation exposure and the risk of leukemia or of various solid tumors, to name a few. In recent years, there have also been many examples of the use of cohort study designs to examine the association between exposures that are difficult to measure, or that may have limited within-cohort exposure variability, and the occurrence of disease. Such examples may involve, for example, physical activity, dietary, environmental, or occupational exposures, and behavioral characteristics. In these settings, cohort studies sometimes yield weak or equivocal results, and multiple cohort studies of the same general association may yield contradictory results. It is important to be able to anticipate the reliability (see **External Validity; Internal Validity**) **power** of cohort studies, to be aware of strategies for enhancing study power and reliability, and to carefully consider optimal research strategies for assessing specific exposure-disease hypotheses.

Basic Cohort Study Elements

Exposure Histories and Disease Rates

A general regression (see **Regression Models**) notation can be used to represent the exposures (and characteristics) to be ascertained in a cohort study. Let $z_1(u)^T = \{z_{11}(u), z_{12}(u), \dots\}$ denote a set of numerically coded variables that describe an individual's characteristics at 'time' u , where, to be specific,

u can be defined as time from selection into the cohort, and ' T ' denotes vector transpose. Let $Z_1(t) = \{z_1(u), u < t\}$ denote the history of such characteristics at times less than t . Note that the 'baseline' exposure data $Z_1(0)$ may include information that pertains to time periods prior to selection into the cohort. Denote by $\lambda\{t; Z_1(t)\}$ the population **incidence** (hazard) rate at time t for a disease of interest, as a function of an individual's preceding 'covariate' history. A typical cohort study goal is the elucidation of the relationship between aspects of $Z_1(t)$ and the corresponding disease rate $\lambda\{t; Z_1(t)\}$. As mentioned above, a single cohort study may be used to examine many such covariate-disease associations.

The interpretation of the relationship between $\lambda\{t; Z_1(t)\}$ and $Z_1(t)$ may well depend on other factors. Let $Z_2(t)$ denote the history up to time t of a set of additional characteristics. If the variates $Z_1(t)$ and $Z_2(t)$ are related among population members at risk for disease at time t and if the disease rate $\lambda\{t; Z_1(t), Z_2(t)\}$ depends on $Z_2(t)$, then an observed relationship between $\lambda\{t; Z_1(t)\}$ and $Z_1(t)$ may be attributable, in whole or in part, to $Z_2(t)$. Hence, toward an interpretation of causality, one can focus instead on the relationship between $Z_1(t)$ and the disease rate function $\lambda\{t; Z_1(t), Z_2(t)\}$, thereby controlling for the 'confounding' influences of Z_2 . In principle, a cohort study needs to control for all pertinent confounding factors in order to interpret a relationship between Z_1 and disease risk as causal. It follows that a good deal must be known about the disease process and disease risk factors before an argument of causality can be made reliably. This feature places a special emphasis on the replication of results in various populations, with the idea that unrecognized or unmeasured confounding factors may differ among populations. As noted above, the principle advantage of a randomized disease prevention trial, as compared to a purely observational study, is that the randomization indicator variable $Z_1 = Z_1(0)$, where here $t = 0$ denotes the time of randomization, is unrelated to the histories $Z_2(0)$ of all confounding factors, whether or not such are recognized or measured.

The choice as to which factors to include in $Z_2(t)$, for values of t in the cohort follow-up period, can be far from being straightforward. For example, factors on a causal pathway between $Z_1(t)$ and disease risk may give rise to 'over adjustment' if included in $Z_2(t)$, since one of the mechanisms

whereby the history $Z_1(t)$ alters disease risk has been conditioned upon. On the other hand, omission of such factors may leave a confounded association, since the relationship between Z_2 and disease risk may not be wholly attributable to the effects of Z_1 on Z_2 .

Cohort Selection and Follow-up

Upon identifying the study diseases of interest and the 'covariate' histories $Z(t) = \{Z_1(t), Z_2(t)\}$ to be ascertained and studied in relation to disease risk, one can turn to the estimation of $\lambda\{t; Z(t)\}$ based on a cohort of individuals selected from the study population. The basic cohort selection and follow-up requirement for valid estimation of $\lambda\{t; Z(t)\}$ is that at any $\{t, Z(t)\}$ a sample that is representative of the population in terms of disease rate be available and under active follow-up for disease occurrence. Hence, conceptually, cohort selection and censoring rates (e.g., loss to follow-up rates) could depend arbitrarily on $\{t, Z(t)\}$, but selection and follow-up procedures cannot be affected in any manner by knowledge about, or perception of, disease risk at specified $\{t, Z(t)\}$.

Covariate History Ascertainment

Valid estimation of $\lambda\{t; Z(t)\}$ requires ascertainment of the individual study subject histories, Z , during cohort follow-up. Characteristics or exposures prior to cohort study enrollment are often of considerable interest, but typically need to be ascertained retrospectively, perhaps using specialized questionnaires, using analysis of biological specimens collected at cohort study entry, or by extracting information from existing records (e.g., employer records or occupational exposures). Postenrollment exposure data may also need to be collected periodically over the cohort study follow-up period to construct the histories of interest. In general, the utility of cohort study analyses depends directly on the extent of variability in the 'covariate' histories Z , and in the ability to document that such histories have been ascertained in a valid and reliable fashion. It often happens that aspects of the covariate data of interest are ascertained with some measurement error, in which case substudies that allow the measured quantities to be related to the underlying variables of interest (e.g., validation

or reliability substudies) may constitute a key aspect of cohort study conduct.

Disease Event Ascertainment

A cohort study needs to include a regular updating of the occurrence times for the disease events of interest. For example, this may involve asking study subjects to report a given set of diagnoses or health-related events (e.g., hospitalization) that initiate a process for collecting hospital and laboratory records to determine whether or not a disease event has occurred. Diagnoses that require considerable judgment may be further adjudicated by a panel of diagnostic experts. While the completeness of outcome ascertainment is a key feature of cohort study quality, the most critical outcome-related cohort study feature concerns whether or not there is differential outcome ascertainment, either in the recognition or the timely ascertainment of disease events of interest, across the exposures or characteristics under study. Differential ascertainment can often be avoided by arranging for outcome ascertainment procedures and personnel to be independent to exposure histories, through document masking and other means.

Data Analysis

Typically, a test of association between a certain characteristic or exposure and disease risk can be formulated in the context of a descriptive statistical model. With occurrence-time data, the Cox regression model [4], which specifies

$$\lambda\{t; Z(t)\} = \lambda_0(t) \exp\{X(t)^T \beta\}, \quad (1)$$

is very flexible and useful for this purpose. In this model, λ_0 is a 'baseline' disease rate model that need not be specified, $X(t)^T = \{X_1(t), \dots, X_p(t)\}$ is a modeled regression vector formed from $Z(t)$, and $\beta^T = (\beta_1, \dots, \beta_p)$ is a corresponding hazard ratio (**relative risk**) parameter to be estimated. Testing and estimation on β is readily carried out using a so-called partial likelihood function [5, 7]. For example, if X_1 defines an exposure variable (or characteristic) of interest, a test of $\beta_1 = 0$ provides a test of the hypothesis of no association between such exposure and disease risk over the cohort follow-up period, which controls for the potential confounding factors

coded by $X_2, \dots, X_p$. Virtually all statistical software packages include tests and **confidence intervals** on β under this model, as well as estimators of the cumulative baseline hazard rate.

The data-analytic methods for accommodating measurement error in covariate histories are less standardized. Various methods are available (e.g., [2]) under a classical measurement model (*see **Measurement: Overview***), wherein a measured regression variable is assumed to be the sum of the target variable plus measurement error that is independent, not only of the targeted value but also of other disease risk factors. With such difficult-to-measure exposures as those related to the environment, occupation, physical activity, or diet, a major effort may need to be expended to develop a suitable measurement model and to estimate measurement model parameters in the context of estimating the association between disease rate and the exposure history of interest.

An Example

While there have been many important past and continuing cohort studies over the past several decades, a particular cohort study in which the author is engaged is the Observational Study component of the Women's Health Initiative (WHI) [10, 18]. This study is conducted at 40 Clinical Centers in the United States, and includes 93 676 postmenopausal women in the age range 50–79 at the time of enrollment during 1993–1998. Cohort enrollment took place in conjunction with a companion multifaceted Clinical Trial among 68 133 postmenopausal women in the same age range, and for some purposes the combined Clinical Trial and Observational Study can be viewed as a cohort study in 161 809 women. Most recruitment took place using population-based lists of age- and gender-eligible women living in proximity to a participating Clinical Center.

Postmenopausal hormone use and nutrition are major foci of the WHI, in relation to disease morbidity and mortality. Baseline collection included a personal hormone history interview, a food frequency questionnaire, a blood specimen (for separation and storage), and various risk factor and health-behavior questionnaires. Outcome ascertainment includes periodic structured self-report of a broad range of health outcomes, document collection by a trained outcome specialist, physician adjudication at each Clinical Center, and subsequent centralized adjudication

for selected outcome categories. Exposure data are updated on a regular basis either through questionnaire or clinic visit. To date, the clinical trial has yielded influential, and some surprising, results on the benefits and risks of postmenopausal hormone therapy [17, 19]. The common context and data collection in the Observational Study and Clinical Trial provides a valuable opportunity to compare results on hormone therapy between the two study designs. A major study is currently being implemented using various objective markers of nutrient consumption toward building a suitable measurement model for calibrating the food frequency nutrient assessments and thereby providing reliable information on nutrient-disease associations. A subset of about 1000 Observational Study participants provided replicate data on various exposures at baseline and at 3 years from enrollment toward allowing for measurement error accommodation in exposure and confounding variables.

Concluding Comment

This entry builds substantially on prior cohort study reviews by the author [13, 14], which provide more detail on study design and analysis choices. There are a number of books and review articles devoted to cohort study methods [1, 2, 3, 6, 8, 9, 11, 12, 15, 16].

Acknowledgment

This work was supported by grant CA-53996 from the US National Cancer Institute.

References

- [1] Breslow, N.E. & Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2: *The Design and Analysis of Cohort Studies*, IARC Scientific Publications No. 82, International Agency for Research on Cancer, Lyon.
- [2] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*, Chapman & Hall, New York.
- [3] Checkoway, H., Pearce, N. & Crawford-Brown, D.J. (1989). *Research Methods in Occupational Epidemiology*, Oxford University Press, New York.
- [4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- [5] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.
- [6] Kahn, H.A. & Sempos, C.T. (1989). *Statistical Methods in Epidemiology*, Oxford University Press, New York.
- [7] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition, John Wiley & Sons.
- [8] Kelsey, J.L., Thompson, W.D. & Evans, A.S. (1986). *Methods in Observational Epidemiology*, Oxford University Press, New York.
- [9] Kleinbaum, D.G., Kupper, L.L. & Morganstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*, Lifetime Learning Publications, Belmont.
- [10] Langer, R.D., White, E., Lewis, C.E., Kotchen, J.M., Hendrix, S.L. & Trevisan, M. (2003). The Women's Health Initiative Observational Study: baseline characteristics of participants and reliability of baseline measures, *Annals of Epidemiology* **13**(95), 107–121.
- [11] Miettinen, O.S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*, Wiley, New York.
- [12] Morganstern, H. & Thomas, D. (1993). Principles of study design in environmental epidemiology, *Environmental Health Perspectives* **101**, 23–38.
- [13] Prentice, R.L. (1995). Design issues in cohort studies, *Statistical Methods in Medical Research* **4**, 273–292.
- [14] Prentice, R.L. (1998). Cohort studies, in *Encyclopedia of Biostatistics*, Vol. 1, P. Armitage & T. Colton, eds, John Wiley & Sons, pp. 770–784.
- [15] Rothman, K.J. (1986). *Modern Epidemiology*, Little, Brown & Co., Boston.
- [16] Willett, W.C. (1998). *Nutritional Epidemiology*, 2nd Edition, Oxford University Press.
- [17] Women's Health Initiative Steering Committee. (2004). Effects of conjugated equine estrogens in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial, *Journal of the American Medical Association* **291**, 1701–1712.
- [18] Women's Health Initiative Study Group. (1998). Design of the Women's Health Initiative clinical trial and observational study, *Controlled Clinical Trials* **19**, 61–109.
- [19] Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative randomized controlled trial, *Journal of the American Medical Association* **288**, 321–333.

(See also **Case-Cohort Studies**)

ROSS L. PRENTICE

Coincidences

BRIAN S. EVERITT

Volume 1, pp. 326–327

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Coincidences

Consider a group of say r people; the probability that none of the people have a birthday (day and month, but not year) in common is given by the following formidable formula

$$\frac{365 \times 364 \times 363 \cdots \times (365 - r + 1)}{365^r}$$

Applying the formula for various values of r leads to the probabilities shown in Table 1.

The reason for including a group of 23 in the table is that it corresponds to a probability of just under a half. Consequently, the probability that *at least two* of the 23 people share a birthday is a little more than a half (.507). Most people, when asked to guess what size group is needed to achieve greater than a 50% chance that at least two of them share a birthday put the figure much higher than 23.

Birthday matches in a group of people is a simple example of a coincidence, a surprising concurrence of events that are perceived as meaningfully related but have no causal connection. (For other examples, enter ‘coincidences’ into *Google*!). Coincidences that self-styled ‘experts’ attach very low probabilities to (such as UFOs, corn circles, and weeping statues of the Virgin Mary) are all old favorites of the tabloid press. But it is not only *National Enquirer* and *Sun* readers who are fascinated by such occurrences. Even the likes of Arthur Koestler and Carl Jung have given coincidences serious consideration. Jung introduced the term *synchronicity* for what he saw as a causal connecting principle needed to explain coincidences,

arguing that such events occur far more frequently than chance allows. But Jung gets little support from one of the most important twentieth-century statisticians, **Ronald Alymer Fisher** who suggests that ‘the one chance in a million’ will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us.

Most statisticians would agree with Fisher (a wise policy because he was rarely wrong, and did not take kindly, or respond quietly, to disagreement!) and explain coincidences as due to the ‘law’ of truly large numbers (*see Laws of Large Numbers*). With a large enough sample, any outrageous thing is likely to happen. If, for example, we use a benchmark of ‘one in a million’ to define a surprising coincidence event, then in the United States, with its population of 280 million (2000) we would expect 280 coincidences a day and more than 100 000 such occurrences a year. Extending the argument from a year to a lifetime and from the population of the United States to that of the world (about 6 billion) means that incredibly remarkable events are almost certain to be observed. If they happen to us or to one of our friends, it is hard not to see them as mysterious and strange. But the explanation is not synchronicity, or something even more exotic, it is simply the action of chance.

Between the double-six throw of two dice and the perfect deal at bridge is a range of more or less improbable events that do sometimes happen – individuals *are* struck by lightening, *do* win a big prize in a national lottery, and *do* sometimes shoot a hole-in-one during a round of golf. Somewhere in this range of probabilities are those coincidences that give us an eerie, spine-tingling feeling, such as dreaming of a particular relative for the first time in years and then waking up to find that this person died in the night. Such coincidences are often regarded as weird when they happen to us or to a close friend, but in reality they are not weird, they are simply rare.

Even the most improbable coincidences *probably* result from the play of random events.

Further Reading

Everitt, B.S. (1999). *Chance Rules*, Springer, New York.

BRIAN S. EVERITT

Table 1 Probabilities that none of r people have a birthday in common

r	Probability that all r birthdays are different
2	0.997
5	0.973
10	0.883
20	0.589
23	0.493
30	0.294
50	0.030
100	0.0000031

Collinearity

GILAD CHEN

Volume 1, pp. 327–328

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Collinearity

Collinearity refers to the intercorrelation among independent variables (or predictors). When more than two independent variables are highly correlated, collinearity is termed *multicollinearity*, but the two terms are interchangeable [4]. Two or more independent variables are said to be collinear when they are perfectly correlated (i.e., $r = -1.0$ or $+1.0$). However, researchers also use the term collinearity when independent variables are highly, yet not perfectly, correlated (e.g., $r > 0.70$), a case formally termed *near collinearity* [4]. Collinearity is generally expressed as the multiple correlation among the i set of independent variables, or R_i^2 .

Perfect collinearity violates the assumption of ordinary least squares regression that there is no perfect correlation between any set of two independent variables [1]. In the context of **multiple linear regression**, perfect collinearity renders regression weights of the collinear independent variables uninterpretable, and near collinearity results in biased parameter estimates and inflated standard errors associated with the parameter estimates [2, 4].

Collinearity is obvious in multiple regression when the parameter estimates differ in magnitude and, possibly, in direction from their respective bivariate correlation and when the standard errors of the estimates are high [4]. More formally, the *variance inflation factor* (VIF) and *tolerance* diagnostics (see **Tolerance and Variance Inflation Factor**)

can help detect the presence of collinearity. VIF_i , reflecting the level of collinearity among an i set of independent variables, is defined as $1 / (1 - R_i^2)$, where R_i^2 is the squared multiple correlations among the i independent variables. A VIF value of 1 indicates no level of collinearity, and values greater than 1 indicate some level of collinearity. Tolerance is defined as $1 - R_i^2$ or $1/VIF$, where values equal to 1 indicate lack of collinearity, and 0 indicates perfect collinearity. Several remedies exist for reducing or eliminating the problems associated with collinearity, including dropping one or more collinear predictors, combining collinear predictors into a single composite, and employing ridge regression or principle components regression (see **Principal Component Analysis**) (e.g., [3]).

References

- [1] Berry, W.D. (1993). *Understanding Regression Assumptions*, Sage, Newbury Park.
- [2] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analyses for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.
- [3] Netter, J., Wasserman, W. & Kutner, M.H. (1990). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, 3rd Edition, Irwin, Boston.
- [4] Pedhazur, E. (1997). *Multiple Regression in Behavior Research*, 3rd Edition, Harcourt-Brace, Fort Worth.

GILAD CHEN

Combinatorics for Categorical Variables

ALKA INDURKHYA

Volume 1, pp. 328–330

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Combinatorics for Categorical Variables

Categorical Variables

Datasets gathered in the behavioral, social, and biomedical sciences using survey instruments tend to consist of categorical responses. In a survey, the response can vary from individual to individual, for example, the question, ‘Have you smoked a cigarette during the past week?’ may be answered using one of the three following responses ‘Yes’, ‘No’, ‘Can’t remember’. The responses are mutually exclusive, that is, each response belongs to only one of the three response categories. It is also referred to in the literature as a nominal variable because the response categories have labels (names) that lack an intrinsic order (see **Scales of Measurement**). Formally, a categorical or nominal variable is defined as a variable that has two or more mutually exclusive groups (or categories) that lack an intrinsic order. An example of a variable that does not have mutually exclusive categories is the psychiatric classification of diagnosis: neurotic, manic-depressive, schizophrenic, and other disorders. An individual could have more than one diagnosis, that is, have co-occurring mental disorders.

Sometimes, a categorical response may be ordered, for example, socioeconomic status may be grouped into ‘low’, ‘medium’, and ‘high’. In this case, the categorical variable is referred to as an *ordinal variable*. Thus, a categorical variable may be either nominal or ordinal. In the social and behavioral sciences, self-report of behavior is typically gathered using an ordinal scale ranging from 1 through 5, with response 1 for ‘never’, 2 for ‘rarely’, 3 for ‘sometimes’, 4 for ‘often’, and 5 representing ‘always’. This illustrates how a numerical structure can be imposed on an ordinal variable, that is, the responses can be mapped onto a numerical scale permitting one to perform numerical operations such as addition and obtain statistics such as the arithmetic mean and standard deviation.

A dichotomous categorical variable is defined as a variable with only two possible responses or outcomes, for example, Yes or No, or 0 or 1. A dichotomous variable is also called a *binomial variable*, for example, gender is a binomial variable. A categorical variable with more than two responses is called

a *multinomial variable*. Marital status is a multinomial variable as the response may be ‘Married’, ‘Separated’, ‘Divorced’, or ‘Single’. A multinomial variable is also referred to as a *polytomous variable*. The ordinal variable with responses mapped onto the numbers 1 through 5 is an example of an ordered polytomous variable.

Data is gathered to test study hypothesis or lend credence to theories, for example, are there more females with internalizing behavioral problems than males in the population? The question can be answered using the counts of males and females in the study sample with/without internalizing behavior problems and reporting the percentages as estimates for the males or females in the population.

Probability and Counting Rules

Probability is defined as the likelihood or chance that a particular response will be given, for example, the chance that the next sampled individual is male. In empirical classical probability, the probability is defined as the ratio of the number of male individuals in the sample to the total number of individuals sampled. This differs from the *a priori* classical definition of probability in that it assumes no *a priori* knowledge of the populations. The *a priori* classical probability is the ratio of the total number of males in the population to the total population. In the social and behavioral sciences, the empirical classical probability is used to estimate and make inferences about population characteristics in the absence of *a priori* knowledge.

Notice that in the definition of probability, one is counting the number of times the desired response, for example, male, is observed. Counting rules come in handy for circumstances where it is not feasible to list all possible ways in which a desired response might be obtained.

Suppose an individual is randomly selected from the population and the gender of the individual noted. This process is repeated 10 times. How would we determine the number of different possible responses, that is, the sequences of males and females?

Counting Rule 1. If any one of k mutually exclusive responses are possible at each of n trials, then the number of possible outcomes is equal to k^n .

Using rule 1 suggests that there are $2^{10} = 1024$ possible sequences of males and females.

2 Combinatorics for Categorical Variables

The second counting rule involves the computation of the number of ways that a set of responses can be arranged in order. Suppose that there are six patients requesting an appointment to see their psychiatrist. What are the total number of ways in which the receptionist may schedule them to see the psychiatrist on the same day?

Counting Rule 2. The number of ways in which n responses can be arranged in order is given by $n! = n(n-1)(n-2)\dots(3)(2)(1)$, where $n!$ is called n factorial and $0!$ is defined to be 1.

An application of Rule 2 shows that the receptionist has $6! = (6)(5)(4)(3)(2)(1) = 720$ ways to schedule them.

But if the psychiatrist can only see 4 patients on that day, in how many ways can the receptionist schedule them in order?

Counting Rule 3. The number of arrangements for k responses selected from n responses in order is $n!/(n-k)!$. This is called the rule of permutations.

Using the rule of permutations, the receptionist has

$$\frac{6!}{(6-4)!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)} = 360 \text{ ways.} \quad (1)$$

But what if the receptionist is not interested in the order but only in the number of ways that any 4 of the 6 patients can be scheduled?

Counting Rule 4. The number of ways in which k responses can be selected from n responses is $n!/k!(n-k)!$. This is called the rule of combinations and the expression is commonly denoted by $\binom{n}{k}$.

The combinations counting rule shows that the receptionist has

$$\frac{6!}{4!(6-4)!} = \frac{6!}{4!2!} = \frac{(6)(5)(4)(3)(2)(1)}{(4)(3)(2)(1)(2)(1)} = 15 \text{ ways.} \quad (2)$$

The main difference between a permutation and a combination is that in the former the order in which

the first four patients out of six call in are each considered distinct, while in the latter the order of the first four is not maintained, that is, the four that call in are considered as the set of individuals that get the appointment for the same day but are not necessarily scheduled in the order they called in.

Example Suppose that a receptionist schedules a total of k patients over n distinct days. What is the probability that t patients are scheduled on a specific day?

First, an application of counting rule 4 to determine the number of ways t patients can be selected out of k gives $\binom{k}{t}$ total ways for choosing $t = 0, 1, 2, \dots, k$ patients scheduled on a specific day. Using counting rule 1, one can compute that the remaining $(k-t)$ patients can be scheduled over the remaining $(n-1)$ ways in a total of $(n-1)^{k-t}$ possible ways.

There are a total of n^k possible ways of randomly scheduling k patients over n days using counting rule 1.

The empirical probability that t patients

are scheduled on a specific day

$$\begin{aligned} & \text{number of ways of scheduling } t \text{ of} \\ & \quad k \text{ people on a specific day} \\ &= \frac{\text{number of ways of scheduling}}{k \text{ people on } n \text{ days}} \\ &= \frac{\binom{k}{t} (n-1)^{k-t}}{n^k}. \end{aligned} \quad (3)$$

This last expression can be rewritten as $\binom{k}{t} (1/n^t) (1 - (1/n))^{k-t}$, which is the empirical form of the binomial distribution.

(See also **Catalogue of Probability Density Functions**)

ALKA INDURKHYA

Common Pathway Model

FRÜHLING RIJSDIJK

Volume 1, pp. 330–331

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Common Pathway Model

The common pathway model, as applied to genetically sensitive data, is a multivariate model in which the covariation between a group of variables is controlled by a single phenotypic latent factor (*see Latent Variable*), with direct paths to each variable [2]. This intermediate factor itself is influenced by genetic and environmental latent factors. The term ‘common’ refers to the fact that the effects of the genetic and environmental factors on all observed variables will impact via this single factor. Because the phenotypic latent factor has no scale, the E_c path (residual variance) is fixed to unity to make the model identified. There is also a set of specific genetic and environmental factors accounting for residual, variable specific variances (see Figure 1). For these specific factors to all have free loadings, the minimal number of variables in this model is three.

For twin data, two identical common pathway models are modeled for each twin with the genetic and environmental factors across twins (both common

and specific) connected by the expected correlations, 1 for MZ twins, 0.5 for DZ twins (see Figure 1).

Similar to the *univariate genetic model*, the MZ and DZ ratio of the cross-twin within variable correlations (e.g., Twin 1 variable 1 and Twin 2 variable 1) will indicate the relative importance of genetic and environmental variance components for each variable. On the other hand, the MZ and DZ ratio of the cross-twin cross-trait correlations (e.g., Twin 1 variable 1 and Twin 2 variable 2) will determine the relative importance of genetic and environmental factors in the covariance between variables (i.e., genetic and environmental correlations). In addition, for any two variables it is possible to derive the part of the phenotypic correlation determined by common genes (which will be a function of both their h^2 (*see Heritability*) and genetic correlation) and by common shared and unique environmental effects (which will be a function of their c^2 and e^2 , and the C and E correlation). For more information on genetic and environmental correlations between variables, see the general section on **multivariate genetic analysis**. Parameter estimates are estimated from the observed

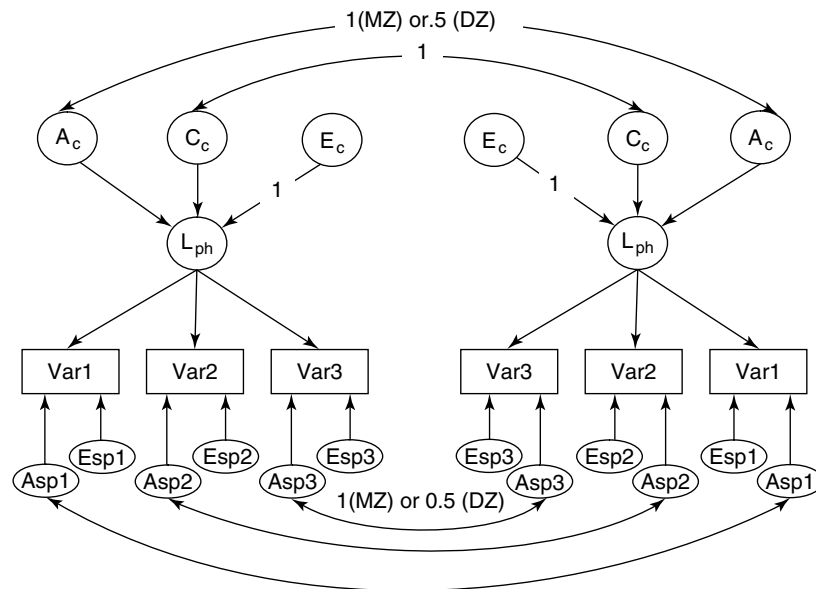


Figure 1 Common pathway model: A_c , C_c , and E_c are the common additive genetic, common shared, and common nonshared environmental factors, respectively. L_{ph} is the latent intermediate phenotypic variable, which influences all observed variables. The factors at the bottom are estimating the variable specific A and E influences. For simplicity, the specific C factors were omitted from the diagram

2 Common Pathway Model

variances and covariances by fitting **structural equation models**. The common pathway model is more parsimonious than the independent pathway model because it estimates fewer parameters.

So what is the meaning and interpretation of this factor model? The common pathway model is a more stringent model than the **independent pathway model**. It hypothesizes that covariation between variables arises purely from their phenotypic relation with the latent intermediate variable. This factor is identical to the factor derived from higher order phenotypic factor analyses, with the additional possibility to estimate the relative importance of genetic and environmental effects of this factor. In contrast, in the independent pathway model, where the common genetic and environmental factors influence the observed variables directly, different clusters of variables for the genetic and environmental factors are possible. This means that some variables could be specified to covary mainly because of shared genetic effects, whereas others covary because of shared environmental effects.

An obvious application of this model is to examine the etiology of **comorbidity**. In an adolescent twin sample recruited through the Colorado Twin Registry and the Colorado Longitudinal Twin Study, conduct disorder and attention deficit hyperactivity disorder, along with a measure of substance experimentation and novelty seeking, were used as indices of a latent *behavioral disinhibition* trait. A common pathway model evaluating the genetic and environmental architecture of this latent phenotype suggested that behavioral disinhibition is highly heritable (0.84), and is not influenced significantly by shared environmental factors. These results suggest that a variety of

adolescent problem behaviors may share a common underlying genetic risk [3].

Another application of this model is to determine the variation in a behavior that is agreed upon by multiple informants. An example of such an application is illustrated for antisocial behavior in 5-year-old twins as reported by mothers, teachers, examiners, and the children themselves [1]. Problem behavior ascertained by consensus among raters in multiple settings indexes cases of problem behavior that are pervasive. Heritability of this pervasive antisocial behavior was higher than any of the informants individually (which can be conceptualized as situational antisocial behavior). In addition, significant informant specific unique environment (including) measurement error was observed.

References

- [1] Arseneault, L., Moffitt, T.E., Caspi, A., Taylor, A., Rijdsdijk, F.V., Jaffee, S., Ablow, J.C. & Measelle, J.R. (2003). Strong genetic effects on antisocial behaviour among 5-year-old children according to mothers, teachers, examiner-observers, and twins' self-reports, *Journal of Child Psychology and Psychiatry* **44**, 832–848.
- [2] McArdle, J.J. & Goldsmith, H.H. (1990). Alternative common-factor models for multivariate biometrical analyses, *Behavior Genetics* **20**, 569–608.
- [3] Young, S.E., Stallings, M.C., Corley, R.P., Krauter, K.S. & Hewitt, J.K. (2000). Genetic and environmental influences on behavioral disinhibition, *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **96**, 684–695.

FRÜHLING RIJSDIJK

Community Intervention Studies

DAVID M. MURRAY

Volume 1, pp. 331–333

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Community Intervention Studies

Introduction

Community intervention trials are an example of the broader class of studies called *group-randomized trials* (GRTs), in which investigators randomize identifiable groups to conditions and observe members of those groups to assess the effects of an intervention [5]. Just as the familiar randomized clinical trial is the gold standard when randomization of individuals to study conditions is possible, the GRT is the gold standard when **randomization** of identifiable groups is required. That situation exists for community-based interventions, which typically operate at a group level, manipulate the social or physical environment, or cannot be delivered to individuals.

An Example

The Tobacco Policy Options for Prevention study was a community intervention trial designed to test the effects of changes in local policies to limit youth access to tobacco [3]. After stratifying on population size and baseline adolescent smoking rate, 14 communities were randomized to the intervention or control condition. The 32-month intervention was designed to change local ordinances to restrict youth access to tobacco, to change retailers' practices regarding provision of tobacco to youth, and to increase enforcement of tobacco age-of-sale laws. Data were collected from students in grades 8–10 and from purchase attempt surveys at retail outlets, both before the intervention and three years later. Daily smoking was reduced by one-third and students reported significantly lower availability of tobacco products from commercial sources [3].

Distinguishing Characteristics

Four characteristics distinguish the GRT [5]. First, the unit of assignment is an identifiable group. Second, different groups are assigned to each condition, creating a nested or hierarchical structure. Third, the units of observation are members of those groups so

that they are nested within both their condition and their group. Fourth, there usually is only a limited number of groups assigned to each condition.

These characteristics create several problems for the design and analysis of GRTs. The major design problem is that a limited number of often-heterogeneous groups makes it difficult for randomization to distribute potential sources of confounding evenly in any single realization of the experiment. This increases the need to employ design strategies that will limit confounding and analytic strategies to deal with confounding where it is detected. The major analytic problem is that there is an expectation for positive **intraclass correlation (ICC)** among observations on members of the same group [4]. That ICC reflects an extra component of variance attributable to the group above and beyond the variance attributable to its members. This extra variation will increase the variance of any group-level statistic beyond what would be expected with random assignment of members to conditions. With a limited number of groups, the degrees of freedom (df) available to estimate group-level statistics are limited. Any test that ignores either the extra variation or the limited df will have an inflated Type I error rate [1].

Potential Design and Analysis Problems and Methods to Avoid Them

For GRTs, there are four sources of bias that should be considered during the planning phase: selection, differential history, differential maturation, and contamination (*see Quasi-experimental Designs*). The first three are particularly problematic in GRTs where the number of units available for randomization is often small. GRTs planned with fewer than 20 groups per condition would be well served to include careful matching or **stratification** prior to randomization to help avoid these biases. Analytic strategies, such as regression adjustment for confounders, can be very helpful in dealing with any bias observed after randomization (*see Regression Models*).

There are two major threats to the validity of the analysis of a GRT, which should be considered during the planning phase: misspecification of the analytic model and low **power**. Misspecification of the analytic model most commonly occurs when the investigator fails to reflect the expected ICC in the analytic model. Low power most commonly occurs

when the design is based on an insufficient number of groups randomized to each condition.

There are several analytic approaches that can provide a valid analysis for GRTs [2, 5]. In most, the intervention effect is defined as a function of a condition-level statistic (e.g., difference in means, rates, or slopes) and assessed against the variation in the corresponding group-level statistic. These approaches included mixed-model ANOVA/ANCOVA for designs having only one or two time intervals (*see Linear Multilevel Models*), random coefficient models for designs having three or more time intervals, and **randomization tests** as an alternative to the model-based methods. Other approaches are generally regarded as invalid for GRTs because they ignore or misrepresent a source of random variation. These include analyses that assess condition variation against individual variation and ignore the group, analyses that assess condition variation against individual variation and include the group as a fixed effect, analyses that assess the condition variation against subgroup variation, and analyses that assess condition variation against the wrong type of group variation. Still other strategies may have limited application for GRTs. For example, the application of **generalized estimating equations (GEE)** and the sandwich method for standard errors requires a total of 40 or more groups in the study, or a correction for the downward bias in the sandwich estimator for standard errors when there are fewer than 40 groups [7].

To avoid low power, investigators should plan a large enough study to ensure sufficient replication, employ more and smaller groups instead of a few large groups, employ strong interventions with a good reach, and maintain the reliability of intervention implementation. In the analysis, investigators should consider regression adjustment for covariates, model time if possible, and consider post hoc stratification.

Excellent treatments on power for GRTs exist, and the interested reader is referred to those sources for additional information. Chapter 9 in the Murray text provides perhaps the most comprehensive treatment of detectable difference, sample size, and power for GRTs [5]. Even so, a few points are repeated here. First, the increase in between-group variance due to the ICC in the simplest analysis is calculated as $1 + (m - 1)ICC$, where m is the number of members per group; as such, ignoring even a small ICC can underestimate standard errors if m is large. Second,

more power is available given more groups per condition with fewer members measured per group than given just a few groups per condition with many members measured per group, no matter the size of the ICC. Third, the two factors that largely determine power in any GRT are the ICC and the number of groups per condition. For these reasons, there is no substitute for a good estimate of the ICC for the primary endpoint, the target population, and the primary analysis planned for the trial, and it is unusual for a GRT to have adequate power with fewer than 8–10 groups per condition. Finally, the formula for the standard error for the intervention effect depends on the primary analysis planned for the trial, and investigators should take care to calculate that standard error, and power, based on that analysis.

Acknowledgment

The material presented here draws heavily on work published previously by David M. Murray [5–7]. Readers are referred to those sources for additional information.

References

- [1] Cornfield, J. (1978). Randomization by group: a formal analysis, *American Journal of Epidemiology* **108**(2), 100–102.
- [2] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*, Arnold, London.
- [3] Forster, J.L., Murray, D.M., Wolfson, M., Blaine, T.M., Wagenaar, A.C. & Hennrikus, D.J. (1998). The effects of community policies to reduce youth access to tobacco, *American Journal of Public Health* **88**(8), 1193–1198.
- [4] Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.
- [5] Murray, D.M. (1998). *Design and Analysis of Group-randomized Trials*, Oxford University Press, New York.
- [6] Murray, D.M. (2000). Efficacy and effectiveness trials in health promotion and disease prevention: design and analysis of group-randomized trials, in *Integrating Behavioral and Social Sciences with Public Health*, N. Schneiderman, J.H. Gentry, J.M. Silva, M.A. Speers & H. Tomes, eds, American Psychological Association, Washington, pp. 305–320.
- [7] Murray, D.M., Varnell, S.P. & Blitstein, J.L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments, *American Journal of Public Health* **94**(3), 423–432.

DAVID M. MURRAY

Comorbidity

S.H. RHEE, JOHN K. HEWITT, R.P. CORLEY AND M.C. STALLINGS

Volume 1, pp. 333–337

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Comorbidity

Comorbidity is the potential co-occurrence of two disorders in the same individual, family etc. According to epidemiological studies (e.g., [9, 17]), comorbidity between psychiatric disorders often exceeds the rate expected by chance alone. Increased knowledge regarding the causes of comorbidity between two psychiatric disorders may have a significant impact on research regarding the classification, treatment, and etiology of both disorders. Therefore, several researchers have proposed alternative models explaining the etiology of comorbidity (e.g., [1, 2, 5], and [16]), and ways to discriminate the correct explanation for comorbidity between two or more alternatives (e.g., [7, 8], and [10]). To date, Klein and Riso have presented the most comprehensive set of alternative comorbidity models, and Neale and Kendler presented the quantitative specifications of Klein and Riso's models (see Table 1). In a series of studies, we examined the validity of common methods used to test alternative comorbidity models [11–14].

Common Methods Used to Test Alternative Comorbidity Models

Klein and Riso's Family Prevalence Analyses.

For each comorbidity model, Klein & Riso [7] presented a set of predictions regarding the prevalence of disorders in the relatives of different groups of probands. They presented a comprehensive set of predictions comparing the prevalence of disorder A-only, disorder B-only, and disorder AB (i.e., both disorders) among the relatives of probands with A-only, B-only, AB, and controls. Several studies have used these predictions to test alternative comorbidity models (e.g., [6, 15], and [18]).

Family Prevalence Analyses in the Literature. Many other researchers (e.g., [3] and [4]) have conducted a subset of the Klein and Riso analyses or analyses very similar to those presented by Klein and Riso [7] without testing their comprehensive set of predictions. Most of the studies in the literature have focused on three comorbidity models, (a) the alternate forms model (i.e., the two comorbid disorders are

alternate manifestations of the same underlying liability), (b) the correlated liabilities model (i.e., there is a significant correlation between the liabilities for the two models), and (c) the three independent disorders model (i.e., the comorbid disorder is a third disorder that is etiologically distinct from either disorder occurring alone).

Neale and Kendler Model-fitting Approach.

Neale and Kendler [8] described 13 alternative models. They illustrated the probabilities for the four combinations of disease state ((a) neither A nor B; (b) A but not B; (c) B but not A; (d) both A and B) for each comorbidity model, then illustrated the probabilities for the 10 combinations of the affected or unaffected status for pairs of relatives for each comorbidity model (e.g., neither A nor B in relative 1 and neither A nor B in relative 2; both A and B in relative 1 and A only in relative 2, etc.). The data that are analyzed in the Neale and Kendler model-fitting approach is simply the frequency tables for the number of relative pairs in each possible combination of a disease state. The observed cell frequencies are compared with the expected cell frequencies (i.e., the probabilities for the 10 combinations of the affected or unaffected status for pairs of relatives) in each comorbidity model. The comorbidity model with the smallest difference between the observed cell frequencies and the expected cell frequencies is chosen as the best fitting model.

Underlying Deficits Approach.

Several researchers have tested alternative comorbidities by comparing the level of underlying deficits of the two comorbid disorders in individuals with neither disorder, A only, B only, and both A and B. For example, Pennington, Groisser, and Welsh [10] examined the comorbidity between reading disability and attention deficit hyperactivity disorder (ADHD), comparing the underlying deficits associated with reading disability (i.e., phonological processes) and the underlying deficits associated with ADHD (i.e., executive functioning) in individuals with neither disorder, reading disability only, ADHD only, and both reading disability and ADHD. Most of the researchers using this approach have made predictions for 5 of the 13 Neale and Kendler comorbidity models. In addition to the three models often tested using family prevalence analyses in the literature (i.e., alternate forms, correlated liabilities, and three independent disorders),

2 Comorbidity

Table 1 Description of the Neale and Kendler models

Name	Abbreviation	Description
Chance	CH	Comorbidity is due to chance.
Alternate forms	AF	Two disorders are alternate manifestations of a single liability.
Random multiformity	RM	Being affected by one disorder directly increases the risk for having the other disorder. (Individuals with disorder A have risk of disorder B increased by probability p and individuals with disorder B have risk of disorder A increased by probability r .)
Random multiformity of A	RMA	Submodel of RM where r is 0.
Random multiformity of B	RMB	Submodel of RM where p is 0.
Extreme multiformity	EM	Being affected by one disorder directly increases the risk for having the other disorder. (Individuals with disorder A have disorder B if they surpass a higher threshold on the liability for disorder A and individuals with disorder B have disorder A if they surpass a higher threshold on the liability for disorder B.)
Extreme multiformity of A	EMA	Submodel of EM where there is no second, higher threshold on the liability for disorder B.
Extreme multiformity of B	EMB	Submodel of EM where there is no second, higher threshold on the liability for disorder A.
Three independent disorders	TD	The comorbid disorder is a disorder separate from either disorder occurring alone.
Correlated liabilities	CL	Risk factors for the two disorders are correlated.
A causes B	ACB	A causes B.
B causes A	BCA	B causes A.
Reciprocal causation	RC	A and B cause each other.

researchers have made predictions regarding the random multiformity of A or random multiformity of B models (i.e., an individual who has one disorder is at an increased risk for having the second disorder, although he or she may not have an elevated liability for the second disorder).

Simulations

In all studies, simulations were conducted to test the validity of the common methods used to test alternative comorbidity models. Data were simulated for each of the 13 Neale and Kendler comorbidity models. In these simulated data, the true cause of comorbidity is known. Then, analyses commonly used to test the alternative comorbidity models were conducted on each of the 13 simulated datasets. If a particular analysis is valid, the predicted result should be found in the data simulated for the comorbidity model, and the predicted result should not be found

in data simulated for other comorbidity models (i.e., the particular analysis should discriminate a particular comorbidity model from alternative hypotheses).

Description of the Results

Klein and Riso's Family Prevalence Analyses.

Most of Klein and Riso's predictions were validated by the simulation results, in that most of their predicted results matched the results in the simulated datasets. However, there were several notable differences between the predicted results and results obtained in the simulated datasets. Some of Klein and Riso's predictions were not obtained in the simulated results because of lack of power in the simulated datasets. Another reason for the discrepancy between the predicted results and the results in the simulated dataset was the predictions' lack of consideration of all possible pathways for the comorbid disorder,

notably the fact that there will be some individuals who have both disorders A and B due to chance.

Family Prevalence Analyses in the Literature. The results of the study examining the validity of family prevalence analyses found in the literature indicate that most of these analyses were not valid. There are some analyses that validly discriminate the alternate forms model from alternative models, but none of the analyses testing the correlated liabilities model and the three independent disorders model were valid. In general, these analyses did not consider the fact that although the predicted results may be consistent with a particular comorbidity model, they also may be consistent with several alternative comorbidity models.

Neale and Kendler Model-fitting Approach. In general, the Neale and Kendler model-fitting approach discriminated the following classes of models reliably: the alternate forms model, the random multiformity models (i.e., random multiformity, random multiformity of A, and random multiformity of B), the extreme multiformity models (i.e., extreme multiformity, extreme multiformity of A, and extreme multiformity of B), the three independent disorders model, and the correlated liabilities models (i.e., correlated liabilities, A causes B, B causes A, and the reciprocal causation). Discrimination within these classes of models was poorer. Results from simulations varying the **prevalences** of the comorbid disorders indicate that the ability to discriminate between models becomes poorer as the prevalence of the disorders decreases, and suggests the importance of considering the issue of **power** when conducting these analyses.

Underlying Deficits Approach. Given adequate power, the method of examining the underlying deficits of comorbid disorders can distinguish between all 13 Neale and Kendler comorbidity models, except the random multiformity, extreme multiformity, and three independent disorders models. As the sample sizes decreased and the magnitude of correlation between the underlying deficits and the symptom scores decreased, the ability to discriminate the correct comorbidity model from alternative hypotheses decreased. Again, the issue of power should be considered carefully.

Conclusions

In a series of simulation studies, we examined the validity of common methods used to test alternative comorbidity models. Although most of Klein and Riso's family prevalence analyses were valid, there were notable discrepancies between their predicted results and results found in the simulated datasets. Some of the family prevalence analyses found in the literature were valid predictors of the alternate forms model, but none were valid predictors of the correlated liabilities or three independent disorders models. The Neale and Kendler model-fitting approach and the method of examining the underlying deficits of comorbid disorders discriminated between several comorbidity models reliably, suggesting that these two methods may be the most useful methods found in the literature. Especially encouraging is the fact that some of the models that cannot be distinguished well using the Neale and Kendler model-fitting approach can be distinguished well by examining the underlying deficits of comorbid disorders, and vice versa. The best approach may be a combination of these two methods.

References

- [1] Achenbach, T.M. (1990/1991). "Comorbidity" in child and adolescent psychiatry: categorical and quantitative perspectives, *Journal of Child and Adolescent Psychopharmacology* **1**, 271–278.
- [2] Angold, A., Costello, E.J. & Erkanli, A. (1999). Comorbidity, *Journal of Child Psychology and Psychiatry* **40**, 57–87.
- [3] Biederman, J., Faraone, S.V., Keenan, K., Benjamin, J., Krifcher, B., Moore, C., Sprich-Buckminster, S., Ugaglia, K., Jellinek, M.S., Steingard, R., Spencer, T., Norman, D., Kolodny, R., Kraus, I., Perrin, J., Keller, M.B. & Tsuang, M.T. (1992). Further evidence of family-genetic risk factors in attention deficit hyperactivity disorder: patterns of comorbidity in probands and relatives in psychiatrically and pediatrically referred samples, *Archives of General Psychiatry* **49**, 728–738.
- [4] Bierut, L.J., Dinwiddie, S.H., Begleiter, H., Crowe, R.R., Hesselbrock, V., Nurnberger, J.I., Porjesz, B., Schuckit, M.A. & Reich, T. (1998). Familial transmission of substance dependence: alcohol, marijuana, cocaine, and habitual smoking, *Archives of General Psychiatry* **55**, 982–988.
- [5] Caron, C. & Rutter, M. (1991). Comorbidity in child psychopathology: concepts, issues and research strategies, *Journal of Child Psychology and Psychiatry* **32**, 1063–1080.

- [6] Donaldson, S.K., Klein, D.N., Riso, L.P. & Schwartz, J.E. (1997). Comorbidity between dysthymic and major depressive disorders: a family study analysis, *Journal of Affective Disorders* **42**, 103–111.
- [7] Klein, D.N. & Riso, L.P. (1993). Psychiatric disorders: problems of boundaries and comorbidity, in *Basic Issues in Psychopathology*, Costello C.G., ed., The Guilford Press, New York, pp. 19–66.
- [8] Neale, M.C. & Kendler, K.S. (1995). Models of comorbidity for multifactorial disorders, *American Journal of Human Genetics* **57**, 935–953.
- [9] Newman, D.L., Moffitt, T.E., Caspi, A., Magdol, L., Silva, P.A. & Stanton, W.R. (1996). Psychiatric disorder in a birth cohort of young adults: prevalence, comorbidity, clinical significance, and new case incidence from ages 11 to 21, *Journal of Consulting and Clinical Psychology* **64**, 552–562.
- [10] Pennington, B.F., Groisser, D. & Welsh, M.C. (1993). Contrasting cognitive deficits in attention deficit hyperactivity disorder versus reading disability, *Developmental Psychology* **29**, 511–523.
- [11] Rhee, S.H., Hewitt, J.K., Corley, R.P. & Stallings, M.C. (2003a). The validity of analyses testing the etiology of comorbidity between two disorders: comparisons of disorder prevalences in families, *Behavior Genetics* **33**, 257–269.
- [12] Rhee, S.H., Hewitt, J.K., Corley, R.P. & Stallings, M.C. (2003b). The validity of analyses testing the etiology of comorbidity between two disorders: a review of family studies, *Journal of Child Psychology and Psychiatry and Allied Disciplines* **44**, 612–636.
- [13] Rhee, S.H., Hewitt, J.K., Corley, R.P., Willcutt, E.G., & Pennington, B.F. (2004). *Testing hypotheses regarding the causes of comorbidity: examining the underlying deficits of comorbid disorders*, Manuscript submitted for publication.
- [14] Rhee, S.H., Hewitt, J.K., Lessem, J.M., Stallings, M.C., Corley, R.P. & Neale, M.C. (2004). The validity of the Neale and Kendler model fitting approach in examining the etiology of comorbidity, *Behavior Genetics* **34**, 251–265.
- [15] Riso, L.P., Klein, D.N., Ferro, T., Kasch, K.L., Pepper, C.M., Schwartz, J.E. & Aronson, T.A. (1996). Understanding the comorbidity between early-onset dysthymia and cluster B personality disorders: a family study, *American Journal of Psychiatry* **153**, 900–906.
- [16] Rutter, M. (1997). Comorbidity: concepts, claims and choices, *Criminal Behaviour and Mental Health* **7**, 265–285.
- [17] Simonoff, E. (2000). Extracting meaning from comorbidity: genetic analyses that make sense, *Journal of Child Psychology and Psychiatry* **41**, 667–674.
- [18] Wickramaratne, P.J. & Weissman, M.M. (1993). Using family studies to understand comorbidity, *European Archives of Psychiatry and Clinical Neuroscience* **243**, 150–157.

S.H. RHEE, JOHN K. HEWITT, R.P. CORLEY
AND M.C. STALLINGS

Compensatory Equalization

PATRICK ONGHENA

Volume 1, pp. 337–338

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Compensatory Equalization

Compensatory equalization refers to a phenomenon in intervention studies (*see* **Clinical Trials and Intervention Studies**) in which comparison groups not obtaining the preferred treatment are provided with compensations that make the comparison groups more equal than originally planned [1, 4]. The simplest case would be an intervention study with a treatment group and a no-treatment control group, in which the people in charge of the intervention program feel the control group is unfairly withheld from a beneficial treatment and provide alternative goods and services to the participants in that control group. The implication is that the difference between the outcomes on a posttest is not a reflection of the difference between the treatment group and the original control group, but rather of the difference between the experimental group and the ‘control + alternative goods and services’ group.

Instead of adding benefits to the control group, compensatory equalization can also result from removing beneficial ingredients from the favored treatment group for reasons of perceived unfairness. This has been called *compensatory deprivation* but because the net effect is the same (*viz.*, equalization), it can be looked upon as a mere variation on the same theme.

Compensatory equalization was first introduced by Cook and Campbell [1] and listed among their threats to **internal validity** (*see* **Quasi-experimental Designs**). However, Shadish, Cook, and Campbell [4] rightly classified this confounder among the threats to construct validity. Internal validity threats are disturbing factors, that can occur even without a treatment, while compensatory equalization is intimately related to the treatment. In fact, compensatory equalization occurs *because* a treatment was introduced and should be considered as an integral part of the conceptual structure of the treatment contrast. This immediately suggests an obvious way of dealing with this confounder: the compensations must be included as part of the treatment construct description [4].

Just like in **compensatory rivalry** and **resentful demoralization** in compensatory equalization as

well, the participants themselves may be the instigators of the bias. This happens if some participants have the impression that they have been put at a disadvantage and force the researchers, the administrators, or the program staff to deliver compensations. Unlike these two other threats to construct validity, however, the involvement of people responsible for the treatment administration (e.g., therapists or teachers) or of external goods and services is crucial. Notice also that these threats to construct validity cannot be ruled out or made implausible by resorting to random assignment of participants to the comparison groups. Because the social comparison during the study itself is the source of the distortion, compensatory equalization may be present in both randomized and nonrandomized studies.

An actual research example of compensatory equalization can be found in Birmingham’s Homeless Project in which standard day care for homeless persons with substance abuse problems was compared with an enhanced day treatment condition [3]. Service providers complained about the inequity that was installed and put additional services at the disposal of the people in the standard day care group. In this way, the outcome for the two groups became more similar than would have been if no additional services were provided.

Researchers who want to avoid or minimize the problem of compensatory equalization might try to isolate the comparison groups in time or space, or make participants or service providers unaware of the intervention being applied (like in double-blind clinical trials). Another strategy, which might be used in combination with these isolation techniques, is to instruct the service providers about the importance of treatment integrity, or more fundamentally, not to use any conditions that can be perceived as unfair. If lack of treatment integrity is suspected, then interviews with the service providers and the participants are of paramount importance. Statistical correction and **sensitivity analysis** [2] might be worth considering, but if major and systematic compensations are blatantly interfering with treatment implementation, it may be more prudent to admit a straight change in the treatment construct itself.

References

- [1] Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Rand-McNally, Chicago.

2 Compensatory Equalization

- [2] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.
- [3] Schumacher, J.E., Milby, J.B., Raczynski, J.M., Engle, M., Caldwell, E.S. & Carr, J.A. (1994). Demoralization and threats to validity in Birmingham's homeless project, in *Critically Evaluating the Role of Experiments*, K.J. Conrad, ed., Jossey Bass, San Francisco, pp. 41–44.
- [4] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

(See also **Adaptive Random Assignment**)

PATRICK ONGHENA

Compensatory Rivalry

KAREN M. CONRAD AND KENDON J. CONRAD

Volume 1, pp. 338–339

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Compensatory Rivalry

Compensatory rivalry is a potential threat to construct validity of the cause in intervention studies (*see Validity Theory and Applications*) [1, 3, 5]. Put more plainly, the causal construct of interest (i.e., intervention) is given to the experimental group, and not to the control group. However, the control group, contrary to the intention of investigators, obtains it. Although it is more likely a potential threat in experimental studies where subjects are randomly assigned to groups, it can operate in quasi-experimental studies as well (e.g., pretest–post–test *see Nonequivalent Group Design*). We limit our discussion to the true experiment.

Compensatory rivalry usually occurs when the study group that is not assigned the experimental treatment feels disadvantaged, disappointed, or left out and decides to obtain a similar intervention on its own. Of course, if the control group receives the intervention, this distorts the construct of ‘control group’ and investigators lose the logic for inferring the cause of the outcomes that are observed.

This threat is sometimes referred to as the *John Henry effect*, named after the legendary railroad worker who died after competing successfully with a steam drill because it threatened his and fellow workers’ jobs [4]. However, it may or may not involve intentional competition or rivalry with the experimental group. Control subjects may simply want to achieve the desirable outcome without regard to comparisons with experimental subjects. However, if something is at stake for the control group relative to the experimental group, then manipulation of the treatment or even of the outcome variable may be of concern. In the case of John Henry, a super performance by an extraordinary individual gave the appearance of successful competition by the control group. In some cases, members of the control group may find other ways to manipulate the outcome, (e.g., **cheating**).

To understand the threat of compensatory rivalry, consider the typical study hypothesis for experimental designs. In the true experiment with random assignment to groups, the groups are assumed to be equivalent on all variables except on exposure to the treatment. In this case, we expect that any differences observed on the posttest outcome measure to

be due to the treatment. We do not want the control group to seek out the treatment on its own since this would reduce the likelihood of us observing a treatment effect if there were one.

Consider as an example an experimental nutrition intervention study to examine the effect of a weight reduction program on obese school aged children. Those students assigned to the control group may be motivated to begin a nutrition program and, along with their parents, may seek out alternative programs. With an outcome measure of lean body mass, for example, the differences observed between the student groups would be lessened because of the alternative independent services received by the control group, and the effectiveness of the treatment would thus be masked.

While we are not likely to prevent compensatory rivalry from operating, we should be aware of its possibility, measure it, and if possible, statistically adjust for it. Minimally, we would want to consider its possibility in interpreting the study findings. Although not a foolproof solution, researchers may consider the use of a delayed-treatment design to minimize the problem of compensatory rivalry. In a delayed-treatment design, the control group waits to receive the treatment until after the experimental group receives it. In the delayed-treatment design, we would measure both groups at baseline, again after the experimental group received the intervention, and then again after the delayed-treatment group received the intervention. For example, in an attempt to minimize compensatory rivalry in a firefighter physical fitness intervention program, a delayed-treatment design was implemented [2]. Even though everyone would eventually receive the program, firefighters in the delayed-treatment group expressed disappointment in not being able to start the program right away. The investigators asked the firefighters in the delayed-treatment group to keep an exercise log while waiting to begin the fitness program. In this way, their self-reported physical activity could be adjusted for in the statistical analyses of program effects. Compensatory rivalry, like other potential threats to the validity of study conclusions, merits our constant vigilance.

References

- [1] Conrad, K.J. & Conrad, K.M. (1994). Re-assessing validity threats in experiments: focus on construct validity,

2 Compensatory Rivalry

- in *Critically Evaluating the Role of Experiments in Program Evaluation*, New Directions for Program Evaluation Series, K.J. Conrad, ed., Jossey-Bass, San Francisco.
- [2] Conrad, K.M., Reichelt, P.A., Meyer, F., Marks, B., Gacki-Smith, J.K., Robberson, J.J., Nicola, T., Rostello, K. & Samo, D. (2004). (manuscript in preparation). Evaluating changes in firefighter physical fitness following a program intervention.
- [3] Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Rand McNally College Publishing Company, Chicago.
- [4] Saretsky, G. (1972). The OEO P.C. experiment and the John Henry effect, *Phi Delta Kappan* **53**, 579–581.
- [5] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, New York.

KAREN M. CONRAD AND KENDON J. CONRAD

Completely Randomized Design

SCOTT E. MAXWELL

Volume 1, pp. 340–341

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Completely Randomized Design

The completely randomized design (CRD) is the most basic type of **experimental design**. As such, it forms the basis for many more complex designs. Nevertheless, it is important in its own right because it is one of the most prevalent experimental designs, not just in the behavioral sciences but also in a wide range of other disciplines. A primary reason for the popularity of this design in addition to its simplicity is that random assignment to treatment conditions provides a strong basis for causal inference. **Sir Ronald A. Fisher** is usually credited with this insight and for developing the foundations of the CRD.

The CRD can be thought of as any design where each individual is randomly assigned to one of two or more conditions, such that the probability of being assigned to any specific condition is the same for each individual. There is no requirement that this probability be the same for every condition. For example, with three conditions, the probabilities of assignment could be .50, .25, and .25 for groups 1, 2, and 3 respectively. From a statistical perspective, there are usually advantages to keeping these probabilities equal, but there are sometimes other considerations that favor unequal probabilities [3].

The CRD is very flexible in that the conditions can differ either qualitatively or quantitatively. The conditions can also differ along a single dimension or multiple dimensions (although in the latter case, the design will typically be conceptualized as a factorial design). In the simplest case, only a single response variable (i.e., dependent variable) is measured for each person, but the design also allows multiple dependent variables.

Data from the CRD are typically analyzed with **analysis of variance (ANOVA)** or **multivariate analysis of variance (MANOVA)** in the case of more than one dependent variable. A standard *F* test can be used to test a null hypothesis that the population means of all groups are equal to one another. Tests of comparisons (i.e., contrasts) (*see Multiple Comparison Procedures*) provide an adjunct to this omnibus *F* test in designs with more than two groups. Statistical tests should frequently be accompanied by

measures of **effect size** which include measures of proportion of explained variance as well as **confidence intervals** for mean differences. Experimental design texts [1, 2] provide further details of data analysis for the CRD.

Recent years have seen increased attention paid to planning an appropriate sample size in the CRD [1, 2]. An adequate sample size is important to achieve acceptable statistical **power** to reject the null hypothesis when it is false and to obtain sufficiently narrow confidence intervals for comparisons of mean differences. In fact, the single biggest disadvantage of the CRD is arguably that it often requires a very large sample size to achieve adequate power and precision. For example, even in the simple case of only 2 groups of equal size and an alpha level of 0.05 (two-tailed), a total sample size of 128 is necessary in order to have a 0.80 probability of detecting a medium effect size. The reason such large sample sizes are often needed is because all sources that make a specific person's score different from the mean score in that person's group are regarded as errors in the ANOVA model accompanying a CRD. This reflects the fact that the CRD relies exclusively on randomization of all relevant influences on the dependent variable instead of attempting to control for these influences. As a consequence, the error variance in the ANOVA associated with the CRD is frequently large, lowering power and precision. Primarily for this reason, more complex designs and analyses such as the **randomized block design** and **analysis of covariance** have been developed to control for extraneous influences and thus increase power and precision.

References

- [1] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [2] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Lawrence Erlbaum Associates, Mahwah.
- [3] McClelland, G.H. (1997). Optimal design in psychological research, *Psychological Methods* 2, 3–19.

SCOTT E. MAXWELL

Computational Models

ROBERT E. PLOYHART AND CRYSTAL M. HAROLD

Volume 1, pp. 341–343

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Computational Models

The Nature of Computational Models

Computational modeling involves developing a set of interrelated mathematical, logical, and statistical propositions to simulate behavior and mimic real-world phenomena [7, 15]. By developing a series of ‘if-then’ statements, modelers can create algorithms that predict certain levels of an output, given specified levels of an input (e.g., *if* unemployment exceeds ‘25%’, *then* we will experience a recession by the year ‘2010’). This technique, shown to be effective in fields such as biology, meteorology, and physics, is slowly being adopted by behavioral scientists. Because of its ability to simulate dynamic, complex, and stochastic phenomena, computational modeling has been touted as the ‘third scientific research discipline’ [7].

To build computational models, Whicker and Sigelman [13] advise that five elements be considered: assumptions, parameters, inputs, outputs, and algorithms. First, the modeler must make a number of assumptions, which, in effect, are the compromises a modeler must make for his/her simulation to work. Among these is the assumption that the model being developed is indeed similar to the behavioral processes people actually experience. Next, the modeler must decide which variables will be examined and manipulated (inputs), which variables will be held constant (parameters), and how inputs will affect the phenomena of interest (outputs). Computational modeling involves using these inputs and outputs to develop a series of if-then statements (algorithms), based on theory and previous empirical research [7]. Essentially, the modeler specifies a number of theory-based environmental and individual conditions that may impact the phenomenon of interest. Once these algorithms are developed, numbers are generated from programs such as Basic, SAS, or SPSS that test the specified relationships. Cognitive psychologists have also developed their own software (e.g., Adaptive Character of Thought Theory, ACT-R) that reflects the underlying theory [2, 3].

Given the degree of control and theoretical rigor, computational modeling is a technique that can greatly inform research and practice. There are a number of benefits to using computational modeling [15]. First, computational modeling allows the

researcher to examine phenomena impossible to study in most field or laboratory settings. For instance, cross-sectional and even many longitudinal studies (which usually span a limited time frame) fall short of fully explaining processes over long periods of time. Second, computational modeling affords the researcher with a high level of precision and control, ensuring that extraneous factors do not affect outcomes. As mentioned, researchers decide which parameters will be fixed and which variables/inputs will be manipulated in each model. Third, computational modeling can effectively handle and model the inherent nesting of individuals within work groups and organizations [6]. Finally, computational modeling is of greatest benefit when grounded in theory and preexisting research. By utilizing established research, computational models can more precisely estimate the magnitude and form (i.e., linear, curvilinear) of the relationship between variables.

Previous research has used modeling techniques to examine phenomena such as the consequences of withdrawal behavior [4] and the impact of faking on personality tests [14]. In this first example, Hanisch et al. [4] used a computer simulation (WORKER) that reproduced a ‘virtual organization’, thus allowing them to test how various organizational and individual factors influenced different withdrawal behaviors, and how these effects may change over time. Ployhart and Ehrhart [10] modeled how racial subgroup differences in test-taking motivation contribute to subgroup test score differences. Cognitive psychologists, such as Anderson [2, 3], have used ACT-R to model human cognition; and in particular, have modeled how humans acquire procedural and declarative knowledge and the acquisition of problem-solving and decision-making skills.

Monte Carlo simulations are a specific type of computational model. Here, the focus is on understanding the characteristics of a statistic (e.g., bias, efficiency, consistency), often under real-world situations or conditions where the assumptions of the statistic are violated. For example, research has examined how various fit indices are affected by model misspecification, for example [5]. Likewise, we may want to understand the accuracy of formula-based cross-validity estimates when there is preselection on the predictors [12]. There have been a wealth of studies examining such issues, and much of what we know about the characteristics of statistics is based on large-scale Monte Carlo studies.

An increasingly common computational model is a hybrid between statistical and substantive questions. This type of model is frequently used when we want to understand the characteristics of a statistic to answer important applied questions. For example, we may want to know the consequences of error variance heterogeneity on tests of differences in slopes between demographic subgroups [1]. LeBreton, Ployhart, and Ladd [8] examined which type of predictor relative importance estimate is most effective in determining which predictors to keep in a regression model. Sackett and Roth [11] demonstrated the effects on adverse impact when combining predictors with differing degrees of intercorrelations and subgroup differences. Murphy [9] documented the negative effect on test utility when the top applicant rejects a job offer. In each of these examples, a merging of substantive and statistical questions has led to a simulation methodology that informs research and practice. The power of computational modeling in such circumstances helps test theories and develop applied solutions without the difficulty, expensive, and frequent impossibility of collecting real-world data.

References

- [1] Alexander, R.A. & DeShon, R.P. (1994). The effect of error variance heterogeneity on the power of tests for regression slope differences, *Psychological Bulletin* **115**, 308–314.
- [2] Anderson, J.R. (1993). Problem solving and learning, *American Psychologist* **48**, 35–44.
- [3] Anderson, J.R. (1996). ACT: a simple theory of complex cognition, *American Psychologist* **51**, 355–365.
- [4] Hanisch, K.A., Hulin, C.L. & Seitz, S.T. (1996). Mathematical/computational modeling of organizational withdrawal processes: benefits, methods, and results, in *Research in Personnel and Human Resources Management*, Vol. 14, G. Ferris, ed., JAI Press, Greenwich, pp. 91–142.
- [5] Hu, L.T. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification, *Psychological Methods* **3**, 424–453.
- [6] Hulin, C.L., Miner, A.G. & Seitz, S.T. (2002). Computational modeling in organizational sciences: contribution of a third research discipline, in *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis*, F. Drasgow & N. Schmitt, eds, Jossey-Bass, San Francisco, pp. 498–533.
- [7] Ilgen, D.R. & Hulin, C.L. (2000). *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline*, American Psychological Association, Washington.
- [8] LeBreton, J.M., Ployhart, R.E. & Ladd, R.T. (2004). Use of dominance analysis to assess relative importance: a Monte Carlo comparison with alternative methods, *Organizational Research Methods* **7**, 258–282.
- [9] Murphy, K.R. (1986). When your top choice turns you down: effect of rejected offers on the utility of selection tests, *Psychological Bulletin* **99**, 133–138.
- [10] Ployhart, R.E. & Ehrhart, M.G. (2002). Modeling the practical effects of applicant reactions: subgroup differences in test-taking motivation, test performance, and selection rates, *International Journal of Selection and Assessment* **10**, 258–270.
- [11] Sackett, P.R. & Roth, L. (1996). Multi-stage selection strategies: a Monte Carlo investigation of effects on performance and minority hiring, *Personnel Psychology* **49**, 1–18.
- [12] Schmitt, N. & Ployhart, R.E. (1999). Estimates of cross-validity for stepwise-regression and with predictor selection, *Journal of Applied Psychology* **84**, 50–57.
- [13] Whicker, M.L. & Sigelman, L. (1991). *Computer Simulation Applications: An Introduction*, Sage Publications, Newbury Park.
- [14] Zickar, M.J. & Robie, C. (1999). Modeling faking at the item-level, *Journal of Applied Psychology* **84**, 95–108.
- [15] Zickar, M.J. & Slaughter, J.E. (2002). Computational modeling, in *Handbook of Research Methods in Industrial and Organizational Psychology*, S.G. Rogelberg, ed., Blackwell Publishers, Walden, pp. 184–197.

ROBERT E. PLOYHART AND CRYSTAL
M. HAROLD

Computer-Adaptive Testing

RICHARD M. LUECHT

Volume 1, pp. 343–350

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Computer-Adaptive Testing

The phrase *computer-adaptive testing* (CAT) describes a wide range of tests administered on a computer, where the test difficulty is specifically targeted to match the proficiency of each examinee or where items are selected to provide an accurate pass/fail or mastery decision using as few items as possible. Applications of CAT can be found for achievement or performance assessments, aptitude tests, certification tests, and licensure tests. Unlike conventional, standardized, paper-and-pencil tests, where all examinees receive a common test form or test battery comprising the same items, in CAT, each examinee may receive a different assortment of items selected according to various rules, goals, or procedures.

There are numerous examples of successful, large-scale CAT testing programs such as the ACCU-PLACER postsecondary placement exams, operated by the College Board [5], the Graduate Record Exam [6], the Armed Service Vocational Aptitude Battery [16], and several licensure or certification tests such as the Novell certification exams and the licensure exam for registered nurses [27].

This entry provides an overview of CAT. Some suggested further readings on CAT include [25], [23], [9]; [7], and [13].

The Basic CAT Algorithm

A CAT algorithm is an iterative process designed to tailor an examination to an individual examinee's ability. CAT employs an item database that typically contains three types of information about each item: (a) the item text and other rendering data that is used by the testing software to present the item to the test taker and capture his or her response; (b) the answer key; and (c) **item response theory** (IRT) item parameter estimates. The IRT item parameter estimates are frequently based on the one-, two-, or three-parameter IRT model (e.g., see [7] and [9]) and must be calibrated to common scale. For example, if the three-parameter model is used, the item database will contain a discrimination parameter estimate, a_i , a difficulty parameter estimate, b_i , and a pseudo-guessing parameter, c_i , for $i = 1, \dots, I$ items in

the database. The database may also contain item exposure control parameters that are used to restrict the overexposure of the best items, as well as various content and other coded item attributes that may be used by the CAT item selection algorithm.

CAT also requires test delivery software – sometimes called the *test driver* – that provides three basic functions. First, the software must render each item on-screen and store the test taker's response(s). Second, the software must be capable of computing provisional IRT proficiency scores for the examinees. These can be **maximum likelihood** scores, Bayes mean scores, or Bayes modal scores (see, for example, references [12] and [21]). IRT scores are computed using the examinee's responses to a series of items and the associated IRT item parameter estimates. For example, the Bayes modal scores [12] can be expressed as

$$\hat{\theta}_{u_{i_1} \dots u_{i_{k-1}}}^{MAP} \equiv \max_{\theta} \{g(\theta|u_{i_1}, \dots, u_{i_{k-1}}) : \theta \in (-\infty, \infty)\} \quad (1)$$

where the value of θ at maximum of the posterior likelihood function, $g(\theta|u_{i_1}, \dots, u_{i_{k-1}})$, is the model estimate of θ . Finally, the software must select the remaining items for each examinee using an adaptive algorithm to maximize the IRT item information at the provisional ability estimate, $\hat{\theta}_{u_{i_1} \dots u_{i_{k-1}}}$. That is, given the unselected items in an item database, R_k , an item is selected to satisfy the function

$$i_k \equiv \max_j \{I_{U_j}(\hat{\theta}_{u_{i_1}, \dots, u_{i_{k-1}}}) : j \in R_k\}. \quad (2)$$

This item selection mechanism is described in more depth in the next section.

A CAT actually begins with a brief 'burn-in' period that usually involves the administration of a very small random sample of items. This sample of items is meant to provide a reasonable starting estimate of the examinee's ability, prior to activating the adaptive algorithm. Following the burn-in period, CAT resolves to an iterative three-step process, cycling through the above three functions. That is, the test delivery software: (a) selects the next item to maximize information at the provisional score; (b) renders the item and captures the examinee's response(s); and (c) updates the provisional score and returns to Step 1. As more items are administered, the CAT software is able to incrementally improve the accuracy of the provisional score for the examinee. The CAT terminates when a stopping rule has been

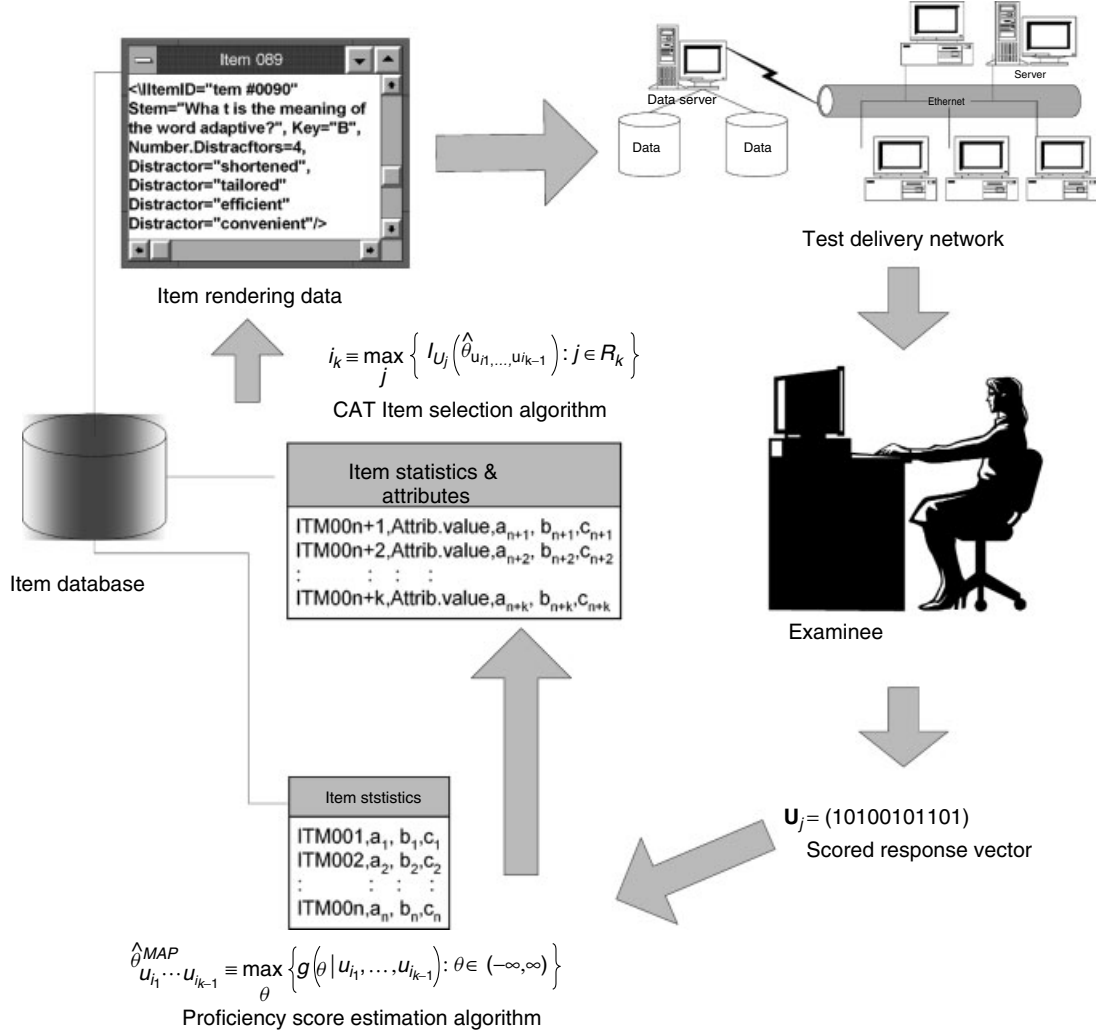


Figure 1 The iterative CAT process

satisfied. Two standard stopping rules for adaptive tests are: (a) a fixed test length has been met or (b) a minimum level of score precision has been satisfied¹. This iterative process of selecting and administering items, scoring, and then selecting more items is depicted in Figure 1.

In Figure 1, the initial items are usually transmitted through a computer network and rendered at the examinee's workstation. The responses are captured and scored by the test delivery software. The scored response vector and the item parameters are then used to update the provisional estimate of θ . That provisional score is

then used by maximum information algorithm, $i_k \equiv \max_j \{ I_{U_j}(\hat{\theta}_{u_1, \dots, u_{i_{k-1}}}) : j \in R_k \}$, to select the next item. The test terminates when either a fixed number of items have been administered or when a particular statistical criterion has been attained.

It is important to realize that each item incrementally improves our statistical confidence about an examinee's unknown proficiency, θ . For example, Figure 2 shows the degree of certainty about an examinee's score after 3 items and again after 50 items. For the sake of this example, assume that we know that this examinee's *true proficiency score* to be $\theta = 1.75$. After administering only 3

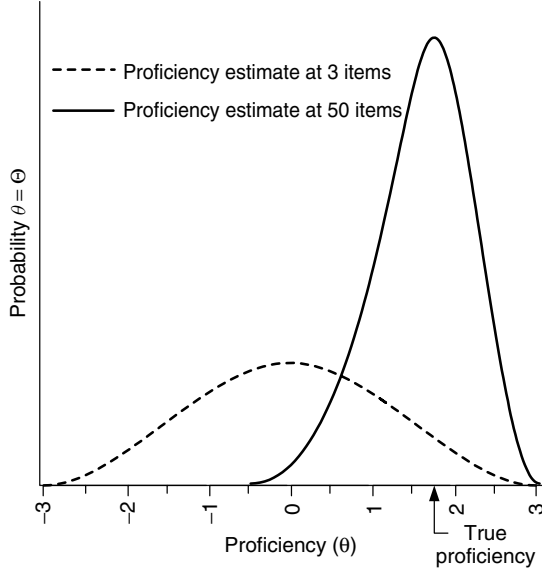


Figure 2 Certainty about proficiency after three and after fifty items

items, our certainty is represented by dotted curve is relatively flat indicating a lack of confidence about the exact location of the *provisional* estimate. However, after administering 50 items, we find that: (a) the *provisional* score estimate is quite close to the *true proficiency*; and (b) our certainty is very high, as indicated by the tall, narrow curve.

IRT Information and Efficiency in CAT

To better understand how the adaptive algorithm actually works, we need to focus on the IRT item and test information functions. Birnbaum [1] introduced the concept of the test information function as a psychometric analysis mechanism for designing and comparing the measurement precision of tests in the context of item response theory (IRT). Under IRT, the conditional measurement error variance, $\text{var}(E|\theta)$, is inversely proportional to the test information function, $I(\theta)$. That is,

$$\text{var}(E|\theta) = [I(\theta)]^{-1} = \frac{1}{\sum_{i=1}^n I_i(\theta)} \quad (3)$$

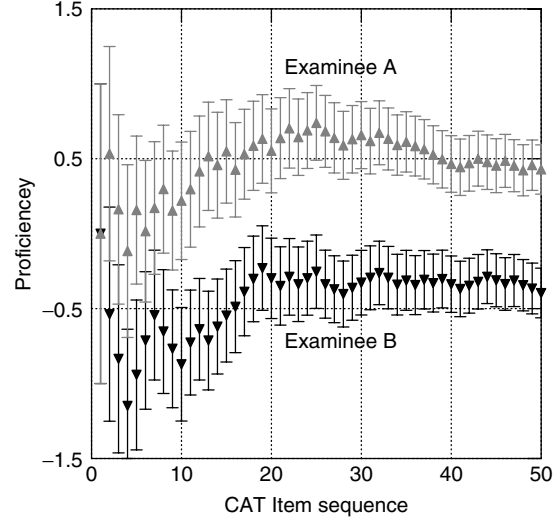


Figure 3 Proficiency scores and standard errors for a 50-item CAT for two hypothetical examinees

where $I_i(\theta)$ is the item information function at some proficiency score of interest, denoted as θ . The exact mathematical form of the information function varies by IRT model. Lord [9] and Hambleton and Swaminathan [7] provide convenient computational formulas for the one-, two-, and three-parameter IRT model information functions.

Equation (3) suggests two important aspects about measurement precision. First, each item contributes some amount of measurement information to the reliability or score precision of the total test. That is, the total test information function is sum of the item information functions. Second, by increasing the test information function, we correspondingly reduce the measurement error variance of the estimated θ score. Simply put, when test information is maximized, measurement errors are minimized.

Figure 3 shows what happens to the provisional proficiency scores and associated standard errors (the square root of the error variance from (3)) for two hypothetical examinees taking a 50-item CAT. The proficiency scale is shown on the vertical axis (-1.5 to $+1.5$). The sequence of 50 adaptively administered items is shown on the horizontal scale. Although not shown in the picture, initially, both examinees start with proficiency estimates near zero. After the first item is given, the estimated proficiency

scores immediately begin to separate ($\blacktriangle$ for Examinee A and $\blacktriangledown$ for Examinee B). Over the course of 50 items, the individual proficiency scores for these two examinees systematically diverge to their approximate true values of +1.0 for Examinee A and -1.0 for Examinee B. The difficulties of the 50 items selected for each examinee CAT would track in a pattern similar to the symbols plotted for the provisional proficiency scores. The plot also indicates the estimation errors present throughout the CAT. The size of each error band about the proficiency score denotes the relative amount of error associated with the scores. Larger bands indicate more error than narrower bands. Near to the left side of the plot the error bands are quite large, indicating fairly imprecise scores. During the first half of the CAT, the error bands rapidly shrink in size. After 20 items or so, the error bands tend to stabilize (i.e., still shrink, but more slowly). This example demonstrates how the CAT quickly reduces error variance and improves the efficiency of a test.

In practice, we can achieve maximum test information in two ways. We can choose highly discriminating items that provide maximum item information within particular regions of the proficiency scale or at specific proficiency scores – that is, we sequentially select items to satisfy (2). Or, we can merely continue adding items to increment the amount of information until a desired level of precision is achieved. Maximizing the test information at each examinee's score is tantamount to choosing a customized, optimally reliable test for each examinee.

A CAT achieves either improvements in relative efficiency or a reduction in test length. Relative efficiency refers to a proportional improvement in test information and can be computed as the ratio of test information functions or reciprocal error variances for two tests (see (3); also see [9]). This relative efficiency metric can be applied to improvements in the accuracy of proficiency scores or to decision accuracy in the context of mastery tests or certification/licensure tests. For example, if the average test information function for a fixed-item test is 10.0 and the average test information function for an adaptive test is 15.0, the adaptive test is said to be 150% as efficient as the fixed-item test. Measurement efficiency is also associated with reductions in test length. For example, if a 20-item adaptive test can provide the same precision as a 40-item nonadaptive test, there is an obvious reduction the amount of test materials

needed and less testing time needed (assuming, of course, that a shorter test ought to take substantially less time than a longer test). Much of the early adaptive testing research reported that typical fixed-length academic achievement tests used could be reduced by half by moving to a computerized adaptive test² [25]. However, that early research ignored the perceptions by some test users – especially in high-stakes testing circles – that short adaptive tests containing only 10 or 20 items could not adequately cover enough content to make valid decisions or uses of scores. Today, CAT designs typically avoid such criticism by using either fixed lengths or at least some minimum test length to ensure basic content coverage.

Nonetheless, CAT does offer improved testing efficiency, which means we can obtain more confident estimates of examinees' performance using fewer items than are typically required on nonadaptive tests. Figure 4 shows an example of the efficiency gains for a hypothetical CAT, compared to a test for which the items were randomly selected. The item characteristics used to generate the test results for Figure 4 are rather typical of most professionally developed achievement tests. The plot shows the *average* standard errors – the square root of the error variance from (3) – over the sequence of 50 items (horizontal axis). The standard errors are averaged for a sizable sample of examinees having different proficiency scores.

In Figure 4, we can more specifically see how the errors decrease over the course of the two tests. It is important to realize that the errors decrease for a randomly selected set of items, too. However, CAT clearly does a better job of more rapidly reducing the errors. For example, at 20 items, the CAT achieves nearly the same efficiency as the 50-item random test; at 50 items, the average standard error for the CAT is approximately half as large as for the random test.

Security Risks in CAT

The risks to the security of computer-based tests are somewhat analogous to the cheating threats faced by gambling casinos or lotteries. Given any type of high stakes (e.g., entrance into graduate school, scholarships, a coveted course placement, a job, a license, a professional certificate), there will be some group of cheaters intent on *beating-the-odds* (of random chance or luck) by employing well

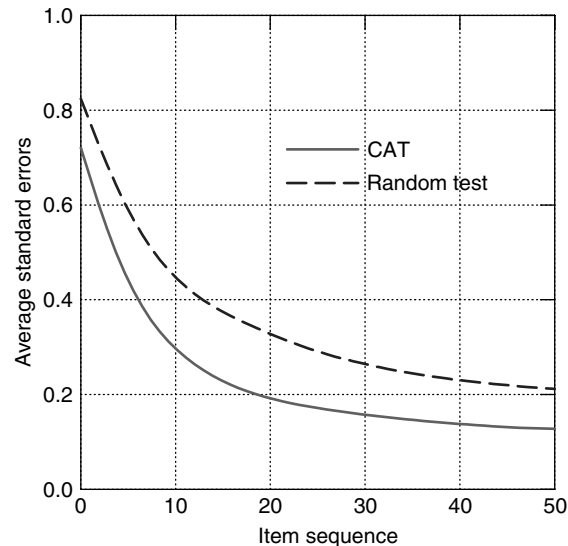


Figure 4 Average standard errors for a 50-item CAT versus 50 randomly selected items

thought out strategies, which provide them with any possible advantage, however slight that may be. One of the most common security risks in high-stakes CAT involves groups of examinees collaborating to memorize and share items, especially when the same item database is active over a long period of time, and testing is nearly continuous during that time period.

Unfortunately, the CAT algorithm actually exacerbates the security risks associated with cheating through systematic memorization of an item database. That is, because the CAT algorithm chooses the items to be maximally informative for each examinee, the most discriminating items are chosen far more often than the less discriminating items. This means that the *effective* item pool will typically be quite small since only a subset of the entire item database is being used. Beyond the bad economic policy of underutilizing an expensive commodity such as a large portion of an item database, cheaters gain the advantage of only needing to memorize and share the most highly exposed items.

Three of the methods for dealing with over-exposure risks in high-stakes CAT are: (a) increasing the size of the active item database; (b) rotating item databases over time (intact or partially); and (c) specifically controlling item exposures as part of the computerized test assembly process. The latter approach involves a modification to the CAT

item selection algorithm. Traditional exposure control modifications cited in the psychometric literature include maximum information item selection with the Simpson–Hetter unconditional item exposure control procedure (see references [8] and [20]), maximum information and Stocking and Lewis (conditional) item exposure control procedure (see [17, 18] and [19]), and maximum information and stochastic (conditional) exposure control procedure (see [14], [15]). An extensive discussion of exposure controls is beyond the scope of this entry.

CAT Variations

In recent years, CAT research has moved beyond the basic algorithm presented earlier in an attempt to generate better strategies for controlling test form quality control and simultaneously reducing exposure risks. Some testing programs are even moving away from the idea of an item as the optimal unit for CAT. Four promising CAT variations are: (a) constrained CAT using shadow tests (CCAT-UST); (b) α -Stratified Computerized Adaptive Testing (AS-CAT); (c) testlet-based CAT (T-CAT); and computer-adaptive multistage testing (CA-MST). These four approaches are summarized briefly, below.

Van der Linden and Reese [24] introduced the concept of a *shadow test* as a method of achieving an optimal CAT in the face of numerous content and other test assembly constraints (also see van der Linden [22]). Under CCAT-UST, a complete test is reassembled following each item administration. This test, called the *shadow test*, incorporates all of the required content constraints, item exposure rules, and other constraints (e.g., cognitive levels, total word counts, test timing requirements, clueing across items), and uses maximization of test information at the examinee's current proficiency estimate as its objective function. The shadow test model is an efficient means for balancing the goals of meeting content constraints and maximizing test information.

A shadow test actually is a special case of content-constrained CAT that explicitly uses automated test assembly (ATA) algorithms for each adaptive item selection. In that regard, this model blends the efficiency of CAT with the sophistication of using powerful linear programming techniques (or other ATA heuristics) to ensure a psychometrically optimal test that simultaneously meets any number of test-level specifications and item attribute constraints. Shadow testing can further incorporate exposure control mechanisms as a security measure to combat some types of cheating [22].

a-Stratified computerized adaptive testing (AS-CAT; [4]) is an interesting modification on the adaptive theme. AS-CAT adapts the test to the examinee's proficiency – like a traditional CAT. However, the AS-CAT model eliminates the need for formal exposure controls and makes use of a greater proportion of the test bank than traditional CAT. As noted earlier, the issue of test bank use is extremely important from an economic perspective (see the section Security Risks in CAT). *a*-Stratified CAT partitions the test bank into ordered layers, based on statistical characteristics of the items (see [4], [3]). First, the items are sorted according to their estimated IRT item discrimination parameters³. Second, the sorted list is partitioned into layers (the strata) of a fixed size. Third, one or more items are selected within each strata by the usual CAT maximum information algorithm. AS-CAT then proceeds sequentially through the strata, from the least to the most discriminating strata. The item selections may or may not be subject to also meeting applicable content specifications or constraints. Chang and Ying reasoned that,

during the initial portion of an adaptive test, less discriminating items could be used since the proficiency estimates have not yet stabilized. This stratification strategy effectively ensures that most discriminating items are saved until later in the test when they can be more accurately targeted to the provisional proficiency scores. In short, the AS-CAT approach avoids wasting the 'high demand' items too early on in the test and makes effective use of the low demand items that, ordinarily, are seldom if ever selected in CAT. Chang, Qian, and Ying [2] went a step further to also block the items based on the IRT difficulty parameters. This modification is intended to deal more effectively with exposure risks when the IRT discrimination and difficulty parameters are correlated with each other within a particular item pool.

One of the principal complaints from examinees about CAT is the inability for them to skip items, or review and change their answers to previously seen items. That is, because the particular sequence of item selections in CAT is dependent on the provisional scores, item review is usually prohibited. To address this shortcoming in CAT, Wainer and Kiely [26] introduced the concept of a *testlet* to describe a subset of items or a 'mini-test' that could be used in an adaptive testing environment. A testlet-based CAT (TB-CAT) involves the adaptive administration of preassembled *sets* of items to an examinee, rather than single items. Examples of testlets include sets of items that are associated with a common reading passage or visual stimulus, or a carefully constructed subset of items that mirrors the overall content specifications for a test. After completing the testlet, the computer scores the items within it and then chooses the next testlet to be administered. Thus, this type of test is adaptive at the testlet level rather than at the item level. This approach allows examinees to skip, review, and change answers within a block of test items. It also allows for content and measurement review of these sets of items prior to operational administration.

It should be clear that testlet-based CATs are only partially adaptive since items within a testlet are administered in a linear fashion. However, TB-CAT offers a compromise between the traditional, nonadaptive format and the purely adaptive model. Advantages of TB-CAT include increased testing efficiency relative to nonadaptive tests; the ability of content experts and sensitivity reviewers to review individual, preconstructed testlets and subtests to

evaluate content quality; and the ability of examinees to skip, review, and change answers to questions within a testlet.

Similar in concept to TB-CAT is computer-adaptive multistage testing (CA-MST). Luecht and Nungester [11] introduced CA-MST under the heading of ‘computer-adaptive sequential testing’ as a framework for managing real-life test construction requirements for large-scale CBT applications (also see [10]). Functionally, CA-MST is a preconstructed, self-administering, multistage adaptive test model that employs testlets as the unit of selection. The primary difference between TB-CAT and CA-MST is that the latter prepackages the testlets, scoring tables, and routing rules for the test delivery software. It is even possible to use number-correct scoring during the real-time administration, eliminating the need for the test delivery software to have to compute IRT-based scoring or select testlets based on a maximum information criterion.

Like TB-CAT, CA-MST uses preconstructed testlets as the fundamental building blocks for test construction and test delivery. Testlets may range in size from several items to well over 100 items. The testlets are usually targeted to have specific statistical properties (e.g., a particular average item difficulty or to match a prescribed IRT information function) and all content balancing is built into the construction of the testlets. As part of the ATA process, the preconstructed testlets will be further prepackaged in small collections called ‘panels’. Each panel contains four to seven (or more) testlets, depending on the panel design chosen – an issue addressed below. Each testlet is explicitly assigned to a particular stage and to a specific route within the panel (easier, moderate, or harder) based upon the average difficulty of the testlet. Multiple panels can be prepared with item overlap precisely controlled across different panels.

CA-MST is adaptive in nature and is therefore more efficient than using fixed test forms. Yet, CA-MST provides explicit control over content validity, test form quality, and the exposure of test materials.

Notes

1. For pass/fail mastery tests that are typically used in certification and licensure testing, a different stopping rule can be implemented related to the desired statistical confidence in the accuracy of the classification decision(s).

2. Although adaptation is clearly important as a psychometric criterion, it is easy sometimes to overstate the real cost-reduction benefits that can be specifically attributed to gains in measurement efficiency. For example, measurement efficiency gains from adaptive testing are often equated with reduced testing time. However, any potential savings in testing time may prove to be unimportant if a computer-based examination is administered at commercial CBT centers. That is, commercial CBT centers typically charge fixed hourly rates per examinee and require a guaranteed [minimum] testing time. Therefore, if the CBT test center vendor negotiates with the test developer for a four-hour test, the same fee may be charged whether the examinee is at the center for two, three, or four hours.
3. See [7] or [9] for a more detailed description of IRT item parameters for multiple-choice questions and related objective response items.

References

- [1] Birnbaum, A. (1968). Estimation of an ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 423–479.
- [2] Chang, H.H., Qian, J. & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing item *b*-blocking, *Applied Psychological Measurement* **25**, 333–342.
- [3] Chang, H.H. & van der Linden, W.J. (2000). A zero-one programming model for optimal stratification of item pools in *a*-stratified computerized adaptive testing, *Paper Presented at The Annual Meeting of the National Council on Measurement in Education*, New Orleans.
- [4] Chang, H.H. & Ying, Z. (1999). *A*-stratified multi-stage computerized adaptive testing, *Applied Psychological Measurement* **23**, 211–222.
- [5] College Board (1993). *Accuplacer: Computerized Placement Tests: Technical Data Supplement*, Author, New York.
- [6] Eignor, D.R., Stocking, M.L., Way, W.D. & Steffen, M. (1993). *Case Studies in Computer Adaptive Test Design Through Simulation (RR-93-56)*, Educational Testing Service, Princeton.
- [7] Hambleton, R.K. & Swaminathan, H.R. (1985). *Item Response Theory: Principles and Applications*, Kluwer Academic Publishers, Boston.
- [8] Hetter, R.D. & Simpson, J.B. (1997). Item exposure control in CAT-ASVAB, in *Computerized Adaptive Testing: From Inquiry to Operation*, W.A. Sands, B.K. Waters & J.R. McBride, eds, American Psychological Association, Washington, pp. 141–144.
- [9] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.

- [10] Luecht, R.M. (2000). Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests, in *Paper Presented at The Meeting of The National Council on Measurement in Education*, New Orleans.
- [11] Luecht, R.M. & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing, *Journal of Educational Measurement* **35**, 229–249.
- [12] Mislevy, R.J. (1986). Bayesian modal estimation in item response models, *Psychometrika* **86**, 177–195.
- [13] Parshall, C.G., Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical Considerations in Computer-based Testing*, Springer, New York.
- [14] Revuela, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing, *Journal of Educational Measurement* **35**, 311–327.
- [15] Robin, F. (2001). *Development and evaluation of test assembly procedures for computerized adaptive testing*, Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- [16] Sands, W.A., Waters, B.K. & McBride, J.R., eds. (1997). *Computerized Adaptive Testing: From Inquiry to Operation*, American Psychological Association, Washington.
- [17] Stocking, M.L. & Lewis, C. (1995). A new method for controlling item exposure in computerized adaptive testing, Research Report No. 95-25, Educational Testing Service, Princeton.
- [18] Stocking, M.L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, *Journal of Educational and Behavioral Statistics* **23**, 57–75.
- [19] Stocking, M.L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT, in *Computerized Adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer Academic Publishers, Boston, pp. 163–182.
- [20] Simpson, J.B. & Hetter, R.D. (1985). Controlling item exposure rates in computerized adaptive tests, in *Paper presented at the Annual Conference of the Military Testing Association*, Military Testing Association, San Diego.
- [21] Thissen, D. & Orlando, M. (2002). Item response theory for items scored in two categories, in *Test Scoring*, D. Thissen & H. Wainer, eds, Lawrence Erlbaum, Mahwah, pp. 73–140.
- [22] van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests, in *Computer-adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer Academic Publishers, Boston, pp. 27–52.
- [23] van der Linden, W.J. & Glas, C.A.W., eds (2000). *Computer-adaptive Testing: Theory and Practice*, Kluwer Academic Publishers, Boston.
- [24] van der Linden, W.J. & Reese, L.M. (1998). A model for optimal constrained adaptive testing, *Applied Psychological Measurement* **22**, 259–270.
- [25] Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format, *Educational Measurement: Issues and Practice* **12**, 15–20.
- [26] Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement* **24**, 185–201.
- [27] Zara, A.R. (1994). An overview of the NCLEX/CAT beta test, in *Paper Presented at the Meeting of the American Educational Research Association*, New Orleans.

(See also **Structural Equation Modeling: Mixture Models**)

RICHARD M. LUECHT

Computer-based Test Designs

APRIL L. ZENISKY

Volume 1, pp. 350–354

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Computer-based Test Designs

Across disciplines and contexts, more and more tests are being implemented as computerized assessments, with examinees viewing and responding to test questions via a desktop PC and, in some cases, also via the Internet. This transition in the administration mode is happening the world over, and many of the benefits of computerization are well documented in the growing computer-based testing (CBT) literature (see edited books by Van der Linden and Glas [20] and Mills et al. [14] for excellent overviews of the state of the art). Briefly, some of the measurement advantages to be realized in operational CBT concern how a computer-based test is implemented [3], and some possible sources of variation include the choice of item type, scoring method, the relative inclusion of multimedia and other technological innovations in the test administration, the procedures for item and item bank development, and test designs. This last issue of test designs, sometimes discussed as test models, refers to structural variations in test administration and how a computerized test will be implemented with respect to whether the specific items presented to an examinee are selected and grouped beforehand or are chosen during the test administration (*see Test Construction: Automated*).

Whereas paper-based tests are by and large static and form-based for very sensible logistical reasons, in computer-based testing (CBT), test developers have the power and memory of the computer at their disposal to be more or less variable during the course of the actual test administration. Such flexibility has helped to make the topic of test designs a significant area for research among test developers, particularly given the clear evidence in the psychometric literature for improved measurement under adaptive test designs in CBT [4].

The possibilities that (a) tests need not be exactly identical in sequence or test length and that (b) alternative designs could be implemented can be traced back to early work on intelligence testing, which was carried out very early in the twentieth century [1]. In these pioneering paper-based tests, both starting and termination points varied across students and were dependent on the responses provided by individual examinees. From that work and later studies by many

researchers including Lord [9, 10], the notion of tailoring tests to individual examinees was further developed, and today the continuum of test designs used in practice with CBT ranges from linear fixed-form tests assembled well in advance of the test administration to tests that are adaptive by item or by sets of items to be targeted at the estimated ability of each examinee individually. Each of these designs possess a variety of benefits and drawbacks for different testing constructs, and making the choice among such designs involves considerable thought and research on the part of a credentialing testing organization about the nature of the construct, the level of measurement precision necessary, and the examinee population.

It is generally established in the measurement literature that there are three families of available test designs for CBT. One of these is not adaptive (the linear fixed-form test design) and the other two are adaptive (multistage test designs and computer-adaptive test designs). Thus, a primary distinction among test designs that can be made concerns the property of being adaptive or not, and the further distinction is whether the test is adaptive at the item level or between sets of items. Traditionally, linear forms have predominated operational testing (both paper-and-pencil and computer-based). However, the advances in research into item response theory (IRT) (*see Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data*) over the years [7, 8, 12] as well as the advent of powerful and inexpensive desktop computers have facilitated implementation of adaptive test models. Such methods are described as adaptive in the sense that the sequence of items or sets of items administered to an individual examinee is dependent on the previous responses provided by the examinee [11].

Computerized Fixed Tests

The first test design, the nonadaptive linear fixed-form test, has been widely implemented in both paper-and-pencil tests and CBTs. In a CBT context, the linear fixed-form test is sometimes referred to as a computerized fixed test, or CFT [16]. CFT involves the case where a fixed set of items is selected to comprise a test form, and multiple parallel test forms may be created to maintain test security and to ensure ample usage of the item bank. In this approach, test

2 Computer-based Test Designs

forms may be constructed well in advance of actual test administration or assembled as the examinee is taking the test. This latter circumstance, commonly referred to as linear-on-the-fly testing, or LOFT, is a special case of CFT that uses item selection algorithms, which do not base item selection on estimated examinee ability; rather, selection of items proceeds relative to other predefined content and other statistical targets [2]. Each examinee receives a unique test form under the LOFT design, but this provides benefits in terms of item security rather than psychometric efficiency [4]. Making parallel forms or introducing some randomization of items across forms are additional methods by which test developers address item exposure and test security concerns in CFT.

Patelis [17] identified some other advantages associated with CFT including (a) the opportunity for examinees to review, revise, and omit items, and (b) the perception that such tests are easier to explain to examinees. At the same time, there are some disadvantages to linear test forms, and these are similar to those arising with paper-based tests. With static forms, each form may be constructed to reflect a range of item difficulty in order to accurately assess examinees of different abilities. Consequently, the scores for some examinees (and especially those at the higher and lower ability levels) may not be as precise as it would be in a targeted test.

The linear test designs possess many benefits for measurement, and depending on the purpose of testing and the degree of measurement precision needed they may be wholly appropriate for many large-scale testing organizations. However, other agencies may be more interested in other test designs that afford them different advantages, such as the use of shorter tests and the capacity to obtain more precise measurement all along the ability distribution and particularly near the cut-score where pass-fail decisions are made in order to classify examinees as masters or nonmasters. The remaining two families of test designs are considered to be adaptive in nature, though they do differ somewhat with respect to structure and format.

Multistage Tests

The second family of test designs, multistage testing (MST), is often viewed as an intermediary step between a linear test and a computer-adaptive test (CAT). As a middle ground, MST combines the

adaptive features of CAT with the opportunity to preassemble portions of tests prior to administration as is done with linear testing [6]. MST designs are generally defined by using multiple sets of items that vary on the basis of difficulty and routing examinees through a sequence of such sets on the basis of the performance on previous sets. With sets varying by difficulty, the particular sequence of item sets that any one examinee is presented with as the test is administered is chosen based on an examinee's estimated ability, and so the test 'form' is likely to differ for examinees of different ability levels. After an examinee finishes each item set, that ability estimate is updated to reflect the new measurement information obtained about that examinee's ability through administration of the item set. In MST terminology, these sets of items have come to be described as *modules* [13] or *testlets* [21], and can be characterized as short versions of linear test forms, where some specified number of individual items are administered together to meet particular test specifications and provide a certain proportion of the total test information. The individual items in a module may be all related to one or more common stems (such as passages or graphics) or be more generally discrete from one another, per the content specifications of the testing program for the test in question. These self-contained, carefully constructed, fixed sets of items are the same for every examinee to whom each set is administered, but any two examinees may or may not be presented with the same sequence of modules. Most of the common MST designs use two or three stages. However, the actual number of stages that could be implemented could be set higher (or lower) given the needs of different testing programs.

As a test design, MST possesses a number of desirable characteristics. Examinees may change answers or skip test items and return to them, prior to actually finishing a module and moving on to another. Of course, after completing a stage in MST, however, the items within that stage are usually scored using an appropriate IRT model and the next stage is selected adaptively, so no return to previous stages can be allowed (though, again, item review within a module at each stage is permissible). Measurement precision may be gained over CFT or LOFT designs without an increase in test length by adapting the exam administration to the performance levels of the examinees [11, 18]. If optimal precision of individual

proficiency estimates is desired, however, empirical studies have similarly demonstrated that scores obtained from MST are not quite as statistically accurate as those from CAT [18].

Computerized-adaptive Tests

In some ways, the third family of test designs, CAT, can be viewed as a special case of the MST model to the extent that CAT can be thought of as an MST made up of n stages with just one item per stage. In both cases, the fundamental principle is to target test administration to the estimated ability of the individual. There are differences, of course: as item selection in CAT is directly dependent on the responses an examinee provides to each item singly, no partial assembly of test forms or stages takes place for a computerized-adaptive test prior to test administration. Furthermore, given that CAT is adaptive at the item level, Lord [11] and Green [5] indicate that this test design provides the most optimal estimation of examinee proficiency all along the ability continuum relative to other test designs. Indeed, CAT has been widely implemented in a variety of testing contexts where precision all along the ability scale is desired, such as admissions testing.

In CAT, if an examinee early on in a test exhibits high ability, that person need not be presented with many items of low difficulty, and conversely, a low-ability examinee would not receive many very hard items. With such efficiency, test length may also be reduced. Other advantages associated with adaptive testing include enhanced test security, testing on demand, individualized pacing of test administration, immediate scoring and reporting of results, and easier maintenance of the item bank [8].

At the same time, CAT is administratively more complex, involves a changed approach to test development and score reporting, which is something of a departure from the procedures used in paper-and-pencil testing, and presents its own security concerns per [15]. Also, an oft-cited shortcoming of CAT from the perspective of examinees is the issue of item review as discussed by [19]. Whereas in traditional paper-based administration, examinees can go back and change answers as they see fit, this is not an option in most implementations of CAT because of the nature of the adaptive algorithm.

Conclusions

The computerization of assessment has facilitated immense flexibility, and the families of designs presented here characterize the range of available options. Each type of test design clearly presents certain benefits for different testing contexts, and the measurement advantages associated with the choice of design for any one situation must be weighed against operational realities. By knowing the alternatives and the properties of each, test developers can use this information to produce tests that are both maximally informative and psychometrically sound given the purpose for testing and the kinds of decisions to be made on the basis of test scores.

References

- [1] Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux~ [New methods for the diagnosis of the intellectual level of abnormals], *LiAnnée Psychologique* **11**, 191–245.
- [2] Carey, P.A. (1999, April). The use of linear-on-the-fly testing for TOEFL reading, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, Montreal.
- [3] Drasgow, F. & Olson-Buchanan, J.B. Eds. (1999). *Innovations in Computerized Assessment*, Erlbaum, Mahwah.
- [4] Folk, V.G. & Smith, R.L. (2002). Models for delivery of computer-based tests, in *Computer-based Testing: Building the Foundation for Future Assessments*, C.N. Mills, M.T. Potenza, J.J. Fremer & W.C. Ward Eds., Lawrence Erlbaum Associates, Mahwah, pp. 41–66.
- [5] Green Jr, B.F. (1983). The promise of tailored tests, in *Principals of Modern Psychological Measurement*, H. Wainer & S. Messick Eds., Lawrence Erlbaum Associates, Hillsdale, pp. 69–80.
- [6] Hambleton, R.K. (2002, April). Test design, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, New Orleans.
- [7] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer, Boston.
- [8] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.
- [9] Lord, F.M. (1970). Some test theory for tailored testing, in *Computer-assisted Instruction, Testing, and Guidance*, Holtzman W.H. Ed., Harper and Row, New York, pp. 139–183.
- [10] Lord, F.M. (1971). A theoretical study of two-stage testing, *Psychometrika* **36**, 227–242.
- [11] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale.

4 Computer-based Test Designs

- [12] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [13] Luecht, R.M. & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing, *Journal of Educational Measurement* **35**(3), 229–249.
- [14] Mills, C.N., Potenza, M.T., Fremer, J.J. & Ward, W.C. eds. (2002). *Computer-based Testing: Building the Foundation for Future Assessments*, Lawrence Erlbaum Associates, Mahwah.
- [15] Mills, C.N. & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing, *Applied Measurement in Education* **9**(4), 287–304.
- [16] Parshall, C.G., Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical Considerations in Computer-based Testing*, Springer, New York.
- [17] Patelis, T. (2000, April). *An Overview of Computer-based Testing (Office of Research and Development Research Notes, RN-09)*, College Board, New York.
- [18] Patsula, L.N. & Hambleton, R.K. (1999, April). A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, Montreal.
- [19] Stone, G.E. & Lunz, M.E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests, *Applied Measurement in Education* **7**, 211–222.
- [20] van der Linden, W.J. & Glas, C.A.W. eds. (2000). *Computerized Adaptive Testing: Theory and Practice*, Kluwer, Boston.
- [21] Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets, *Journal of Educational Measurement* **24**(3), 185–201.

Further Reading

- Wise, S.L. (1996, April). A critical analysis of the arguments for and against item review in computerized adaptive testing, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, New York.

APRIL L. ZENISKY

Computer-based Testing

TIM DAVEY

Volume 1, pp. 354–359

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Computer-based Testing

As most broadly defined, computer-based tests (CBTs) include not just tests administered on computers or workstations but also exams delivered via telephones, PDAs, and other electronic devices [1]. There have been three main reasons for test developers to move beyond conventional paper-and-pencil administration. The first is to change the nature of what is being measured [2, 10]. The second is to improve measurement precision or efficiency [18, 19]. The third is to make test administration more convenient for examinees, test sponsors, or both.

Changed Measurement

Standardized tests are often criticized as artificial and abstract, measuring performance in ways divorced from real-world behaviors [5]. At least some of this criticism is due to the constraints that paper-based administration imposes upon test developers. Paper is restricted to displaying static text and graphics, offers no real means of interacting with the examinee, and sharply limits the ways in which examinees can respond. Computers can free test developers from these restrictions. They can present sound and motion, interact dynamically with examinees, and accept responses through a variety of modes. For example,

- A CBT assessing language proficiency can measure not just how well an examinee can read and write but also their ability to comprehend spoken language, speak, and even converse [14].
- A CBT measuring proficiency with a software package can allow examinees to interact directly with that software to determine or even generate their responses.
- A selection test for prospective air traffic controllers can allow examinees to interact with and control simulated air traffic in a realistic environment [12].
- A science test can allow students to design and conduct simulated experiments as a means of responding.
- A medical certification exam can allow examinees to interactively evaluate, diagnose, treat, and manage simulated patients [3].

As these examples illustrate, a CBT can be a richer, more realistic experience that allows more direct measurement of the traits in question.

Improved Measurement Precision and Efficiency

An important type of CBT is termed *adaptive* because of the test's ability to tailor itself to uniquely suit an examinee's level of performance. As an *adaptive test* proceeds, answers to earlier questions determine which questions are asked later. The test, therefore, successively changes as the examinee's performance level is revealed [11, 18, 19].

At least three types of adaptive tests can be distinguished, but all consist of two basic steps: item selection and score estimation. Both are repeated each time an item (or set of items) is presented and answered. The first step determines the most appropriate item or set of items to administer given what is currently known about the examinee's performance level. Items or sets are selected from a *pool* containing many more items than any single examinee sees. Sets may comprise either items that share some natural connection to each other or items that are more arbitrarily linked. An example of the former is a reading passage to which several items are attached.

The second step uses the response or responses to the item or items previously presented to refine the score or performance estimate so that the next item or set presented can be more appropriate still. This cycle continues until either a specified number of items have been administered or some measure of score precision is reached. The process is represented schematically by Figure 1.

All of the adaptive testing strategies select items and assemble tests to best meet some or all of three

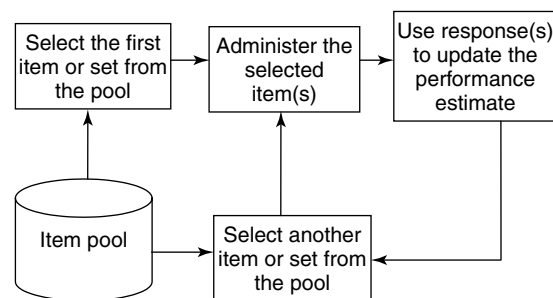


Figure 1 The adaptive testing process

goals, which usually conflict with one another [11]. The first is to maximize test efficiency by measuring examinees to appropriate levels of precision with as few items as possible. The competing adaptive testing strategies have evolved largely because different definitions can be attached to terms like ‘efficiency,’ ‘appropriate,’ and ‘precision.’ In any case, achieving this goal can allow an adaptive test to match or better the precision of a conventional test that is several times longer.

The second goal is that each examinee’s test be properly balanced in terms of item substance or content. This is important to ensure that tests are content valid and meet both examinees’ and score users’ subjective expectations of what a proper test should measure. The intent is to force adaptive tests to meet proper test-construction standards despite their being assembled ‘on-the-fly’ as the test proceeds [15].

A third goal is to control or balance the rates at which various items in the pool are administered [16]. The concern is that without such control, a small number of items might be administered very frequently while others rarely or never appear.

The potential conflicts between these goals are many. For example, imposing strict content standards is likely to lower test precision by forcing the selection of items with less optimal measurement properties. Protecting the administration rates of items with exceptional measurement properties will have a similar effect on precision. Every adaptive test must therefore strike a balance between these goals. The three basic testing strategies that will be described do so in fundamentally different ways.

CATs and MSTs

The first two types of adaptive tests to be described share a common definition of test precision. Both the *computerized adaptive test* (CAT) and the *multi-stage test* (MST) attempt to accurately and efficiently estimate each examinee’s location on a continuous performance or score scale. This goal will be distinguished below from that of *computerized classification tests*, which attempt to accurately assign each examinee to one of a small number of performance strata. Where CATs and MSTs differ is in the way this goal of maximum precision is balanced with the competing interests of controlling test content and the rates at which various items are administered.

In essence, the two strategies differ in the way and the extent to which the testing process is permitted to adapt.

The CAT selects items from the pool individually or in small sets. A wide range of item selection criteria have been proposed. Some of these operate from different definitions of precision; others try to recognize the difficulty inherent in making optimal decisions when information about examinee performance is incomplete and possibly misleading. Still others more explicitly subordinate measurement precision to the goals of balancing content and item exposure.

Test scoring methods also vary widely, although nearly all are based on **item response theory** (IRT) [8]. These procedures assume that all items in the pool are properly characterized and lie along a single IRT proficiency scale. A test score is then an estimate of the examinee’s standing on this same scale. **Maximum likelihood** and Bayesian estimation methods (see **Bayesian Item Response Theory Estimation**) are most commonly used. However, a number of more exotic estimation procedures have been proposed, largely with the goal of increased statistical robustness [17].

Because of the flexibility inherent in the process, the way in which a CAT unfolds is very difficult to predict. If the item pool is reasonably large (and most researchers recommend the pool contain at least 8–10 times more items than the test length), the particular combination of items administered to any examinee is virtually unique [9]. This variation has at least three sources. First, different items are most appropriate for different examinees along the proficiency scale. In general, easy items are most appropriate for low-scoring examinees while harder items are reserved for more proficient examinees. Second, each response an examinee makes can cause the test to move in a new direction. Correct answers generally lead to harder questions being subsequently selected while wrong answers lead to easier questions in the future. Finally, most item selection procedures incorporate a random element of some sort. This means that even examinees of similar proficiency who respond in similar ways are likely to see very different tests.

Although the changing and unpredictable nature of the CAT is the very essence of test adaptation, it can also be problematic. Some item selection procedures can ‘paint themselves into a corner’ and have no choice but to administer a test that fails to conform to all test-construction rules. Measurement precision

and overall test quality can also differ widely across examinees. Because tests are assembled in real time and uniquely for each examinee, there is obviously no opportunity for forms to be reviewed prior to administration. All of these concerns contributed to the development of multistage testing.

The MST is a very constrained version of CAT, with these constraints being imposed to make the testing process more systematic and predictable [7]. Development of an MST begins by assembling all of the available pool items into a relative handful of short tests, often called *testlets*, some of which target specific proficiency levels or ranges [20]. Content and item exposure rate considerations can be taken into account when assembling each testlet. A common practice is to assemble each testlet as a miniature version of an entire form.

Test administration usually begins by presenting each examinee with a testlet that measures across a wide proficiency range. The testlet is presented intact, with no further selection decision made until the examinee has completed all of its items. Once the initial testlet is completed, performance is evaluated and a selection decision is made. Examinees who performed well are assigned a second testlet that has been assembled to best measure higher proficiency ranges. Examinees who struggled are assigned a testlet largely comprising easier items. The logic inherent in these decisions is the same as that employed by the CAT, but selection decisions are made less frequently and the range of options for each decision is sharply reduced (since there are usually far fewer testlets available than there are items in a CAT pool). Scoring and routing decisions can be made based either on IRT methods similar to those used in CAT, or on conventional number-right scores. The former offers theoretical and psychometric advantages; the latter is far simpler operationally.

MSTs differ in the number of levels (or choice points) that each examinee is routed through. The number of proficiency-specific testlets available for selection at each level also differs. In simpler, more restrictive cases – those involving fewer levels and fewer testlets per level – it is quite possible to construct and review all of the test forms that could possibly be administered as combinations of the available elements. In all cases, the particular form administered via an MST is far more predictable than the outcome of a CAT. The price paid for increased predictability is a loss of flexibility and a decrease in

test efficiency or precision. However, this decrease can be relatively minor in some cases.

Computerized Classification Tests

Also called a *computerized mastery test*, the *computerized classification test* (CCT) is based on a very different premise [6, 13]. Rather than trying to position each examinee accurately on a proficiency scale, the CCT instead tries to accurately sort examinees into broad categories. The simplest example is a test that assigns each examinee to either of two classes. These classes may be labeled master versus nonmaster, pass versus fail or certified versus not certified. Classification is based around one or more *decision thresholds* positioned along the proficiency scale.

The CCT is an attractive alternative for the many testing applications that require only a broad grouping of examinees. Because it is far easier to determine whether an examinee is above or below a threshold than it is to position that examinee precisely along the continuous scale, a CCT can be even shorter and more efficient than a CAT. CCTs also lend themselves naturally to being of variable length across examinees. Examinees whose proficiency lies well above or below a decision threshold can be reliably classified with far fewer items than required by examinees who lie near that threshold.

CCTs are best conducted using either of three item selection and examinee classification methods. The first makes use of *latent class* IRT models, which assume a categorical rather than continuous underlying proficiency scale [4]. These models naturally score examinees through assignments to one of the latent classes or categories.

A second approach uses the *sequential probability ratio test* (SPRT), which conducts a series of likelihood ratio tests that lead ultimately to a classification decision [13]. The SPRT is ideally suited to tests that vary in length across examinees. Each time an item is administered and responded to, the procedure conducts a statistical test that can have either of two outcomes. The first is to conclude that an examinee can be classified with a stated level of confidence given the data collected so far. The second possible outcome is that classification cannot yet be confidently made and that testing will need to continue. The test ends either when a classification is made with confidence or some maximum number of items

4 Computer-based Testing

have been administered. In the latter case, the decision made at the end of the test may not reach the desired level of confidence. Although IRT is not necessary to administering a test under the SPRT, it can greatly increase test precision or efficiency.

The third strategy for test administration and scoring uses Bayesian decision theory to classify examinees [6]. Like the SPRT, Bayes methods offer control over classification precision in a variable length test. Testing can therefore continue until a desired level of confidence is reached. Bayes methods have an advantage over the SPRT in being more easily generalized to classification into more than two categories. They are also well supported by a rich framework of statistical theory.

Item selection under classification testing can be very different from that under CAT. It is best to select items that measure best at the classification thresholds rather than target examinee proficiency. Naturally, it is much easier to target a stationary threshold than it is to hit a constantly changing proficiency estimate. This is one factor in CCT's improved efficiency over CAT. However, the primary factor is that the CCT does not distinguish between examinees who are assigned the same classification. They are instead considered as having performed equally. In contrast, the CAT is burdened with making such distinctions, however small. The purpose of the test must, therefore, be considered when deciding whether the CAT or the CCT is the most appropriate strategy.

Operational Convenience

The third benefit of computerized testing is operational convenience for both examinees and test sponsors. These conveniences include:

Self-proctoring

Standardized paper and pencil tests often require a human proctor to distribute test booklets and answer sheets, keep track of time limits, and collect materials after the test ends. Administering a CBT can be as simple as parking an examinee in front of a computer. The computer can collect demographic data, orient the examinee to the testing process, administer and time the test, and produce a score report at the conclusion. Different examinees can sit side by side taking different tests with different time limits for

different purposes. With conventional administration, these two examinees would likely need to be tested at different times or in different places.

Mass Customization

A CBT can flexibly customize itself uniquely to each examinee. This can go well beyond the sort of adaptivity discussed above. For example, a CBT can choose and administer each examinee only the appropriate components of a large battery of tests. Another example would be a CBT that extended the testing of failing examinees in order to provide detailed diagnostic feedback useful for improving subsequent performance. A CBT also could select or adjust items based on examinee characteristics. For example, spelling and weight and measurement conventions can be easily matched to the location of the examinee.

Reach and Speed

Although a CBT can be administered in a site dedicated to test administration, it can also be delivered anywhere and anytime a computer is available. Examinees can test individually at home over the Internet or in large groups at a centralized testing site.

It is also possible to develop and distribute a CBT much faster than a paper test can be formatted, printed, boxed and shipped. This can allow tests to change rapidly in order to keep up with fast-changing curricula or subject matter.

Flexible Scheduling

Many CBT testing programs allow examinees to test when they choose rather than requiring them to select one of several periodic mass administrations. Allowing examinees to schedule their own test date can be more than just a convenience or an invitation to procrastinate. It can also provide real benefits and efficiencies. For example, examinees in a training program can move directly to certification and employment without waiting for some distant test administration date to arrive.

Immediate Scoring

Many CBTs are able to provide examinees with a score report immediately upon conclusion of the test.

This is particularly important when coupled with flexible scheduling. This can allow examinees to meet tight application deadlines, move directly to employment, or simply decide that their performance was substandard and register to retest.

Summary

Any description of computer-based testing is almost certain to be out-of-date before it appears in print. Things are changing quickly and will continue to do so. Over the last two decades, CBT has evolved from a largely experimental procedure under investigation to an operational procedure employed by hundreds of testing programs serving millions of examinees each year [2]. Even faster growth can be expected in the future as technology becomes a more permanent fixture of everyone's lives. Testing on computer may eventually become an even more natural behavior than testing on paper has ever been.

References

- [1] Bennett, R.E. (1998). *Reinventing Assessment*, Educational Testing Service, Princeton.
- [2] Bennett, R.E. (2002). Inexorable and inevitable: the continuing story of technology and assessment, *Journal of Technology, Learning, and Assessment* 1(1).
- [3] Clauser, B.E., Subhiyah, R.G., Nungester, R.J., Ripkey, D.R., Clyman, S.G. & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts, *Journal of Education Measurement* 32, 397–415.
- [4] Dayton, C.M. (1999). *Latent Class Scaling Analysis*, Sage, Newbury Park.
- [5] Gardner, H. (1991). Assessment in context: the alternative to standardized testing, in *Cognitive Approaches to Assessment*, B. Gifford & M.C. O'Connor, eds, Kluwer Academic Publishers, Boston.
- [6] Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement* 14, 367–386.
- [7] Lord, F.M. (1971). A theoretical study of two-stage testing, *Psychometrika* 36, 227–242.
- [8] Lord, F.M. (1980). *Applications of Item Response Theory to Testing Problems*, Lawrence Erlbaum, Hillsdale.
- [9] Mills, C. & Stocking, M. (1996). Practical issues in large scale computerized adaptive testing, *Applied Measurement in Education* 9(4), 287–304.
- [10] Parshall, C.G., Davey, T. & Pashley, P.J. (2000). Innovative item types for computerized testing, in *Computerized Adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer Academic Publishers, Norwell, pp. 129–148.
- [11] Parshall, C.G., Spray, J.A., Kalohn, J.C. & Davey, T. (2002). *Practical Considerations in Computer-based Testing*, Springer, New York.
- [12] Ramos, R.A., Heil, M.C. & Manning, C.A. (2001). *Documentation of Validity for the AT-SAT Computerized Test Battery* (DOT/FAA/AM-01/5), US Department of Transportation, Federal Aviation Administration, Washington.
- [13] Reckase, M.D. (1983). A procedure for decision making using tailored testing, in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, D.J. Weiss, ed., Academic Press, New York, pp. 237–255.
- [14] Rosenfeld, M., Leung, S. & Oltman, P.K. (2001). The reading, writing, speaking and listening tasks important for success at the undergraduate and graduate levels, TOEFL Monograph Series Report No. 21, Educational Testing Service, Princeton.
- [15] Stocking, M.L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing, *Applied Psychological Measurement* 17, 277–292.
- [16] Simpson, J.B. & Hetter, R.D. (1985). Controlling item exposure rates in computerized adaptive testing, *Proceedings of the 27th annual meeting of the Military Testing Association*, Naval Personnel Research and Development Center, San Diego, pp. 973–977.
- [17] Thissen, D. & Wainer, H., eds (2001). *Test Scoring*, Lawrence Erlbaum, Hillsdale.
- [18] Van der Linden, W.J. & Glas, C.A.W., eds (2000). *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers, Boston.
- [19] Wainer, H., ed. (1990). *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum, Hillsdale.
- [20] Wainer, H., Bradlow, E.T. & Du, Z. (2000). Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing, in *Computerized Adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer Academic Publishers, Boston, pp. 245–269.

TIM DAVEY

Concordance Rates

EDWIN J.C.G. VAN DEN OORD

Volume 1, pp. 359–359

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Concordance Rates

In genetic studies of medical conditions, the degree of similarity between twins is often calculated using concordance rates. These concordance rates are useful to study the relative importance of genes and environment. Monozygotic (MZ) twins share all of their genes, whereas dizygotic (DZ) twins share on an average half of their genes. Therefore, a higher concordance rate for MZ twins than for DZ twins suggests genetic influences.

Two concordance rates are commonly calculated: the pairwise and the probandwise concordance rate. In the pairwise concordance rate, each twin pair is counted as one unit. The concordance rate is then simply equal to the proportion of twin pairs in the total sample in which both twins have the disorder. The proband is the member of a twin pair who qualified the pair for inclusion in the study. It is possible for both members of a twin pair

to be independently selected. In the probandwise concordance rate, these pairs would be counted twice. For example, assume a total sample of 10 twin pairs. If there would be 6 twin pairs where both members of the pair would have the disease, the pairwise concordance rate would be $6/10 = 0.6$. Now assume that 2 of the 6 twin pairs with two affected twins were independently selected for the study. These pairs would be counted twice, and the 4 other pairs with a single proband once. This gives as the numerator of the probandwise concordance rate $2 \times 2 + 4 = 8$. The denominator equals the sum of the numerator plus the discordant twin pairs counted once. In our example, the pairwise probandwise concordance rate would therefore equal $8/(8 + 4) = 0.67$.

*(See also **Correlation Issues in Genetics Research; Heritability**)*

EDWIN J.C.G. VAN DEN OORD

Conditional Independence

NICHOLAS T. LONGFORD

Volume 1, pp. 359–361

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Conditional Independence

Stochastic independence is a principal assumption in many models. Although often restrictive, it is very convenient because it is a special case of association. Without assuming independence, the covariance or correlation (dependence) structure of the observations has to be specified. Conditional independence is a natural way of relaxing the assumption of independence. Regression models (*see Multiple Linear Regression*) can be regarded as models of conditional independence, especially when the values of the covariates are not specified in advance but are, together with the values of the outcomes, subject to the vagaries of the sampling and data-generation processes.

Variables that induce (by conditioning) independence of outcome variables need not be observed. For example, scores on the sections of a test are usually correlated. Although the subjects have strengths and weaknesses in the different domains of the test represented by its sections, those who obtain high scores on one section tend to obtain high scores also on other sections. However, the correlations among the scores are reduced substantially if considered over the subjects with a specific value of a background variable, such as past performance or the dominant latent ability (trait). Indeed, many models for educational scores are based on conditional independence of the responses given the subject's ability [3]. With continuous (normally distributed) outcomes, **factor analysis** [2], postulates models in which the outcomes are conditionally independent given a few **latent variables**.

According to [1], every probability (statement) is conditional, either explicitly or because of the context and assumptions under which it is formulated. By the same token, every distribution is conditional, and therefore every statement of independence, referring to distributions, is conditional, and has to be qualified by the details of conditioning. Two random variables that are conditionally independent given the value of a third variable need not be conditionally independent given the value of a fourth variable. When they are independent, additional conditioning does not upset their (conditional) independence.

Operating with conditional distributions is relatively easy, facilitated by the identity

$$f(x, y|z) = \frac{f(x, y, z)}{f(z)}, \quad (1)$$

where f is the generic notation for a density or probability of the (joint) distribution given by its arguments. Hence, conditional independence of X and Y , given Z , is equivalent to

$$f(x, y, z)f(z) = f(x, z)f(y, z). \quad (2)$$

However, one of the difficulties with conditional distributions is that their densities or probabilities often have forms very different from their parent (unconditional, or less conditional) distributions. The normal and multinomial distributions are important exceptions (*see Catalogue of Probability Density Functions*).

Suppose y and $\mathbf{x}$ are jointly normally distributed with mean vector $\mu = (\mu_y, \mu_x)$ and variance matrix

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{y,x} \\ \sigma_{x,y} & \Sigma_x \end{pmatrix}. \quad (3)$$

If the values of $\mathbf{x}$ are regarded as fixed the outcomes y are associated with a univariate random sample $(y_i; \mathbf{x}_i)$, $i = 1, \dots, n$. They appear to be dependent, as similar values of $\mathbf{x}$ are associated with similar values of y . However, after conditioning on $\mathbf{x}$,

$$(y|\mathbf{x}) \sim \mathcal{N}(\mu_y + \sigma_{y,x} \Sigma_x^{-1}(\mathbf{x} - \mu_x), \sigma_y^2 - \sigma_{y,x} \Sigma_x^{-1} \sigma_{x,y}), \quad (4)$$

the outcomes are *conditionally* independent given $\mathbf{x}$, although their expectations depend on $\mathbf{x}$. The property of the normal distribution that the conditional distribution of $(y|\mathbf{x})$ belongs to the same class as the distributions of y and $(y, \mathbf{x})$ without conditioning is not shared by other classes of distributions. Together with the closed form of the density of the normal distribution, and its closure with respect to addition, this makes the normal distribution the assumption of choice in many analyses.

Thus far, we were concerned with conditional independence in connection with observational units. A related area of statistical research is conditional independence of variables without such a reference. Well-established examples of this are Markov processes, (*see Markov Chains*) [4], and **time series** processes in general.

2 Conditional Independence

In Markov processes, a sequence of variables indexed by discrete or continuous time t , X_t , has the property that variables X_{t_1} and X_{t_3} are conditionally independent given an intermediate variable X_{t_2} , $t_1 < t_2 < t_3$. The concept of conditional independence is essential in the definitions of other time series models, such as autoregressive and moving-average and their combinations, because they involve (residual) contributions independent of the past.

Search for conditional independence is an important preoccupation in many applications, because it enables a simpler description or explanation of and better insight into the studied processes. In graphical models, [5], sets of random vectors are represented by graphs in which *vertices* (variables) V are connected (associated) by *edges* E . A graph is defined as $G = (V, E)$. An important convention supporting

the interpretation of such graphs is that two sets of variables, A and B , are conditionally independent given C , if and only if there is no connection between the vertices corresponding to A and B after all the vertices that involve a vertex from C are erased. A complex graph is much easier to study through sets of such conditionally independent and conditioning variables. The edges may be associated with arrows that represent the direction of causality.

A simple example is drawn in Figure 1. It represents the graph

$$(\{A, B, C, D, E\}, \{AC, AD, BC, CD, CE\}). \quad (5)$$

The circles A–E represent random variables, and lines are drawn between variables that are not conditionally independent. By removing all the lines that connect C with the other variables, B and E are not connected with A or D ; the random vectors and variables $\{A, D\}$, $\{B\}$, and $\{E\}$ are mutually conditionally independent given C .

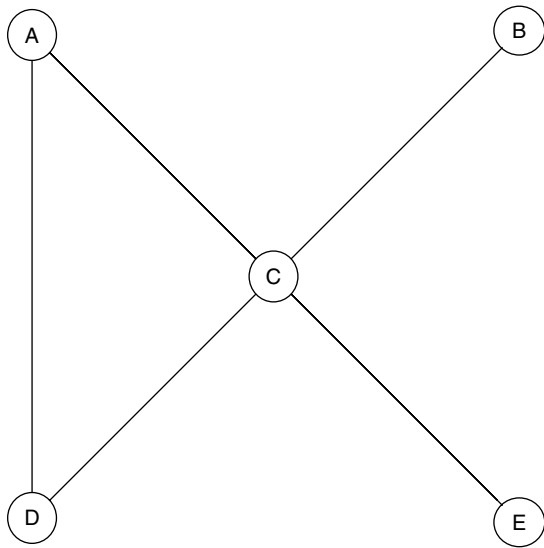


Figure 1 An example of a graphical model

References

- [1] de Finetti, B. (1974). *Theory of Probability: A Critical Introductory Treatment*, Vol. 1, Translated by A. Machí & A. Smith, John Wiley & Sons, Chichester.
- [2] Lawley, D.N. & Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd Edition, Butterworths, London.
- [3] Lord, F.M. (1980). *Application of Item Response Theory to Practical Problems*, Addison-Wesley, Reading.
- [4] Sharpe, M. (1988). *General Theory of Markov Processes*, Academic Press, New York.
- [5] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, New York.

NICHOLAS T. LONGFORD

Conditional Standard Errors of Measurement

YE TONG AND MICHAEL J. KOLEN

Volume 1, pp. 361–366

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Conditional Standard Errors of Measurement

Introduction

Errors of measurement for test scores generally are viewed as random and unpredictable. Measurement errors reduce the usefulness of test scores and limit the extent to which test results can be generalized. Conditional standard errors of measurement (conditional SEMs) index the amount of measurement error involved when measuring a particular person's proficiency using a test. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) recommends that conditional SEMs be reported by test publishers. Conditional SEMs can be used to construct a **confidence interval** on the score an individual earns and serve as one approach for reporting information about reliability.

Theoretically, conditional SEMs are conditioned on persons. For practical reasons, the reported conditional SEMs usually are conditioned on observed scores, because it is often impractical to report different conditional SEMs for people with the same observed test score. Many approaches for estimating conditional SEMs are based on three test theories: **classical test theory (CTT)**, **item response theory (IRT)**, and **generalizability theory (GT)**. These test theories make various assumptions about measurement errors. In this entry, conditional SEMs based on CTT and GT are considered in detail.

Classical Test Theory

Under classical test theory (Feldt & Brennan, [6]), the assumption is made that the observed score (X) is composed of a true score (T) and an error score (E) and $X = T + E$; test forms are *classically parallel*; correlation with an external criterion is the same for all parallel forms, and the expected error scores across test forms or populations of examinees are zero (i.e., $1/K \sum_{f=1}^K E_f = 0$, $1/N \sum_{p=1}^N E_p = 0$, with K being number of test forms and N number of examinees in the population).

For classically parallel forms the expected values and variances of the observed scores are assumed identical across forms. These assumptions are often

relaxed in practice. For *tau-equivalent forms*, the same expected scores are assumed across forms, but such forms are allowed to have different variances. For *essentially tau-equivalent forms*, expected scores across forms are allowed to differ by a constant and variances can differ across forms. For *congeneric forms*, expected scores are assumed to be linearly related, but true and error score variances can be different. For *classically congeneric forms*, error variance is assumed to follow classical theory, in addition to the congeneric forms assumptions (see **Tau-Equivalent and Congeneric Measurements**).

Based on these assumptions, a framework of estimating reliability and standard errors (both conditional and total) was developed. In the following discussion, different methods of estimating conditional SEMs using classical test theory and its extensions are presented.

Thorndike and Mollenkopf Approaches

Thorndike [19] proposed a method to estimate conditional SEMs, assuming tau-equivalent forms. Let X_1 be the observed score on the first half of a test form and X_2 the second half of the test form. The observed scores can be decomposed as

$$\begin{aligned} X_1 &= T + E_1 + c_1 \\ X_2 &= T + E_2 + c_2, \end{aligned} \quad (1)$$

where c_1 and c_2 are constants under tau-equivalent forms assumption. The theory is relatively straightforward: the variance of half-test difference ($\text{Var}(X_1 - X_2)$) equals the error variance of the total test ($\text{Var}(X_1 + X_2)$), because errors are assumed to be uncorrelated. That is,

$$\text{Var}(X_1 - X_2) = \text{Var}(X_1 + X_2). \quad (2)$$

By grouping individuals into a series of short intervals and estimating the total score and the variance of difference scores, the conditional SEMs for all intervals can be estimated.

For extreme intervals where not many examinees score, Thorndike's method can produce erratic estimates of conditional SEMs. Mollenkopf [18] proposed a regression technique to smooth out such irregularities and to make the estimates more stable. Still assuming tau-equivalent forms, consider the following quantity for a person:

$$Y_p = [(X_1 - X_2)_p - (\bar{X}_1 - \bar{X}_2)]^2, \quad (3)$$

2 Conditional Standard Errors of Measurement

when $\bar{X}_1$ is the mean of X_1 and $\bar{X}_2$ the mean of X_2 . The mean of this quantity estimates the variance of the half-test difference, which is an estimate for the total test error variance for the group. In the regression, Y is considered the dependent variable being predicted by $X_1 + X_2$ using polynomial regression. (see **Polynomial Model**) One potential complication with this method is to choose a certain degree of the polynomial. The lowest degree that fits the data is recommended in practice [6].

Lord's Binomial Error Model

Perhaps the best-known approach of calculating conditional SEMs was proposed by Lord [13, 14] based on the binomial error model. Under the binomial error model, each test form is regarded as a random set of n independent and dichotomously scored items. Each examinee is assumed to have a true proportion score (ϕ_p); the error for an individual p can therefore be defined as $X_p - n(\phi_p)$. Under such a conceptualization, the error variance conditional on a person over the population of test forms is

$$\sigma_{E.X_p}^2 = n(\phi_p)(1 - \phi_p). \quad (4)$$

Using the estimate of ϕ_p obtained through observed scores, $\hat{\phi}_p = X_p/n$, the estimate of the error variance for a person can be calculated:

$$\hat{\sigma}_{E.X_p}^2 = \frac{(n - X_p)X_p}{n - 1}. \quad (5)$$

The square root of this quantity yields the estimated conditional SEM for the person p .

By definition, persons with the same observed score on a given test have the same error variance under this model. One potential problem with this error model is that it fails to address the fact that test developers construct forms to be more similar to one another than would be expected if items were randomly sampled from a large pool of items. Therefore, this estimator typically produces overestimates of the conditional SEMs. Keats [8] proposed a correction factor for this binomial error model that proved to be quite effective.

Stratified Binomial Error

Feldt [5] modified Lord's binomial error model. His approach assumes that the test consists of a stratified

random sample of items where the strata are typically based on content classifications. Let $n_1, n_2, \dots$ be the number of items drawn from each stratum and $X_{p1}, X_{p2}, \dots$ be the observed score of person p on strata $1, 2, \dots$. In this case, the estimated total test error variance for person p is

$$\hat{\sigma}_{E.X_p}^2 = \sum_{h=1}^m \frac{(n_h - X_{ph})X_{ph}}{n_h - 1}. \quad (6)$$

Errors are assumed independent across strata and m is the total number of strata on the test.

When the number of items associated with any particular stratum is small, it can lead to instability in estimating the conditional SEMs using this method. Furthermore, for examinees with the same observed scores, the estimates of their conditional SEMs may be different, which might pose practical difficulties in score reporting.

Strong True Score Theory

Strong true score models can also be considered extensions of the classical test theory. In addition to assuming a binomial-related model for error score, typically a distributional form is assumed for true proportion-correct scores (ϕ). The most frequently used distribution form for true proportion scores is the beta distribution (see **Catalogue of Probability Density Functions**) whose random variable ranges from 0 to 1. The conditional error distribution $\Pr(X = i | \phi)$ is typically assumed to be binomial or compound binomial. Under these assumptions, the observed score then will follow a beta-binomial or compound beta-binomial distribution ([9], [15], and [16]).

Using Lord's two-term approximation of the compound binomial distribution, a general form of the error variance [15] can be estimated using

$$\hat{\sigma}_{E|x_p}^2 = \frac{x_p(n - x_p)}{n - 1} \times \left[1 - \frac{n(n - 1)S_{\bar{X}_i}^2}{\hat{\mu}_X(n - \hat{\mu}_X) - S_{X_p}^2 - nS_{\bar{X}_i}} \right]. \quad (7)$$

In 7, x_p is an observed score, $S_{\bar{X}_i}^2$ the observed variance of item difficulties and $S_{X_p}^2$ the observed variance of all examinees.

Generalizability Theory

Generalizability theory also assumes that observed test scores are composed of true score and error scores. A fundamental difference between generalizability theory and classical test theory is that under CTT, multiple sources of errors are confounded whereas under GT, errors are disentangled. GT typically involves two studies: G (*generalizability*) studies and D (*decision*) studies. The purpose of a G study is to obtain estimates of variance components associated with a universe of admissible observations. D studies emphasize the estimation, use, and interpretation of variance components for decision-making with well-specified measurement procedures [4]. **Linear models and Analysis Of Variance (ANOVA)** approaches are used to estimate errors and reliability.

Depending on the sources of errors in which the investigator is interested, different designs can be chosen for G and D studies. The simplest G study design is the person cross item ($p \times i$) design. For this design, all the examinees take the same set of items. Under the $p \times i$ design, an observed score is defined as $X_{pi} = \mu + v_p + v_i + v_{pi}$, where μ is the grand mean in the population and v_p , v_i , and v_{pi} are uncorrelated error effects related to the person, the item, and the person cross the item. The variances of these effects ($\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$) are called *variance components* and are of major concern in GT. Reliability and standard errors can be calculated from estimated variance components.

The corresponding D study is the $p \times I$ design. The linear decomposition of an examinee's average score, X_{pI} , over n' items, is $X_{pI} = \bar{X}_p = \mu + v_p + v_I + v_{pI}$, where the n' items for the D study are considered 'randomly parallel' of the n items in the G study. Under GT, by convention, average scores over a sample of conditions are indicated by uppercase letters. Other more complicated designs for both studies exist and are not further described in this entry. Brennan [4] gives a detailed description of different designs and the linear models related to each of the designs.

There are two kinds of errors under GT: *absolute error* and *relative error*. Both error variances can be defined in terms of the variance components obtained from a G study and the sample sizes used in a D study (n'). Absolute error refers to the difference between a person's observed score and universe score ($\Delta_{pI} =$

$X_{pI} - \mu_p$); relative error refers to the difference between a person's observed deviation score and universe deviation score ($\delta_{pI} = (X_{pI} - \mu_I) - (\mu_p - \mu)$). It can be shown that the absolute error is always larger or equal to the relative error, depending on the design of the D study. Accordingly, there are also two kinds of conditional SEMs in GT: Conditional absolute SEM and conditional relative SEM.

Brennan [4] argued that depending on the designs involved in GT, conditional SEMs are defined or estimated somewhat differently. Because of the complications associated with unbalanced designs, the conditional SEMs are only considered for the balanced design in this entry. A balanced design has no missing data, and the sample size is constant for each level of a nested facet.

Conditional Absolute Standard Error of Measurement

Conditional absolute SEM is straightforward with balanced designs: it is the standard error of the within-person mean. For a $p \times i$ design, absolute error for person p is $\Delta_p = X_{pI} - \mu_p$, the difference between the person's observed mean score across items (X_{pI}) and the mean score for the universe of items (μ_p). The associated error variance is:

$$\sigma^2(\Delta_p) = \text{var}(X_{pI} - \mu_p | p), \quad (8)$$

which is the variance of the mean over the number of items (n') for a D study for person p . An unbiased estimator is

$$\hat{\sigma}^2(\Delta_p) = \frac{\sum_i (X_{pi} - X_{pI})^2}{n'(n - 1)}, \quad (9)$$

the square root of which is an estimator of the conditional absolute SEM.

When all items are scored dichotomously and the number of items is the same for G and D studies (i.e., $n = n'$), the conditional absolute SEM is the same as Lord's conditional SEM. The extension of estimating conditional absolute SEMs for multifacet designs are illustrated in Brennan ([4], pp. 164–165).

Conditional Relative Standard Error of Measurement

Relative error for person p is $\delta_p = (X_{pI} - \mu_I) - (\mu_p - \mu)$, the difference between observed deviation

4 Conditional Standard Errors of Measurement

score and universe deviation score. Using results from Jarjoura [7], Brennan [3] showed that when the sample size for persons is large, an approximate estimator of the conditional relative SEM is

$$\hat{\sigma}(\delta_p) \doteq \sqrt{\hat{\sigma}^2(\Delta_p) + \frac{\hat{\sigma}^2(i)}{n'} - \frac{2\text{cov}(X_{pi}, X_{Pi|p})}{n'}}, \quad (10)$$

where $\text{cov}(X_{pi}, X_{Pi|p})$ is the observed covariance over items between examinee p 's item scores and the item mean scores. The observed covariance is not necessarily 0.

For multifacet designs, estimators exist for estimating the conditional relative SEMs, but they are rather complicated ([4], pp. 164–165). For practical use, the following formula often provides an adequate estimate:

$$\hat{\sigma}(\delta_p) = \sqrt{\hat{\sigma}^2(\Delta_p) - [\hat{\sigma}^2(\Delta) - \hat{\sigma}^2(\delta)]}. \quad (11)$$

Relationship Between Conditional SEMs Based on GT and CTT

Empirical investigation of the methods suggests that the estimates of conditional SEMs are fairly close to each other when these methods are applied to standardized achievement tests [2], [17]. These methods strongly support the conclusion that conditional SEMs vary as a curvilinear function of the observed score and the true score, on the raw score metric.

When scores are dichotomously scored and the sample sizes are the same for G and D studies, the conditional SEMs calculated using Lord's binomial error model yields the same result as the conditional absolute SEMs, even when the underlying test models differ. The conditional relative SEMs are harder to estimate [4], [12].

Score Scales

Number-correct raw scores on standardized tests typically are transformed to scale scores. Such transformations are often nonlinear. Some examples of nonlinear transformations are developmental standard scores, stanines, ACT scores, and SAT scores. The conditional SEMs calculated typically need to be on the same score scale on which

the test score is reported. To convert raw scores to scale scores, rounding and truncation are often involved.

On the raw score scale, the conditional SEM tends to be large for middle scores and small for extreme scores. If the raw-to-scale score transformation is linear, then the scale score reliability will remain the same and the conditional SEM of the scale scores is a multiple of the conditional SEM of raw scores and the relative magnitude stays the same. However, a nonlinear transformation can change the relative magnitude of conditional SEMs. For example, a transformation that stretches the two ends and compresses the middle of the score distribution can make the conditional SEM fairly consistent across the score scale. With an even more extreme transformation, the conditional SEMs can be made relatively large at the two extremes and small in the middle [10].

Practical Applications

The Standards for Educational and Psychological Testing requires that the conditional SEMs be reported (AERA, APA, & NCME, [1], p. 35, Standard 2.14). It is recommended that

'conditional standard errors of measurement should be reported at several score levels if consistency cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.'

Conditional SEMs provide a 'confidence band' for the observed test score an individual earns.

Because of the requirements in the Test Standards, testing programs report conditional SEMs. When the ACT Assessment was rescaled in 1989 [11], an arcsine transformation was applied on the raw scores, which resulted in conditional SEMs of the scale scores being relatively constant across all score levels for all the subtests.

For some other test batteries, such as SAT scores, conditional SEMs are not forced to be constant across the score levels. Instead, the testing programs typically report conditional SEMs in their test manuals for a number of score levels.

Summary and Conclusion

Conditional SEMs provide important information on the amount of error in observed test scores. Many approaches have been proposed in the literature to estimate conditional SEMs. These approaches are based on different test theory models and assumptions. However, these methods typically produce fairly consistent results when applied to standardized achievement tests. There are no existing rules stating which method to use in the estimation of conditional SEMs. Practitioners can choose a certain method that aligns the best with the assumptions made and the characteristics of their tests. On the raw score scale, conditional SEMs are relatively large in magnitude at the two ends and small in the middle. However, conditional SEMs for scale scores can take on a variety of forms depending on the function of raw-to-scale score transformations that are used.

References

- [1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington.
- [2] Blixt, S.L. & Shama, D.B. (1986). An empirical investigation of the standard error of measurement at different ability levels, *Educational and Psychological Measurement* **46**, 545–550.
- [3] Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in generalizability theory, *Applied Psychological Measurement* **22**, 307–331.
- [4] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.
- [5] Feldt, L.S. (1984). Some relationships between the binomial error model and classical test theory, *Educational and Psychological Measurement* **44**, 883–891.
- [6] Feldt, L.S. & Brennan, R.L. (1989). Reliability, in *Educational Measurement*, R.L. Linn, ed., Macmillan, New York.
- [7] Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments, *Applied Psychological Measurement* **10**, 175–186.
- [8] Keats, J.A. (1957). Estimation of error variances of test scores, *Psychometrika* **22**, 29–41.
- [9] Keats, J.A. & Lord, F.M. (1962). A theoretical distribution for mental test scores, *Psychometrika* **27**, 215–231.
- [10] Kolen, M.J., Hanson, B.A. & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores, *Journal of Educational Measurement* **29**, 285–307.
- [11] Kolen, M.J. & Hanson, B.A. (1989). Scaling the ACT assessment, in *Methodology Used In Scaling The ACT Assessment and P-ACT+*, R.L. Brennan, ed., ACT, Iowa City, pp. 35–55.
- [12] Lee, W., Brennan, R.L. & Kolen, M.J. (2000). Estimators of conditional scale-score standard errors of measurement: a simulation study, *Journal of Educational Measurement* **37**, 1–20.
- [13] Lord, F.M. (1955). Estimating test reliability, *Educational and Psychological Measurement* **15**, 325–336.
- [14] Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement* **17**, 510–521.
- [15] Lord, F.M. (1965). A strong true score theory with applications, *Psychometrika* **30**, 239–270.
- [16] Lord, F.M. (1969). Estimating true-score distribution in a psychological testing (An empirical Bayes estimation problem), *Psychometrika* **34**, 259–299.
- [17] Lord, F.M. (1984). Standard errors of measurement at different score levels, *Journal of Educational Measurement* **21**, 239–243.
- [18] Mollenkopf, W.G. (1949). Variation of the standard error of measurement, *Psychometrika* **14**, 189–229.
- [19] Thorndike, R.L. (1950). Reliability, in *Educational Measurement*, E.F. Lindquist ed., American Council on Education, Washington, pp. 560–620.

YE TONG AND MICHAEL J. KOLEN

Confidence Intervals

CHRIS DRACUP

Volume 1, pp. 366–375

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Confidence Intervals

Behavioral scientists have long been criticized for their overreliance on null hypothesis significance testing in the statistical analysis of their data [4, 13, 17, 21, 22, 28]. A null hypothesis is a precise statement about one or more population parameters from which, given certain assumptions, a probability distribution for a relevant sample statistic can be derived (*see Sampling Distributions*). As its name implies, a null hypothesis is usually a hypothesis that no real effect is present (two populations have identical means, two variables are not related in the population, etc.). A decision is made to reject or not reject such a null hypothesis on the basis of a significance test applied to appropriately collected sample data. A significant result is taken to mean that the observed data are not consistent with the null hypothesis, and that a real effect is present. The significance level represents the conditional probability that the process would lead to the rejection of the null hypothesis given that it is, in fact, true (*see Classical Statistical Inference: Practice versus Presentation*).

Critics have argued that rejecting or not rejecting a null hypothesis of, say, no difference in population means, does not constitute a full analysis of the data or advance knowledge very far. They have implored researchers to report other aspects of their data such as the size of any observed effects. In particular, they have advised researchers to calculate and report confidence intervals for effects of interest [22, 34]. However, it would be a mistake to believe that by relying more on confidence intervals, behavioral scientists were turning their backs on significance testing. Confidence intervals can be viewed as a generalization of the null hypothesis significance test, but in order to understand this, the definition of a null hypothesis has to be broadened somewhat. It is, in fact, possible for a null hypothesis to specify an effect size other than zero, and such a hypothesis can be subjected to significance testing provided it allows the derivation of a relevant sampling distribution. To distinguish such null hypotheses from the more familiar null hypothesis of no effect, the term 'nil hypothesis' was introduced for the latter [6]. Whilst the significance test of the nil null hypothesis indicates whether the data are consistent with a hypothesis that says

there is no effect (or an effect of zero size), a confidence interval indicates all those effect sizes with which the data are and are not consistent.

The construction of a confidence interval will be introduced by considering the estimation of the mean of a normal population of unknown variance from the information contained in a simple random sample from that population. The formula for the confidence interval is a simple rearrangement of that of the single sample *t* Test (*see Catalogue of Parametric Tests*):

$$\bar{X} + t_{0.025, n-1} \left(\frac{\hat{s}}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + t_{0.975, n-1} \left(\frac{\hat{s}}{\sqrt{n}} \right) \quad (1)$$

where $\bar{X}$ is the sample mean, $\hat{s}$ is the unbiased sample standard deviation, n is the sample size, $(\hat{s}/\sqrt{n})$ is an estimate of the standard error of the mean, and $t_{0.025, n-1}$ and $t_{0.975, n-1}$ are the critical values of *t* with $n - 1$ degrees of freedom that cut off the bottom and top 2.5% of the distribution respectively. This is the formula for a 95% confidence interval for the population mean of a normal distribution whose standard deviation is unknown. In repeated applications of the formula to simple random samples from normal populations, the constructed interval will contain the true population mean on 95% of occasions, and not contain it on 5% of occasions. Intervals for other levels of confidence are easily produced by substituting appropriate critical values. For example, if a 99% confidence interval was required, $t_{0.005, n-1}$ and $t_{0.995, n-1}$ would be employed.

The following example will serve to illustrate the calculation. A sample of 20 participants was drawn at random from a population (assumed normal) and each was measured on a psychological test. The sample mean was 283.09 and the unbiased sample standard deviation was 51.42. The appropriate critical values of *t* with 19 degrees of freedom are -2.093 and $+2.093$. Substituting in (1) gives the 95% confidence interval for the mean of the population as $259.02 \leq \mu \leq 307.16$. It will be clear from the formula that the computed interval is centered on the sample mean and extends a number of estimated standard errors above and below that value. As sample size increases, the exact number of estimated standard errors above and below the mean approaches 1.96 as the *t* distribution approaches the standardized normal distribution. Of course, as sample size increases, the estimated standard error of the mean decreases

2 Confidence Intervals

and the confidence interval gets narrower, reflecting the greater accuracy of estimation resulting from a larger sample.

The Coverage Interpretation of Confidence Intervals

Ninety five percent of intervals calculated in the above way will contain the appropriate population mean, and 5% will not. This property is referred to as ‘coverage’ and it is a property of the procedure rather than a particular interval. When it is claimed that an interval contains the estimated population mean with 95% confidence, what should be understood is that the procedure used to produce the interval will give 95% coverage in the long run. However, for any particular interval that has been calculated, it will be either one of the 95% that contain the population mean or one of the 5% that do not. There is no way of knowing which of these two is true. According to the frequentist view of probability (*see Probability: Foundations of*) from which the method derives [24], for any particular confidence interval the population is either included or not included in it. The probability of inclusion is therefore either 1 or 0. The confidence interval approach does not give the probability that the true population mean will be in the particular interval constructed, and the term ‘confidence’ rather than ‘probability’ is employed to reinforce the distinction. Despite the claims of the authors of many elementary textbooks [8], significance tests do not result in statements about the probability of the truth of a hypothesis, and confidence intervals, which derive from the same statistical theory, cannot do so either. The relationship between confidence intervals and significance testing is examined in the next section.

The Significance Test Interpretation of Confidence Intervals

The lower limit of the 95% confidence interval calculated above is 259.02. If the sample data were used to conduct a two-tailed one sample t Test at the 0.05 significance level to test the null hypothesis that the true population mean was equal to 259.02, it would yield a t value that was just nonsignificant. However, any null hypothesis that specified that the population mean was a value less than 259.02 would

be rejected by a test carried out on the sample data. The situation is similar with the upper limit of the calculated interval, 307.16. The term ‘plausible’ has been applied to values of a population parameter that are included in a confidence interval [10, 29]. According to this usage, the 95% confidence interval calculated above showed any value between 259.02 and 307.16 to be a ‘plausible’ value for the mean of the population, as such values would not be rejected by a two-tailed test at the 0.05 level carried out on the obtained sample data.

One-sided Confidence Intervals

In some circumstances, interest is focused on estimating only the highest (or lowest) plausible value for a population parameter. The distinction between two-sided intervals and one-sided intervals mirrors that between two-tailed and one-tailed significance tests. The formula for one of the one-sided 95% confidence intervals for the population mean, derived from the one-tailed single sample t Test is:

$$-\infty \leq \mu \leq \bar{X} + t_{0.950, n-1} \left(\frac{\hat{s}}{\sqrt{n}} \right) \quad (2)$$

Applying (2) to the example data above yields:

$$-\infty \leq \mu \leq 302.97 \quad (3)$$

In the long run, intervals calculated in this way will contain the true population value for 95% of appropriately collected samples, but they are not popular in the behavioral sciences for the same reasons that one-tailed tests are unpopular (*see Classical Statistical Inference: Practice versus Presentation*).

Central and Noncentral Confidence Intervals

The two-sided and one-sided 95% confidence intervals calculated above represent the extremes of a continuum. With the two-sided interval, the 5% rejection region of the significance test was split equally between the two tails giving limits of 259.02 and 307.16. In the long run, such intervals are exactly as likely to miss the true population mean by having a lower bound that is too high as by having an upper bound that is too low. Intervals with this property are known as ‘central’. However, with the

one-sided interval, which ranged from $-\infty$ to 302.97, it is impossible for the interval to lie above the population mean, but in the long run 5% of such intervals will fall below the population mean. This is the most extreme case of a noncentral interval.

It is possible to divide the 5% rejection region in an infinite number of ways between the two tails. All that is required is that the tails sum to 0.05. If the lower rejection region was made to be 1% and the upper region made to be 4% by employing $t_{0.010, n-1}$ and $t_{0.960, n-1}$ as the critical values, then 95% of intervals calculated in this way would include the true population mean and 5% would not. Such intervals would be four times more likely to miss the population mean by being too low than by being too high.

Whilst arguments for employing noncentral confidence intervals have been put forward in other disciplines [15], central confidence intervals are usually employed in the behavioral sciences. In a number of important applications, such as the construction of confidence intervals for population means (and their differences), central confidence intervals have the advantage that they are shorter than noncentral intervals and therefore give the required level of confidence for the shortest range of plausible values.

The Confidence Interval for the Difference Between Two Population Means

If the intention is to compare means, then the appropriate approach is to construct a confidence interval for the difference in population means using a rearrangement of the formula for the two independent samples t Tests (*see Catalogue of Parametric Tests*):

$$\begin{aligned}
 & (\bar{X}_1 - \bar{X}_2) + t_{0.025, n_1+n_2-2} \\
 & \times \sqrt{\left(\frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\
 & \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{0.975, n_1+n_2-2} \\
 & \times \sqrt{\left(\frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (4)
 \end{aligned}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means, $\hat{s}_1$ and $\hat{s}_2$ are the unbiased sample standard deviations, and n_1 and n_2 are the two sample sizes. This is

the formula for the 95% confidence interval for the difference between the means of two normal populations, which are assumed to have the same (unknown) variance. When applied to data from two appropriately collected samples, over the course of repeated application, the upper and lower limits of the interval will contain the true difference in population means for 95% of those applications. The formula is for a central confidence interval that is equally likely to underestimate as overestimate the true difference in population means. Since the sampling distribution involved is symmetrical, this is the shortest interval providing 95% confidence.

If (4) is applied to the sample data that were introduced previously ($\bar{X}_1 = 283.09$, $\hat{s}_1 = 51.42$, and $n_1 = 20$) and to a second sample ($\bar{X}_2 = 328.40$, $\hat{s}_2 = 51.90$, and $n_2 = 21$), the 95% confidence interval for the difference in the two population means ranges from -77.96 to -12.66 (which is symmetrical about the observed difference between the sample means of -45.31). Since this interval does not contain zero, the difference in population means specified by the nil null hypothesis, it is readily apparent that the nil null hypothesis can be rejected at the 0.05 significance level. Thus, the confidence interval provides all the information provided by a two-tailed test of the nil null hypothesis at the corresponding conventional significance level.

In addition, the confidence interval implies that a two-tailed independent samples t Test carried out on the observed data would not lead to the rejection of any null hypothesis that specified the difference in population means was equal to a value in the calculated interval, at the 0.05 significance level. This information is not provided by the usual test of the nil null hypothesis, but is a unique contribution of the confidence interval approach. However, the test of the nil null hypothesis of no population mean difference yields an exact P value for the above samples of 0.008. So the nil null hypothesis can be rejected at a stricter significance level than the 0.05 implied by the 95% confidence interval. So whilst the confidence interval approach provides a test of all possible null hypotheses at a single conventional significance level, the significance testing approach provides a more detailed evaluation of the extent to which the data cast doubt on just one of these null hypotheses.

If, as in this example, all the 'plausible' values for the difference in the means of two populations are of the same sign, then Tukey [32] suggests that

it is appropriate to talk of a ‘confident direction’ for the effect. Here, the 95% confidence interval shows that the confident direction of the effect is that the first population mean is less than the second.

The Confidence Interval for the Difference Between Two Population Means and Confidence Intervals for the Two Individual Population Means

The 95% confidence interval for the mean of the population from which the first sample was drawn has been previously shown to range from 259.03 to 307.16. The second sample yields a 95% confidence interval ranging from 304.77 to 352.02. Values between 304.77 and 307.16 are common to both intervals and therefore ‘plausible’ values for the means of both populations. The overlap of the two intervals would be visually apparent if the intervals were presented graphically, and the viewer might be tempted to interpret the overlap as demonstrating the absence of a significant difference between the means of the two samples. However, the confidence interval for the difference between the population means did not contain zero, so the two sample means do differ significantly at the 0.05 level. This example serves to illustrate that caution is required in the comparison and interpretation of two or more confidence intervals. Confidence intervals and significance tests derived from the data from single samples are often based on different assumptions from those derived from the data from two (or more) samples. The former, then, cannot be used to determine the results of the latter directly. This problem is particularly acute in repeated measures designs [11, 19] (*see Repeated Measures Analysis of Variance*).

Confidence Intervals and Posterior Power Analysis

The entry on **power** in this volume demonstrates convincingly that confidence intervals cannot by themselves provide a substitute for prospective power analysis. However, they can provide useful insights into why a result was not significant, traditionally the role of posterior power analysis. The fact that a test result is not significant at the 0.05 level means that the nil null effect size lies within the 95%

confidence interval. Is this because there is no real effect or is it because there is a real effect but the study’s sample size was inadequate to detect it? If the confidence interval is narrow, the study’s sample size has been shown to be sufficient to provide a relatively accurate estimate of the true effect size, and, whilst it would not be sensible to claim that no real effect was present, it might confidently be claimed that no large effect was present. If, however, the confidence interval is wide, then this means that the study’s sample size was not large enough to provide an accurate estimate of the true effect size. If, further, the interval ranges from a large negative limit to a large positive limit, then it would be unwise to claim anything except that more information is needed.

Confidence Intervals Based on Approximations to a Normal Distribution

As sample sizes increase, the sampling distributions of many statistics tend toward a normal distribution with a standard error that is estimated with increasing precision. This is true, for example, of the mean of a random sample from some unknown nonnormal population, where an interval extending from 1.96 estimated standard errors below the observed sample mean to the same distance above the sample mean gives coverage closer and closer to 95% as sample size increases. A rough and ready 95% confidence interval in such cases would range from two estimated standard errors below the sample statistic to two above. Intervals ranging from one estimated standard error below the sample statistic to one above would give approximately 68% coverage, and this reasoning lies behind the promotion of mean and error graphs by the editors of some journals [18]. Confidence intervals for population correlation coefficients are often calculated in this way, as Fisher’s z transformation of the sample Pearson correlation (r) tends toward a normal distribution of known variance as sample size increases (*see Catalogue of Parametric Tests*). The procedure leads to intervals that are symmetrical about the transformed sample correlation (z_r), but the symmetry is lost when the limits of the interval are transformed back to r .

These are large sample approximations, however, and some caution is required in their interpretation. The use of the normal approximation to the binomial distribution in the construction of a confidence

interval for a population proportion will be used to illustrate some of the issues.

Constructing Confidence Intervals for the Population Proportion

Under repeated random sampling from a population where the population proportion is π , the sample proportion, p , follows a discrete, binomial distribution (see **Catalogue of Probability Density Functions**), with mean π and standard deviation $\sqrt{\pi(1-\pi)/n}$. As n increases, provided π is not too close to zero or one, the distribution of p tends toward a normal distribution. In these circumstances, $\sqrt{p(1-p)/n}$ can provide a reasonably good estimate of $\sqrt{\pi(1-\pi)/n}$, and an approximate 95% confidence interval for π is provided by

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1.96\sqrt{\frac{p(1-p)}{n}} \quad (5)$$

In many circumstances, this may provide a serviceable confidence interval. However, the approach uses a continuous distribution to approximate a discrete one and makes no allowance for the fact that $\sqrt{p(1-p)/n}$ cannot be expected to be exactly equal to $\sqrt{\pi(1-\pi)/n}$. Particular applications of this approximation can result in intervals with an impossible limit (less than zero or greater than one) or of zero width (when p is zero or one).

A more satisfactory approach to the question of constructing a confidence interval for a population proportion, particularly when the sample size is small, is to return to the relationship between the plausible parameters included in a confidence interval and statistical significance. According to this, the lower limit of the 95% confidence interval should be that value of the population proportion, π_L , which if treated as the null hypothesis, would just fail to be rejected in the upper tail of a two-tailed test at the 0.05 level on the observed data. Similarly, the upper limit of the 95% confidence interval should be that value of the population proportion, π_U , which if treated as the null hypothesis, would just fail to be rejected in the lower tail of a two-tailed test at the 0.05 level on the observed data. However, there are particular problems in conducting two-tailed tests on a discrete variable like the sample proportion [20]. As a result, the approach taken is to work in terms of one-tailed tests

and to find the value of the population proportion π_L , which if treated as the null hypothesis, would just fail to be rejected by a one-tailed test (in the upper tail) at the 0.025 level on the observed data, and the value π_U , which if treated as the null hypothesis, would just fail to be rejected by a one-tailed test (in the lower tail) at the 0.025 level on the observed data [3].

The approach will be illustrated with a small sample example that leads to obvious inconsistencies when the normal approximation is applied. A simple random sample of Chartered Psychologists is drawn from the British Psychological Society's Register. Of the 20 psychologists in the sample, 18 are female. The task is to construct a 95% confidence interval for the proportion of Chartered Psychologists who are female.

Using the normal approximation method, (5) would yield:

$$\begin{aligned} 0.9 - 1.96\sqrt{\frac{0.9(1-0.9)}{20}} &\leq \pi \leq 0.9 \\ &+ 1.96\sqrt{\frac{0.9(1-0.9)}{20}} \\ 0.77 &\leq \pi \leq 1.03 \end{aligned} \quad (6)$$

The upper limit of this symmetrical interval is outside the range of the parameter. Even if it was treated as 1, it is clear that if this were made the null hypothesis of a one-tailed test at the 0.025 level, then it must be rejected on the basis of the observed data, since if the population proportion of females really was 1, the sample could not contain any males at all, but it did contain two.

The exact approach proceeds as follows. That value of the population proportion, π_L , is found such that under this null hypothesis the sum of the probabilities of 18, 19 and 20 females in the sample of 20 would equal .025. Reference to an appropriate table [12] or statistical package (e.g., Minitab) yields a value of 0.683017. Then, that value of the population proportion, π_U , is found such that under this null hypothesis the sum of the probabilities of 0, 1, 2, ..., 18 females in a sample of 20 would be .025. The desired value is 0.987651.

So the exact 95% confidence interval for the population proportion of females is:

$$0.683017 \leq \pi \leq 0.987651 \quad (7)$$

If a one-tailed test at the 0.025 level was conducted on any null hypothesis that specified that the population proportion was less than 0.683017, it would

be rejected on the basis of the observed data. So also would any null hypothesis that specified that the population proportion was greater than 0.987651.

In contrast to the limits indicated by the approximate method, those produced by the exact method will always lie within the range of the parameter and will not be symmetrical around the sample proportion (except when this is exactly 0.5). The exact method yields intervals that are conservative in coverage [23, 33], and alternatives have been suggested. One approach questions whether it is appropriate to include the full probability of the observed outcome in the tail when computing π_L and π_U . Berry and Armitage [3] favor **Mid-P** confidence intervals where only half the probability of the observed outcome is included. The Mid-P intervals will be rather narrower than the exact intervals, and will result in a rather less conservative coverage. However, in some circumstances the Mid-P coverage can fall below the intended level of confidence, which cannot occur with the exact method [33].

Confidence Intervals for Standardized Effect Size Measures

Effect size can be expressed in a number of ways. In a simple two condition experiment, the population effect size can be expressed simply as the difference between the two population means, $\mu_1 - \mu_2$, and a confidence interval can be constructed as illustrated above. The importance of a particular difference between population means is difficult to judge except by those who are fully conversant with the measurement scale that is being employed. Differences in means are also difficult to compare across studies that do not employ the same measurement scale, and the results from such studies prove difficult to combine in **meta-analyses**. These considerations led to the development of standardized effect size measures that do not depend on the particular units of a measurement scale. Probably the best known of these measures is Cohen's d . This expresses the difference between the means of two populations in terms of the populations' (assumed shared) standard deviation:

$$d = \frac{\mu_1 - \mu_2}{\sigma} \quad (8)$$

Standardized effect size measures are unit-less, and therefore, it is argued, are comparable across

studies employing different measurement scales. A d value of 1.00 simply means that the population means of the two conditions differ by one standard deviation. Cohen [5] identified three values of d (0.2, 0.5, and 0.8) to represent small, medium, and large effect sizes in his attempts to get psychologists to take power considerations more seriously, though the appropriateness of 'canned' effect sizes has been questioned by some [14, 16].

Steiger and Fouladi [31] have provided an introduction to the construction of confidence intervals for standardized effects size measures like d from their sample estimators. The approach, based on noncentral probability distributions, proceeds in much the same way as that for the construction of an exact confidence interval for the population proportion discussed above. For a 95% confidence interval, those values of the noncentrality parameter are found (by numerical means) for which the observed effect size is in the bottom 0.025 and top 0.025 of the sampling distributions. A simple transformation converts the obtained limiting values of the noncentrality parameter into standardized effect sizes. The calculation and reporting of such confidence intervals may serve to remind readers that observed standardized effect sizes are random variables, and are subject to sampling variability like any other sample statistic.

Confidence Intervals and the Bootstrap

Whilst the central limit theorem might provide support for the use of the normal distribution in constructing approximate confidence intervals in some situations, there are other situations where sample sizes are too small to justify the process, or sampling distributions are suspected or known to depart from normality. The **Bootstrap** [7, 10] is a numerical method designed to derive some idea of the sampling distribution of a statistic without recourse to assumptions of unknown or doubtful validity. The approach is to treat the collected sample of data as if it were the population of interest. A large number of random samples are drawn with replacement from this 'population', each sample being of the same size as the original sample. The statistic of interest is calculated for each of these 'resamples' and an empirical sampling distribution is produced. If a large enough number of resamplings is performed, then a 95% confidence interval can be produced directly from these

by identifying those values of the statistic that correspond to the 2.5th and 97.5th percentiles. On other occasions, the standard deviation of the statistic over the 'resamples' is calculated and this is used as an estimate of the true standard error of the statistic in one of the standard formulae. Various ways of improving the process have been developed and subjected to some empirical testing. With small samples or with samples that just by chance are not very representative of their parent population, the method may provide rather unreliable information. Nonetheless, Efron, and Tibshirani [10] provide evidence that it works well in many situations, and the methodology is becoming increasingly popular.

Bayesian Certainty Intervals

Confidence intervals derive from the frequentist approach to statistical inference that defines probabilities as the limits in the long run of relative frequencies (*see* **Probability: Foundations of**). According to this view, the confidence interval is the random variable, not the population parameter [15, 24]. In consequence, when a 95% confidence interval is constructed, the frequentist position does not allow statements of the kind 'The value of the population parameter lies in this interval with probability 0.95'. In contrast, the Bayesian approach to statistical inference (*see* **Bayesian Statistics**) defines probability as a subjective or psychological variable [25, 26, 27]. According to this view, it is not only acceptable to claim that particular probabilities are associated with particular values of a parameter; such claims are the very basis of the inference process. Bayesians use Bayes's Theorem (*see* **Bayesian Belief Networks**) to update their prior beliefs about the probability distribution of a parameter on the basis of new information. Sometimes, this updating process uses elements of the same statistical theory that is employed by a frequentist and the final result may appear to be identical to a confidence interval. However, the interpretation placed on the resulting interval could hardly be more different.

Suppose that a Bayesian was interested in estimating the mean score of some population on a particular measurement scale. The first part of the process would involve attempting to sketch a prior probability distribution for the population mean, showing

what value seemed most probable, whether the distribution was symmetrical about this value, how spread out the distribution was, perhaps trying to specify the parameter values that contained the middle 50% of the distribution, the middle 90%, and so on. This process would result in a subjective prior distribution for the population mean. The Bayesian would then collect sample data, in very much the same way as a Frequentist, and use Bayes's Theorem to combine the information from this sample with the information contained in the prior distribution to compute a posterior probability distribution. If the 2.5th and 97.5th percentiles of this posterior probability distribution are located, they form the lower and upper bounds on the Bayesian's 95% certainty interval [26].

Certain parallels and distinctions can be drawn between the confidence interval of the Frequentist and the certainty interval of the Bayesian. The center of the Frequentist's confidence interval would be the sample mean, but the center of the Bayesian's certainty interval would be somewhere between the sample mean and the mean of the prior distribution. The smaller the variance of the prior distribution, the nearer the center of the certainty interval will be to that of the prior distribution. The width of the Frequentist's confidence interval would depend only on the sample standard deviation and the sample size, but the width of the Bayesian's certainty interval would be narrower as the prior distribution contributes a 'virtual' sample size that can be combined with the actual sample size to yield a smaller posterior estimate of the standard error as well as contribute to the degrees of freedom of any required critical values.

So, to the extent that the Bayesian had any views about the likely values of the population mean before data were collected, these views will influence the location and width of the 95% certainty interval. Any interval so constructed is to be interpreted as a probability distribution for the population mean, permitting the Bayesian to make statements like, 'The probability that μ lies between L and U is X.' Of course, none of this is legitimate to the Frequentist who distrusts the subjectivity of the Bayesian's prior distribution, its influence on the resulting interval, and the attribution of a probability distribution to a population parameter that can only have one value.

If the Bayesian admits ignorance of the parameter being estimated prior to the collection of the sample, then a uniform prior might be specified which says that any and all values of the population parameter are

equally likely. In these circumstances, the certainty interval of the Bayesian and the confidence interval of the Frequentist will be identical in numerical terms. However, an unbridgeable gulf will still exist between the interpretations that would be made of the interval.

Confidence Intervals in Textbooks and Computer Packages

Despite the recognition that they have been neglected in the past, most statistics textbooks for behavioral scientists give rather little space to confidence intervals (compared to significance tests). Smithson has gone some way to filling the gap, though his earlier and more elementary book [29] was less than accurate on a number of topics [9]. His more recent book [30] has a wider range and is more accurate, but it may be at too advanced a level for many behavioral scientists. At the moment, the best texts seem to be coming from medicine. Altman et al. [2] provide a fairly simple introduction to confidence intervals as well as detailed instructions on how to construct them for a number of important parameters. A disc is provided with the text that contains useful programs for computing intervals from raw data or summary statistics. Armitage et al. [1] offer a well-balanced blend of significance tests and confidence intervals at a level that should be accessible to most graduates from the behavioral sciences.

A search of the World Wide Web reveals many sites offering confidence interval calculations for means, variances, correlations, and proportions. These must be approached with some caution as they are not always explicit about just what methodology is being employed. Like the textbooks, the commercial statistical packages commonly used by behavioral scientists tend to neglect confidence intervals compared to significance tests. Packages tend to provide confidence intervals for means and mean differences (including some simultaneous intervals) and for the parameters of regression analysis. There is limited cover of proportions, less of correlations, and hardly any of variances.

It will be some time yet before textbooks and software provide adequate support for the guideline of the American Psychological Association Task Force on Statistical Inference [34] that:

Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients of association or variation whenever possible. [p. 599]

References

- [1] Armitage, P., Berry, G. & Matthews, J.N.S. (2002). *Statistical Methods in Medical Research*, 4th Edition, Blackwell, Oxford.
- [2] Altman, D.G., Machin, D., Bryant, T.N. & Gardner, M.J., eds (2000). *Statistics with Confidence*, 2nd Edition, BMJ Books, London.
- [3] Berry, G. & Armitage, P. (1995). Mid-P confidence intervals: a brief review, *The Statistician* **44**, 417–423.
- [4] Carver, R.P. (1978). The case against statistical significance testing, *Harvard Educational Review* **48**, 378–399.
- [5] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum Associates, Hillsdale.
- [6] Cohen, J. (1994). The earth is round ($p < .05$), *American Psychologist* **49**, 997–1003.
- [7] Davison, A.C. & Hinkley, D.V. (2003). *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.
- [8] Dracup, C. (1995). Hypothesis testing: what it really is, *Psychologist* **8**, 359–362.
- [9] Dracup, C. (2000). Book review: Statistics with confidence, *British Journal of Mathematical and Statistical Psychology* **53**, 333–334.
- [10] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [11] Estes, W.K. (1997). On the communication of information by displays of standard errors and confidence intervals, *Psychonomic Bulletin and Review* **4**, 330–341.
- [12] Geigy Scientific Tables (1982). Vol. 2: *Introduction to Statistics, Statistical Tables, Mathematical Formulae*, 8th Edition, Ciba-Geigy, Basle.
- [13] Hunter, J.E. (1997). Needed: a ban on the significance test, *Psychological Science* **8**, 3–7.
- [14] Jiroutek, M.R., Muller, K.E., Kupper, L.L. & Stewart, P.W. (2003). A new method for choosing sample size for confidence interval-based inferences, *Biometrics* **59**, 580–590.
- [15] Kendall, M.G. & Stuart, A. (1979). *The Advanced Theory of Statistics. Vol. 2. Inference and Relationship*, 4th Edition, Griffin, London.
- [16] Lenth, R.V. (2001). Some practical guidelines for effective sample size determination, *American Statistician* **55**, 187–193.
- [17] Loftus, G.R. (1993a). A picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age, *Behavior Research Methods, Instruments, & Computers* **25**, 250–256.
- [18] Loftus, G.R. (1993b). Editorial comment, *Memory and Cognition* **21**, 1–3.

-
- [19] Loftus, G.R. & Masson, M.E.J. (1994). Using confidence intervals in within-subjects designs, *Psychonomic Bulletin & Review* **1**, 476–480.
- [20] Macdonald, R.R. (1998). Conditional and unconditional tests of association in 2×2 tables, *British Journal of Mathematical and Statistical Psychology* **51**, 191–204.
- [21] Meehl, P.E. (1967). Theory testing in psychology and in physics: a methodological paradox, *Philosophy of Science* **34**, 103–115.
- [22] Meehl, P.E. (1997). The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum Associates, Mahwah.
- [23] Neyman, J. (1935). On the problem of confidence intervals, *Annals of Mathematical Statistics* **6**, 111–116.
- [24] Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society, A* **236**, 333–380.
- [25] Novick, M.R. & Jackson, P.H. (1974). *Statistical Methods for Educational and Psychological Research*, McGraw-Hill, New York.
- [26] Phillips, L.D. (1993). *Bayesian Statistics for Social Scientists*, Nelson, London.
- [27] Pruzek, R.M. (1997). An introduction to Bayesian inference and its applications, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum Associates, Mahwah.
- [28] Schmidt, F.L. (1996). Statistical significance and cumulative knowledge in psychology: implications for the training of researchers, *Psychological Methods* **1**, 115–129.
- [29] Smithson, M. (2000). *Statistics with Confidence*, Sage, London.
- [30] Smithson, M. (2003). *Confidence Intervals*, Sage, Thousand Oaks.
- [31] Steiger, J.H. & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical models, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum Associates, Mahwah.
- [32] Tukey, J. (1991). The philosophy of multiple comparisons, *Statistical Science* **6**, 100–116.
- [33] Vollset, S.E. (1993). Confidence intervals for a binomial proportion, *Statistics in Medicine* **12**, 809–824.
- [34] Wilkinson, L. and The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations, *American Psychologist* **54**, 594–604.

CHRIS DRACUP

Confidence Intervals: Nonparametric

PAUL H. GARTHWAITE

Volume 1, pp. 375–381

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Confidence Intervals: Nonparametric

Introduction

Hypothesis testing and constructing **confidence intervals** are among the most common tasks in statistical inference. The two tasks are closely linked. Suppose interest focuses on a parameter θ and a null hypothesis specifies that θ_0 is the value taken by θ . Then, the null hypothesis is rejected by a two-tailed test at significance level α if, and only if, θ_0 is not contained in the $100(1 - \alpha)\%$ central confidence interval for θ . Most methods of forming confidence intervals exploit this relationship, including those described later in this section. However, the first methods we consider – bootstrap methods – are an exception and determine confidence intervals directly from percentiles of a constructed ‘bootstrap’ distribution.

Bootstrap Confidence Intervals

The fundamental steps of the standard **bootstrap** method are as follows.

1. The data are used as a nonparametric estimate of the distribution from which they were drawn. Specifically, suppose the data consist of n observations $y_1, \dots, y_n$, that were drawn from a population with distribution (cumulative distribution function) F . Then, the estimate of F states that every member of the population takes one of the values $y_1, \dots, y_n$ and is equally likely to take any one of these values. Let $\hat{F}$ denote the estimate of F .
2. The sampling method that gave the data is replicated, but using $\hat{F}$ rather than F . Thus, n observations are selected at random *but with replacement* from $y_1, \dots, y_n$. For example, if the original data consist of five values, 4.7, 3.6, 3.9, 5.2, 4.5, then a set of resampled values might be 3.6, 5.2, 3.9, 5.2, 3.9. The resampled values are referred to as a *bootstrap sample*. Let $y_1^*, \dots, y_n^*$ denote the values in the resample.
3. Suppose θ is the population parameter of interest and that $\hat{\theta}$ is the estimate of θ obtained from

$y_1, \dots, y_n$. An estimate of θ , say $\hat{\theta}^*$, is constructed from the bootstrap sample in precisely the same way as $\hat{\theta}$ was obtained from the original sample.

4. Steps 2 and 3 are repeated many times, so as to yield a large number of bootstrap samples, from each of which an estimate $\hat{\theta}^*$ is obtained. The **histogram** of these estimates is taken as an approximation to the *bootstrap probability distribution* of $\hat{\theta}^*$. Bootstrap inferences about θ are based on $\hat{\theta}$ and this distribution.

There are two common methods of forming bootstrap confidence intervals. One is Efron’s *percentile method* [3], which equates **quantiles** of the bootstrap distribution of $\hat{\theta}^*$ to the equivalent quantiles of the distribution of $\hat{\theta}$. The other method is termed the *basic bootstrap method* by Davison and Hinkley [2]. This assumes that the relationship between $\hat{\theta}$ and F is similar to the relationship between $\hat{\theta}^*$ and $\hat{F}$. Specifically, it assumes that the distribution of $\hat{\theta} - \theta$ (where $\hat{\theta}$ is the random quantity) is approximately the same as the distribution of $\hat{\theta}^* - \hat{\theta}$, where $\hat{\theta}^*$ is the random quantity. Let $\theta^*(\alpha)$ denote the α -level quantile of the bootstrap distribution. Then, the percentile method specifies that the $100(1 - 2\alpha)\%$ equal-tailed confidence interval is

$$(\theta^*(\alpha), \theta^*(1 - \alpha))$$

whereas the basic bootstrap method specifies the $100(1 - 2\alpha)\%$ equal-tailed confidence interval as

$$(\hat{\theta} - \{\theta^*(1 - \alpha) - \hat{\theta}\}, \hat{\theta} + \{\hat{\theta} - \theta^*(\alpha)\}).$$

As an example, we will consider a set of data about law school students that has been widely used with the percentile method (e.g., [3], where the data are reported). The data consist of the average scores on two criteria of students entering 15 law schools. One score is LSAT (a national test similar to the Graduate Record Exam) and the other is GPA (undergraduate grade point average). The average scores for each law school are given in Table 1.

The two scores are clearly correlated and have a sample correlation of 0.776. A confidence interval for the population correlation can be determined using bootstrap methods by applying the above procedure, as follows.

1. Define y_i to be the pair of values for the i th law school (e.g., $y_1 = (576, 3.39)$). $\hat{F}$ states that

2 Confidence Intervals: Nonparametric

Table 1 Average LSAT and GPA for students entering 15 law schools

Law school	Average LSAT	Average GPA
1	576	3.39
2	635	3.30
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3.00
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

each y_i has a probability of $1/15$ of occurring whenever a law school is picked at random.

- As the original sample size was 15, a bootstrap sample is chosen by selecting 15 observations from $\hat{F}$.
- θ is the true population correlation between LSAT and GPA across all law schools. In the original data, the sample correlation between LSAT and GPA is $\hat{\theta} = 0.776$; $\hat{\theta}^*$ is their correlation in a bootstrap sample.
- Steps 2 and 3 are repeated many times to obtain an estimate of the bootstrap distribution of $\hat{\theta}^*$.

For the law school data, we generated 10 000 bootstrap samples of 15 observations and determined the sample correlation for each. A histogram of the 10 000 correlations that were obtained is given in Figure 1.

The histogram shows that the distribution of bootstrap correlations has marked skewness. The sample size is small and a 95% confidence interval for θ would be very wide, so suppose a central 80% confidence interval is required. One thousand of the 10 thousand values of $\hat{\theta}^*$ were less than 0.595 and 1000 were above 0.927, so $\hat{\theta}(0.025) = 0.595$ and $\hat{\theta}(0.975) = 0.927$. The percentile method thus yields $(0.595, 0.927)$ as an 80% confidence interval for θ , while the basic bootstrap method gives an interval of $(0.776 - (0.927 - 0.776), 0.776 + (0.776 - 0.595)) = (0.625, 0.957)$. The difference between the two confidence intervals is substantial because of the skewness of the bootstrap distribution.

Whether the percentile method is to be preferred to the basic bootstrap method, or *vice-versa*, depends upon characteristics of the actual distribution of $\hat{\theta}$. If there is some transformation of $\hat{\theta}$ that gives a symmetric distribution, then the percentile method is optimal. Surprisingly, this is the case even if we do not know what the transformation is. As an example, for the sample correlation coefficient there is a transformation that gives an approximately symmetric distribution (Fisher's $\tanh^{-1}$ transformation). Hence, in the above example the percentile confidence interval is to be preferred to the basic bootstrap interval, even though no transformations were used. Some situations where the basic bootstrap method is optimal are described elsewhere in this encyclopedia (see **Bootstrap Inference**).

A variety of extensions have been proposed to improve the methods, notably the bias-corrected percentile method [4] and the bootstrap- t method [8], which are modifications of the percentile method and basic bootstrap method, respectively. It should be mentioned that the quantity that should be resampled for bootstrapping is not always obvious. For instance, in regression problems the original data

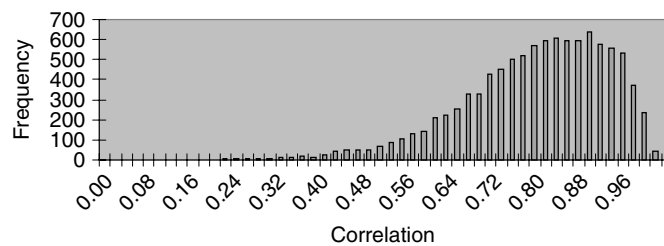


Figure 1 Histogram of bootstrap sample correlations for law school data

points are sometimes resampled (analogous to the above example), while sometimes a regression model is fitted and the residuals from the model are resampled.

Confidence Intervals from Permutation Tests

Permutation tests (*see* **Permutation Based Inference**) and randomization tests (*see* **Randomization Based Tests**) are an appealing approach to hypothesis testing. They typically make fewer distributional assumptions than parametric tests and usually they are just slightly less powerful (*see* **Power**) than their parametric alternatives. Some permutation tests are used to test hypotheses that do not involve parameters, such as whether observations occur at random or whether two variables are independent. This may seem natural, since without parametric assumptions there are few parameters to test. However, quantities such as population means are well-defined without making distributional assumptions and permutation tests may be used to test hypotheses about their value or the difference between two population means, for example. If a permutation test or randomization test can be used to test whether some specified value is taken by a parameter, then the test also enables a confidence interval for the parameter to be constructed.

In a permutation test, the first step is to choose a plausible test statistic for the hypothesis under consideration and to determine the value it takes for the observed data. Then, we find permutations of the data such that the probability of each permutation can be determined under the null hypothesis, H_0 ; usually the permutations are chosen so that each of them is equally probable under H_0 . The value of the test statistic is then calculated for each permutation and a P value is evaluated by comparing the value of the test statistic for the actual data with the probability distribution of the statistic over all the permutations. H_0 is rejected if the observed value is far into a tail of

the probability distribution. In practice, the number of permutations may be so large that the test statistic is only evaluated for a randomly chosen subset of them, in which case the permutation test is often referred to as a randomization test.

Suppose interest centers on a scalar parameter θ and that, for any value θ_0 we specify, a permutation or randomization test can be used to test the hypothesis that θ takes the value θ_0 . Also, assume one-sided tests can be performed and let α_1 be the P value from the test of $H_0: \theta = \theta_L$ against $H_1: \theta > \theta_L$ and let α_2 be the P value from the test of $H_0: \theta = \theta_U$ against $H_1: \theta < \theta_U$. Then, from the relationship between hypothesis tests and confidence interval, (θ_L, θ_U) is a $100(1 - \alpha_1 - \alpha_2)\%$ confidence interval for θ . Usually we want to specify α_1 and α_2 , typically choosing them to each equal 0.025 so as to obtain a central 95% confidence interval. Then, the task is to find values θ_L and θ_U that give these P values.

In practice, the values of θ_L and θ_U are often found by using a simple trial-and-error search based on common sense. An initial ‘guess’ is made of the value of θ_L . Denoting this first guess by L_1 , a permutation test is conducted to test $H_0: \theta = L_1$ against $H_1: \theta > L_1$. Usually between 1000 and 5000 permutations would be used for the test.

Taking account of the P value from this first test, a second guess of the value of θ_L is made and another permutation test conducted. This sequence of trial-and-error continues until the value of θ_L is found to an adequate level of accuracy, (when the P value of the test will be close to α_1). A separate search is conducted for θ_U .

As an example, we consider data from a study into the causes of schizophrenia [12]. Twenty-five hospitalized schizophrenic patients were treated with antipsychotic medication and, some time later, hospital staff judged ten of the patients to be psychotic and fifteen patients to be nonpsychotic. Samples of cerebrospinal fluid were taken from each patient and assayed for dopamine b -hydroxylase activity (nmol/(ml)(hr)/mg). Results are given in Table 2.

Table 2 Dopamine levels of schizophrenic patients

Patients judged psychotic (X):	0.0150	0.0204	0.0208	0.0222	0.0226
	0.0245	0.0270	0.0275	0.0306	0.0320
Patients judged nonpsychotic (Y):	0.0104	0.0105	0.0112	0.0116	0.0130
	0.0145	0.0154	0.0156	0.0170	0.0180
	0.0200	0.0200	0.0210	0.0230	0.0252

4 Confidence Intervals: Nonparametric

The sample totals are $\Sigma x_i = 0.2426$ and $\Sigma y_i = 0.2464$, giving sample means $\bar{x} = 0.02426$ and $\bar{y} = 0.1643$. We assume the distributions for the two types of patients are identical in shape, differing only in their locations, and we suppose that a 95% confidence interval is required for $\theta = \mu_x - \mu_y$, the difference between the population means. The point estimate of θ is $\hat{\theta} = 0.02426 - 0.1643 = 0.00783$.

Let $\theta_L = 0.006$ be the first guess at the lower confidence limit. Then, a permutation test of $H_0: \theta = 0.006$ against $H_1: \theta > 0.006$ must be conducted. To perform the test, 0.006 is subtracted from each measurement for the patients judged psychotic. Denote the modified values by $x'_1, \dots, x'_{10}$ and their mean by $\bar{x}'$. Under H_0 , the distribution of X' and Y are identical, so a natural test statistic is $\bar{x}' - \bar{y}$, which should have a value close to 0 if H_0 is true. Without permutation the value of $\bar{x}' - \bar{y}$ equals $0.00783 - 0.006 = 0.00183$. A permutation of the data simply involves relabeling 10 of the values $x'_1, \dots, x'_{10}, y_1, \dots, y_{15}$ as X -values and relabeling the remaining 15 values as Y -values. The difference between the mean of the relabeled X -values and the mean of the relabeled Y -values is the value of the test statistic for this permutation. Denote this difference by d .

The number of possible permutations is the number of different ways of choosing 10 values from 25 and exceeds 3 million. This number is too large for it to be practical to evaluate the test statistics for all permutations, so instead a random selection of 4999 permutations was used, giving 5000 values of the test statistic when the value for the unpermuted data (0.00183) is included. Nine hundred and twelve of these 5000 values were equal to or exceeded 0.00183, so H_0 is rejected at the 18.24% level of significance. The procedure was repeated using further guesses/estimates of the lower confidence limit. The values examined and the significance levels from the hypothesis tests are shown in Table 3.

The value of 0.0038 was taken as the lower limit of the confidence interval and a corresponding search for the upper limit gave 0.0120 as its estimate. Hence the permutation method gave (0.0038, 0.0120) as the 95% confidence interval for the mean difference in dopamine level between the two types of patient.

A trial-and-error search for confidence limits requires a substantial number of permutations to be performed and is clearly inefficient in terms of computer time. (The above search for just the lower confidence limit required 25 000 permutations). Human input to give the next estimate is also required several times during the search. A much more automated and efficient search procedure is based on the Robbins–Monro process. This search procedure has broad application and is described below under *Distribution-free confidence intervals*. Also for certain simple problems there are methods of determining permutation intervals that do not rely on search procedures. In particular, there are methods of constructing permutation intervals for the mean of a symmetric distribution [9], the difference between the locations of two populations that differ only in their location [9] and the regression coefficient in a simple regression model [5].

Distribution-free Confidence Intervals

The most common distribution-free tests are rank-based tests and methods of forming confidence intervals from some specific rank-based tests are described elsewhere in this encyclopedia. Here we simply give the general method and illustrate it for the Mann–Whitney test, before considering a distribution-free method that does not involve ranks.

The Mann–Whitney test compares independent samples from two populations. It is assumed that the populations have distributions of similar shape but whose means differ by θ . For definiteness, let μ_x and μ_y denote the two means and let $\theta = \mu_x - \mu_y$. The Mann–Whitney test examines the null hypothesis that θ equals a specified value, say θ_0 , testing it against a one- or two-sided alternative hypothesis. (If the null hypothesis is that the population distributions are identical, then $\theta_0 = 0$.) The mechanics of the test are to subtract θ_0 from each of the sample values from the population with mean μ_x . Let $x'_1, \dots, x'_m$ denote these revised values and let $y_1, \dots, y_n$ denote the values from the other sample. The combined set of values $x'_1, \dots, x'_m, y_1, \dots, y_n$, are then ranked from smallest to largest and the sum of the ranks of the x' -values is

Table 3 Sequence of confidence limits and their corresponding significance levels

Estimate of confidence limit:	0.006	0.004	0.003	0.0037	0.0038
Significance level:	18.24	3.04	1.42	2.34	2.54

determined. This sum (or some value derived from its value and the sample sizes) is used as the test statistic. For small sample sizes the test statistic is compared with tabulated critical values and for larger sample sizes an asymptotic approximation is used.

The correspondence between confidence intervals and hypothesis tests can be used to form confidence intervals. For an equal-tailed $100(1 - 2\alpha)\%$ confidence interval, values θ_U and θ_L are required such that $H_0: \theta = \theta_L$ is rejected in favour of $H_1: \theta > \theta_L$ at a P value of α and $H_0: \theta = \theta_U$ is rejected in favour of $H_1: \theta < \theta_U$ at a P value, again, of α . With rank-based tests, there are only a finite number of values that the test statistic can take (rather than a continuous range of values) so it is only possible to find P values that are close to the desired value α . An advantage of rank-based tests, though, is that often there are quick computational methods for finding values θ_L and θ_U that give the P values closest to α . The broad strategy is as follows.

1. For a one-sample test, order the data values from smallest to largest and, for tests involving more than one sample, order each sample separately.
2. Use statistical tables to find critical values of the test statistic that correspond to the P values closest to α .
3. The extreme data values are the ones that increase the P value. For each confidence limit, separately determine the set of data values or combinations of data values that together give a test statistic that equals the critical value or is just in the critical region. The combination of data values that were last to be included in this set determine the confidence interval.

The description of Step 3 is necessarily vague as it varies from test to test. To illustrate the method, we consider the data on schizophrenic patients given in Table 1, and use the Mann–Whitney test to derive a 95% confidence interval for θ , the difference in dopamine b -hydroxylase activity between patients judged psychotic and those judged nonpsychotic. The data for each group of patients have already been ordered according to size, so Step 1 is not needed. The two sample sizes are ten and fifteen and as a test statistic we will use the sum of the ranks of the psychotic patients. From statistical tables (e.g., Table A.7 in [1]), 95 is the critical value at significance level $\alpha = 0.025$ for a one-tailed test of $H_0: \theta = \theta_U$ against $H_1: \theta < \theta_U$. Let $X' = X - \theta_U$. If

the X' -values were all smaller than all the Y -values, the sum of their ranks would be $10(10 + 1)/2 = 55$. As long as forty or fewer of the $X' - Y$ differences are positive, H_0 is rejected, as $40 + 55$ equals the critical value. At the borderline between ‘rejecting’ and ‘not rejecting’ H_0 , the 40th largest $X' - Y$ difference is 0. From this it follows that the upper confidence limit is equal to the 40th largest value of $X - Y$. The ordered data values in Table 1 simplify the task of finding the 40th largest difference. For example, the $X - Y$ differences greater than 0.017 are the following 12 combinations: $X = 0.0320$ in conjunction with $Y = 0.0104, 0.0105, 0.0112, 0.0116, 0.0130$ or 0.0145 ; $X = 0.0306$ in conjunctions with $Y = 0.0104, 0.0105, 0.0112, 0.0116$ or 0.0130 ; $X = 0.0275$ in conjunction with $Y = 0.0104$. A similar count shows that the 40th largest $X - Y$ difference is 0.0120, so this is the upper limit of the confidence interval. Equivalent reasoning shows that the lower limit is the 40th smallest $X - Y$ difference, which here takes the value 0.0035. Hence, the Mann–Whitney test yields (0.0035, 0.0120) as the central 95% confidence interval for θ .

A versatile method of forming confidence intervals is based on the Robbins–Monro process [11]. The method can be used to construct confidence intervals in one-parameter problems, provided the mechanism that gave the real data could be simulated if the parameter’s value were known. We first consider this type of application before describing other situations where the method is useful.

Let θ denote the unknown scalar parameter and suppose a $100(1 - 2\alpha)\%$ equal-tailed confidence interval for θ is required. The method conducts a separate sequential search for each endpoint of the confidence interval. Suppose, first, that a search for the upper limit, θ_U , is being conducted, and let U_i be the current estimate of the limit after i steps of the search. The method sets θ equal to U_i and then generates a set of data using a mechanism similar to that which gave the real data. From the simulated data an estimate of θ is determined, $\hat{\theta}_i$ say. Let $\hat{\theta}^*$ denote the estimate of θ given by the actual sample data. The updated estimate of θ_U , U_{i+1} , is given by

$$U_{i+1} = \begin{cases} U_i - c\alpha/i, & \text{if } \hat{\theta}_i > \hat{\theta}^* \\ U_i + c(1 - \alpha)/i, & \text{if } \hat{\theta}_i \leq \hat{\theta}^* \end{cases} \quad (1)$$

where c is a positive constant that is termed the *step-length constant*.

Table 4 Mark-recapture data for a population of frogs

Sample	Sample size	Number of recaptures	No. of marked frogs in population before sample
1	109	—	—
2	133	15	109
3	138	30	227
4	72	23	335
5	134	47	384
6	72	33	471

The method may be thought of as stepping from one estimate of θ_U to another and c governs the magnitude of the steps. If U_i is equal to the $100(1 - \alpha_i)\%$ point, the expected distance stepped is

$$(1 - \alpha_i)(-c\alpha/i) + \alpha_i c(1 - \alpha)/i = c(\alpha_i - \alpha)/i,$$

which shows that each step reduces the expected distance from θ_U . A predetermined number of steps are taken and the last U_i is adopted as the estimate of θ_U . An independent search is carried out for the lower limit, θ_L . If L_i is the estimate after i steps of the search, then L_{i+1} is found as

$$L_{i+1} = \begin{cases} L_i + c\alpha/i, & \text{if } \hat{\theta}_i < \hat{\theta}^* \\ L_i - c(1 - \alpha)/i, & \text{if } \hat{\theta}_i \geq \hat{\theta}^* \end{cases} \quad (2)$$

This method of forming confidence intervals was developed by Garthwaite and Buckland [7]. They suggest methods of choosing starting values for a search and of choosing the value of the step-length constant. They typically used 5000 steps in the search for an endpoint.

As a practical example where the above method is useful, consider the data in Table 4. These are from a multiple-sample mark-recapture experiment to study a population of frogs [10]. Over a one-month period, frogs were caught in six random samples and, after a sample had been completed, the frogs that had been caught were marked and released. Table 4 gives the numbers of frogs caught in each sample, the number of these captures that were recaptures and the number of frogs that has been marked before the sample. One purpose of a mark-recapture experiment is to estimate the population size, θ say. A point estimate of θ can be obtained by maximum likelihood but most methods of forming confidence intervals require asymptotic approximations that may be inexact. The method based on the Robbins–Monro process can be applied, however, and will give exact intervals. At the

i th step in the search for an endpoint, a population size is specified (U_i or L_i) and equated to θ . Then it is straightforward to simulate six samples of the sizes given in Table 4 and to record the number of recaptures. The estimate of θ based on this resample is $\hat{\theta}_i$ and its value determines the next estimate of the endpoint. Garthwaite and Buckland [7] applied the method to these data and give (932, 1205) as a 95% confidence interval for the population size.

The procedure developed by Garthwaite and Buckland can also be used to form confidence intervals from randomization tests [6]. It is assumed, of course, that the randomization test examines whether a scalar parameter θ takes a specified value θ_0 and that, for any θ_0 , the hypothesis $H_0: \theta = \theta_0$ could be tested against one-sided alternative hypotheses. In the search for the upper limit, suppose U_i is the estimate of the limit after i steps of the Robbins–Monro search. Then the mechanics for a randomization test of the hypothesis $H_0: \theta = U_i$ against $H_1: \theta < U_i$ are followed, except that only a *single* permutation of the data is taken.

Let T_i denote the value of the test statistic from this permutation and let T_1^* denote its value for the actual data (before permutation). The next estimate of the limit is given by

$$U_{i+1} = \begin{cases} U_i - c\alpha/i, & \text{if } T_i > T_1^* \\ U_i + c(1 - \alpha)/i, & \text{if } T_i \leq T_1^* \end{cases} \quad (3)$$

where c is the step-length constant defined earlier and a $100(1 - 2\alpha)\%$ confidence interval is required. A predetermined number of steps are taken and the last U_i is adopted as the estimate of the upper limit. An equivalent search is conducted for the lower limit. An important feature of the search process is that only one permutation is taken at each hypothesized value, rather than the thousands that are taken in unsophisticated search procedures. Typically,

5000 permutations are adequate for estimating each confidence limit.

Most hypothesis tests based on ranks may be viewed as permutation tests in which the actual data values are replaced by ranks. Hence, Robbins–Monro searches may be used to derive confidence intervals from such tests. This approach can be useful if the ranks contain a large number of ties – rank tests typically assume that there are no ties in the ranks and the coverage of the confidence intervals they yield may be uncertain when this assumption is badly violated.

References

- [1] Conover, W.J. (1980). *Practical Nonparametric Statistics*, John Wiley, New York.
- [2] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge.
- [3] Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation, *American Statistician* **37**, 36–48.
- [4] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London.
- [5] Gabriel, K.R. & Hall, W.J. (1983). Rerandomization inference on regression and shift effects: computationally feasible methods, *Journal of the American Statistical Association* **78**, 827–836.
- [6] Garthwaite, P.H. (1996). Confidence intervals from randomization tests, *Biometrics* **52**, 1387–1393.
- [7] Garthwaite, P.H. & Buckland, S.T. (1992). Generating Monte Carlo confidence intervals by the Robbins–Monro process, *Applied Statistics* **41**, 159–171.
- [8] Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Annals of Statistics* **16**, 927–963.
- [9] Maritz, J.S. (1995). *Distribution-Free Statistical Methods*, 2nd Edition, Chapman & Hall, London.
- [10] Pyburn, W.F. (1958). Size and movement of a local population of cricket frogs, *Texas Journal of Science* **10**, 325–342.
- [11] Robbins, H. & Monro, S. (1951). A stochastic approximation method, *Annals of Mathematical Statistics* **22**, 400–407.
- [12] Sternberg, D.E., Van Kammen, D.P. & Bunney, W.E. (1982). Schizophrenia: dopamine b-hydroxylase activity and treatment response, *Science* **216**, 1423–1425.

PAUL H. GARTHWAITE

Configural Frequency Analysis

ALEXANDER VON EYE

Volume 1, pp. 381–388

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Configural Frequency Analysis

Configural Frequency Analysis and the Person-orientation of Research

Cross-classification of categorical variables can be analyzed from two perspectives. The more frequently used one is *variable-oriented*. Researchers ask questions concerning the relationships among the variables that span the cross-classification. These questions concern, for example, association and dependency structures [5] (*see Contingency Tables*). However, a new perspective, contrasting with the variable-oriented perspective, has emerged since the beginning of the 1990s. This perspective is *person-oriented* [2, 19]. Under this perspective, researchers ask whether subgroups exist that show group-specific patterns of variable relationships.

Consider the following data example [16, 17]. In a study on the development of aggression in adolescents, Finkelstein, von Eye, and Preece [3] asked whether aggression could be predicted from physical pubertal development (PPD). PPD was measured using the Tanner measure. For the following purposes, we use the data from the 1985 and 1987 waves, when the adolescents were 13 and 15 years old. The Tanner measure was scaled to have four levels, with one indicating prepubertal and four indicating physically mature. For the following example, we use the data from 64 respondents. The cross-classification of the 1985 with the 1987 data appears in Table 1.

When looking at Table 1, we notice that, at neither point in time are there adolescents at the level of prepubertal physical development (category 1 does not appear in the table). This is not a surprise, considering the age of the respondents. The cross-classification contains, therefore, only 3×3 instead

of 4×4 cells. We now ask first, from a variable-oriented perspective, whether an association between the measures from 1985 and 1987 exists. We find a Pearson $X^2 = 24.10$, which indicates a significant association ($df = 4; p < 0.01$).

From a person-oriented perspective, we can ask different questions. For example, we ask whether there exist particular patterns that strongly contradict the null hypothesis of independence of the two observations and, thus, shed light on the developmental process. Using Configural Frequency Analysis (CFA; [13, 16, 17]), we show, later in this entry, that particular meaningful patterns stand out.

Concepts and Methods of Configural Frequency Analysis (CFA)

CFA is applied mostly in exploratory contexts. Using CFA, researchers ask whether particular cells stand out in the sense that they contain either more or fewer cases than expected based on *a priori* specified assumptions. If more cases are found than expected, the cell (also called *configuration*) that contains these cases is said to constitute a *CFA type*. If a cell contains fewer cases than expected, it is said to constitute a *CFA antitype*. The *a priori* assumptions are specified in the form of a *base model* (explained in the next paragraph). The null hypothesis for the decision as to whether a cell constitutes a type or an antitype is $H_0 : E[m_i] = e_i$, where i indexes the cells of the cross-classification, e_i is the expected frequency for Cell i , m_i is the observed frequency for Cell i , and E is the expectancy.

To perform a CFA, researchers proceed in five steps. The following sections introduce these five steps.

CFA Base Models and Sampling Schemes

CFA Base Models. The first step involves the selection and specification of a *CFA base model*, and the estimation of the expected cell frequencies. Base models take into account all effects that are NOT of interest for the interpretation of types and antitypes. Types and antitypes then contradict these base models and reflect the processes that the researchers are interested in. Base models can be derived from a number of backgrounds. Most frequently, researchers specify base models in the form of log-frequency

Table 1 Cross-classification of the Tanner scores from 1985 and 1987

		Tanner scores in 1987			
		2	3	4	Totals
Tanner scores in 1985	2	3	17	15	35
	3	0	0	16	16
	4	0	0	13	13
	Totals	3	17	44	64

2 Configurational Frequency Analysis

models (see **Log-linear Models**), $\log E = X\lambda$, where E is the vector of model frequencies, X is the indicator matrix that contains the effects that define the base model, and λ is a parameter vector.

To give an example of a base model, consider the data in Table 1. In tables of this sort, researchers rarely need to ask the question whether an association between the earlier and the later observations exists. The X^2 -analysis confirms what everybody either knows or assumes: there is a strong association. It is, therefore, the goal of CFA to explore the cross-classification, and to identify those cells that deviate from the assumption of independence. These cells not only carry the association, they also define the developmental process that the researchers attempt to capture in this study. Later in this entry, we present a complete CFA of this table. In brief the main effects **log-linear model** of variable independence can be a CFA base model.

Another example of a log-linear CFA base model is that of Prediction CFA (P-CFA). This variant of CFA is used to identify patterns of predictor categories that go hand-in-hand with patterns of criteria categories. For example, one can ask whether particular patterns of categories that describe how students do (or do not do) their homework allow one to predict particular patterns of categories that describe the students' success in academic subjects. In this situation, the researchers are not interested in the associations among the predictors, and they are not interested in the associations among the criterion variables. Therefore, all these associations are part of the base model. It is saturated in the predictors, and it is saturated in the criteria. However, the researchers are interested in predictor-criterion relationships. Therefore, the base model proposes that the predictors are unrelated to the criteria. P-CFA types and antitypes indicate where this assumption is violated. P-CFA types describe where criterion patterns can be predicted from predictor patterns. P-CFA antitypes describe which criterion patterns do not follow particular predictor patterns.

A second group of base models uses prior information to determine the expected cell frequencies. Examples of prior information include population parameters, theoretical probabilities that are known, for instance, in coin toss or roulette experiments, or probabilities of transition patterns (see [16], Chapter 8).

A third group of base models determines expected cell frequencies based on distributional assumptions. For example, von Eye and Bogat [20] proposed estimating the expected cell probabilities based on the assumption that the categorized variables that span a cross-classification follow a multinormal distribution. CFA tests can then be used to identify those sectors that deviate from multinormality most strongly.

The first group of base models is log-linear. The latter two are not log-linear, thus illustrating the flexibility of CFA as a method of analysis of cross-classifications.

Sampling Schemes and Their Relation to CFA Base Models. When selecting a base model for CFA, first, the variable relationships the researchers are (not) interested in must be considered. This issue was discussed above. Second, the sampling scheme must be taken into account. The sampling scheme determines whether a base model is admissible (see **Sampling Issues in Categorical Data**). In the simplest case, sampling is multinomial (see **Catalogue of Probability Density Functions**), that is, cases are randomly assigned to all cells of the cross-classification. Multinomial sampling is typical in observational studies. There are no constraints concerning the univariate or multivariate marginal frequencies.

However, researchers determine marginal frequencies often before data collection. For example, in a comparison of smokers with nonsmokers, a study may sample 100 smokers and 100 nonsmokers. Thus, there are constraints on the marginals. If 100 smokers are in the sample, smokers are no longer recruited, and the sample of nonsmokers is completed. In this design, cases are no longer randomly assigned to the cells of the entire table. Rather, the smokers are randomly assigned to the cells for smokers, and the nonsmokers are assigned to the cells for nonsmokers. This sampling scheme is called *product-multinomial*. In univariate, product-multinomial sampling, the constraints are on the marginals of one variable. In multivariate, product-multinomial sampling, the constraints are on the marginals of multiple variables, and also on the cross-classifications of these variables. For example, the researchers of the smoker study may wish to include in their sample 50 female smokers from the age bracket between 20 and 30, 50 male smokers from the same age bracket, and so forth.

In log-linear modeling, the sampling scheme is of lesser concern, because it has no effect on the estimated parameters. In contrast, from the perspective of specifying a base model for CFA, taking the sampling scheme into account is of particular importance. In multinomial sampling, there are no constraints on the marginal probabilities. Therefore, any base model can be considered, as long as the estimated expected cell frequencies sum to equal the sample size. In contrast, if one or more variables are observed under a product-multinomial sampling scheme, constraints are placed on their marginal probabilities, and the marginal frequencies of these variables must be reproduced. This can be guaranteed by base models that include the main effects (and the interactions) of the (multiple) variables observed under the product-multinomial sampling scheme [16, 2003].

Concepts of Deviation from Independence

CFA types and antitypes result when the discrepancies between observed and expected frequencies are large. It is interesting to see that several definitions of such discrepancies exist [6]. Two of these definitions have been discussed in the context of CFA [22]. The definition that is used in the vast majority of applications of log-linear modeling and CFA is that of the *marginal-free* residual. This definition implies that the marginal probabilities are taken into account when the discrepancy between the observed and the expected cell probabilities is evaluated. For example, the routine standardized residual, $r_i = (m_i - e_i)/\sqrt{e_i}$, its equivalent $X^2 = r_i^2$, and the correlation ρ are marginal-dependent, and so is Goodman's [6] weighted log-linear interaction, which can be interpreted as the log-linear interaction, λ , with the marginal probabilities as weights. The second definition leads to measures that are *marginal-dependent*. This definition implies that the marginal probabilities have an effect on the magnitude of the measure that describes the discrepancy between the observed and the expected cell frequencies. Sample measures that are marginal-dependent are the log-linear interaction, λ , and the **odds ratio**, θ (see [21]). Von Eye et al. [22] showed that the pattern of types and antitypes that CFA can unearth can vary depending on whether marginal-free or marginal-dependent definitions of deviation from independence are used.

Selection of a Significance Test

Significance tests for CFA can be selected based on the following five characteristics:

- (1) *exact* versus *approximative* tests (see **Exact Methods for Categorical Data**); the **binomial test** is the most frequently used exact test;
- (2) statistical **power** [18, 23]; Lehman's [12] approximative hypergeometric test has the most power;
- (3) differential sensitivity of test to types and antitypes; most tests are more sensitive to types when samples are small, and more sensitive to antitypes when samples are large;
- (4) *sampling scheme*; any of the tests proposed for CFA can be used when sampling is multinomial; however, the exact and the approximative hypergeometric (see **Catalogue of Probability Density Functions**) tests (e.g., [12]) require product-multinomial sampling.
- (5) use for *inferential* or *descriptive* purposes; typically, tests are used for inferential purposes; odds ratios and correlations are often also used for descriptive purposes.

The following three examples of tests used in CFA illustrate the above characteristics. We discuss the binomial test, the Pearson X^2 component test, and Lehman's [12] approximative hypergeometric test. Let m_i denote the frequency that was observed for Cell i , and e_i the estimated expected cell frequency. Then, the Pearson X^2 test statistic is

$$X_i^2 = \frac{(m_i - e_i)^2}{e_i}. \quad (1)$$

This statistic is approximately distributed as χ^2 , with $df = 1$. For the binomial test, let the estimate of the probability of Cell i be $p_i = e_i/N$. The tail probability for m_i is then

$$b_i = \sum_{l=m_i}^N \binom{N}{l} p_i^l (1 - p_i)^{N-l} \quad (2)$$

For a CFA test of the form used for the Pearson component test, Lehman [12] derived the exact variance,

$$\sigma_i^2 = N p_i [1 - p_i - (N - 1)(p_i - \tilde{p}_i)], \quad (3)$$

4 Configural Frequency Analysis

where p_i is defined as for the binomial test, and

$$\tilde{p}_i = \frac{\prod_{j=1}^d (N_j - 1)}{(N - 1)^d}, \quad (4)$$

is where d is the number of variables that span the cross-classification, and j indexes these variables. Using the exact variance, the Lehmacher cell-specific test statistic can be defined as

$$z_i^L = \frac{m_i - e_i}{\sigma_i}. \quad (5)$$

Lehmacher's z is, for large samples, standard normally distributed. A continuity correction has been proposed that prevents Lehmacher's test from being nonconservative in small samples. The correction requires subtracting 0.5 from the numerator if $m > e$, and adding 0.5 to the numerator if $m < e$.

These three test statistics have the following characteristics:

- (1) The *Pearson X^2 component test* is an approximative test with average power. It is more sensitive to types when samples are small, and more sensitive to antitypes when samples are large. It can be applied under any sampling scheme, and is mostly used for inferential purposes.
- (2) The *binomial test* is exact, and can be used under any sampling scheme.
- (3) *Lehmacher's test* is approximative. It has the most power of all tests that have been proposed for use in CFA. It is more sensitive to types when samples are small, and more sensitive to antitypes when samples are large. The only exceptions are **2 × 2 tables**, where Lehmacher's test always identifies exactly the same number of types and antitypes. The test requires that the sampling scheme be product-multinomial, and can be used only when the CFA base model is the log-linear main effect model.

Protection of the Significance Threshold α

Typically, CFA applications are exploratory and examine all cells of a cross-classification. This strategy implies that the number of tests performed on the same data set can be large. If the number of tests is large, the α -level needs to be protected for two reasons (see **Multiple Comparison Procedures**). First,

there is *capitalizing on chance*. If the significance threshold is α , then researchers take the risk of committing an α -error at each occasion a test is performed. When multiple tests are performed, the risk of α comes with each test. In addition, there is a risk that the α -error is committed twice, three times, or, in the extreme case, each time a test is performed.

The second reason is that the tests in a CFA are dependent. Von Weber, Lautsch, and von Eye [23] showed that the results of three of the four tests that CFA can possibly perform in a 2×2 table are completely dependent on the results of the first test. In larger tables, the results of each cell-specific test are also dependent on the results of the tests performed before, but to a lesser extent. In either case, the α -level needs to be protected for the CFA tests to be valid.

A number of methods for α -protection has been proposed. The most popular and simplest method, termed *Bonferroni adjustment* (see **Multiple Comparison Procedures**), yields an adjusted significance level, α^* , that (a) is the same for each test, and (b) takes the total number of tests into account. The protected α -level is $\alpha^* = \alpha/c$, where c is the total number of tests performed. For example, let a cross-classification have $2 \times 3 \times 3 \times 2 = 36$ cells. For this table, one obtains the Bonferroni-protected $\alpha^* = 0.0014$. Obviously, this new, protected threshold is extreme, and it will be hard to find types and antitypes.

Therefore, less radical procedures for α protection have been devised. An example of these is Holm's [9] procedure. This approach takes into account (1) the maximum number of tests to be performed, and (2) the number of tests already performed before test i . In contrast to the Bonferroni procedure, the protected significance threshold α_i^* is not constant. Specifically, the Holm procedure yields the protected significance level

$$\alpha_i^* = \frac{\alpha}{c - i + 1}, \quad (6)$$

where i indexes the i th test that is performed. Before applying Holm's procedure, the test statistics have to be rank-ordered such that the most extreme statistic is examined first. Consider the first test; for this test, the Holm-protected α is $\alpha^* = \alpha/(c - 1 + 1) = \alpha/c$. This threshold is identical to the one used for the first test under the Bonferroni procedure. However, for the second test under Holm, we obtain $\alpha^* = \alpha/(c - 1)$, a

threshold that is less extreme and prohibitive than the one used under Bonferroni. For the last cell in the rank order, we obtain $\alpha^* = \alpha/(c - c + 1) = \alpha$. Testing under the Holm procedure concludes after the first null hypothesis prevails. More advanced methods of α -protection have been proposed, for instance, by Keselman, Cribbie, and Holland [10].

Identification and Interpretation of Types and Antitypes

Performing CFA tests usually leads to the identification of a number of configurations (cells) as constituting types or antitypes. The interpretation of these types and antitypes uses two types of information. First, types and antitypes are interpreted based on the *meaning of the configuration* itself. Consider, for example, the data in Table 1. Suppose Configuration 2-3 constitutes a type (below, we will perform a CFA on these data to determine whether this configuration indeed constitutes a type). This configuration describes those adolescents who develop from a Tanner stage 2 to a Tanner stage 3. That is, these adolescents make progress in their physical pubertal development. They are neither prepubertal nor physically mature, but they develop in the direction of becoming mature. Descriptions of this kind indicate the meaning of a type or antitype.

The second source of information is included in the decisions as they were processed in the five steps of CFA. Most important is the definition of the base model. If the base model suggests variable independence, as it does in the log-linear main effect base model, types and antitypes suggest local associations. If the base model is that of prediction CFA, types and antitypes indicate pattern-specific relationships between predictors and criteria. If the base model is that of two-group comparison, types indicate the patterns in which the two groups differ significantly.

The characteristics of the measure used for the detection of types and antitypes must also be considered. Measures that are marginal-free can lead to different harvests of types and antitypes than measures that are marginal-dependent.

Finally, the **external validity** of types and antitypes needs to be established. Researchers ask whether the types and antitypes that stand out in the space of the variables that span the cross-classification under study do also stand out, or

can be discriminated in the space of variables not used in the CFA that yielded them. For example, Görtelmeyer [7] used CFA to identify six types of sleep problems. To establish external validity, the author used ANOVA methods to test hypotheses about mean differences between these types in the space of personality variables.

Data Examples

This section presents two data examples.

Data example 1: Physical pubertal development.

The first example is a CFA of the data in Table 1. This table shows the cross-tabulation of two ratings of physical pubertal development in a sample of 64 adolescents. The second ratings were obtained two years after the first. We analyze these data using first order CFA. The base model of first order CFA states that the two sets of ratings are independent of each other. If types and antitypes emerge, we can interpret them as indicators of local associations or, from a substantive perspective, as transition patterns that occur more often or less often than expected based on chance.

To search for the types, we use Lehmacher's test with continuity correction and Holm's procedure of α -protection. Table 2 presents the results of CFA.

The overall Pearson X^2 for this table is 24.10 ($df = 4$; $p < 0.01$), indicating that there is an association between the two consecutive observations of physical pubertal development. Using CFA, we now ask whether there are particular transition patterns that stand out. Table 2 shows that three types and two antitypes emerged. Reading from the top of the table, the first type is constituted by Configuration 2-3. This configuration describes adolescents who develop from an early adolescent pubertal stage to a late adolescent pubertal stage. Slightly over nine cases had been expected to show this pattern, but 17 were found, a significant difference. The first antitype is constituted by Configuration 2-4. Fifteen adolescents developed so rapidly that they leaped one stage in the one-year interval between the observations, and showed the physical development of a mature person. However, over 24 had been expected to show this development. From this first type and this first antitype, we conclude that the development by one stage is normative, and the development by more

6 Configural Frequency Analysis

Table 2 First order CFA of the pubertal physical development data in Table 1

Wave 85 87	Frequencies		Lehmacher's z	$p(z)^a$	Holm's α^*	Type/antitype?
	Observed	Expected				
22	3	1.64	1.013	.1555	.01666	
23	17	9.30	4.063	$<\alpha^*$	.00625	T
24	15	24.06	-4.602	$<\alpha^*$	.00555	A
32	0	0.75	-0.339	.3674	.025	
33	0	4.25	-2.423	.0075	.00833	A
34	16	11.00	2.781	.0027	.00714	T
42	0	0.61	-0.160	.4366	.05	
43	0	3.45	-2.061	.0196	.0125	
44	13	8.94	2.369	.0089	.01	T

^a $<\alpha^*$ indicates that the first four decimals are zero.

than one stage can be observed, but less often than chance.

If development by one stage is normative, the second transition from a lower to a higher stage in Table 2, that is, the transition described by Configuration 3-4, may also constitute a type. The table shows that this configuration also contains significantly more cases than expected.

Table 2 contains one more antitype. It is constituted by Configuration 3-3. This antitype suggests that it is very unlikely that an adolescent who has reached the third stage of physical development will still be at this stage two years later. This lack of stability, however, applies only to the stages of development that adolescents go through before they reach the mature physical stage. Once they reach this stage, development is completed, and stability is observed. Accordingly, Configuration 4-4 constitutes a type.

Data example 2. The second data example presents a reanalysis of a data set published by [4]. A total of 181 high school students processed the 24 items of a cube comparison task. The items assess the students' spatial abilities. After the cube task, the students answered questions concerning the strategies they had used to solve the task. Three strategies were used in particular: mental rotation (R), pattern comparison (P), and change of viewpoint (V). Each strategy was scored as either 1 = not used, or 2 = used. In the following sample analysis, we ask whether there are gender differences in strategy use. Gender was scaled as 1 = females and 2 = males. The analyses are performed at the level of individual responses.

The base model for the following analyses (1) is saturated in the three strategy variables that are used to distinguish the two gender groups, and (2) assumes that there is no relationship between gender and strategies used. If discrimination types emerge, they indicate the strategy patterns in the two gender groups differ significantly. For the analyses, we use Pearson's X^2 -test, and the Bonferroni-adjusted $\alpha^* = .05/8 = 0.00625$. Please note that the numerator for the calculation of the adjusted α is 8 instead of 16, because, to compare the gender groups, one test is performed for each strategy pattern (instead of two, as would be done in a one-sample CFA). Table 3 displays the results of 2-group CFA.

The results in Table 3 suggest strong gender differences. The first discrimination type is constituted by Configuration 111. Twenty-five females and five males used none of the three strategies. This difference is significant. The second discrimination type is constituted by Configuration 122. Thirteen males and 63 females used the pattern comparison and the viewpoint change strategies. Discrimination type 221 suggests that the rotational and the pattern comparison strategies were used by females in 590 instances, and by males in 872 instances. The last discrimination type emerged for Configuration 222. Females used all three strategies in 39 instances; males used all three strategies in 199 instances.

We conclude that female high school students differ from male high school students in that they either use no strategy at all, or use two. If they use two strategies, the pattern comparison strategy is one of them. In contrast, male students use all

Table 3 Two-group CFA of the cross-classification of rotation strategy (R), pattern comparison strategy (P), viewpoint strategy (V), and gender (G)

Configuration	m	e	statistic	p	Type?
1111	25	11.18			
1112	5	18.82	27.466	.000000	Discrimination Type
1121	17	21.99			
1122	42	37.01	1.834	.175690	
1211	98	113.29			
1212	206	190.71	3.599	.057806	
1221	13	28.69			
1222	64	48.31	13.989	.000184	Discrimination Type
2111	486	452.78			
2112	729	762.22	5.927	.014911	
2121	46	52.55			
2122	95	88.45	1.354	.244633	
2211	590	544.83			
2212	872	917.17	10.198	.001406	Discrimination Type
2221	39	88.69			
2222	199	149.31	47.594	.000000	Discrimination Type

three strategies significantly more often than female students.

Summary and Conclusions

CFA has found applications in many areas of the social sciences, recently, for instance, in research on child development [1, 14], and educational psychology [15]. Earlier applications include analyses of the control beliefs of alcoholics [11]. CFA is the method of choice if researchers test hypotheses concerning *local associations* [8], that is, associations that hold in particular sectors of a cross-classification only. These hypotheses are hard or impossible to test using other methods of categorical data analysis.

Software for CFA can be obtained free from von-eye@msu.edu. CFA is also a module of the software package SLEIPNER, which can be downloaded, also free, from <http://www.psychology.su.se/sleipner/>.

References

- [1] Aksan, N., Goldsmith, H.H., Smider, N.A., Essex, M.J., Clark, R., Hyde, J.S., Klein, M.H. & Vandell, D.L. (1999). Derivation and prediction of temperamental types among preschoolers, *Developmental Psychology* **35**, 958–971.
- [2] Bergman, L.R. & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology, *Development and Psychopathology* **9**, 291–319.
- [3] Finkelstein, J.W., von Eye, A. & Preece, M.A. (1994). The relationship between aggressive behavior and puberty in normal adolescents: a longitudinal study, *Journal of Adolescent Health* **15**, 319–326.
- [4] Glück, J. & von Eye, A. (2000). Including covariates in configural frequency analysis. *Psychologische Beiträge* **42**, 405–417.
- [5] Goodman, L.A. (1984). *The Analysis of Cross-Classified Data Having Ordered Categories*, Harvard University Press, Cambridge.
- [6] Goodman, L.A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data, *Journal of the American Statistical Association* **86**, 1085–1111.
- [7] Görtelmeyer, R. (1988). *Typologie Des Schlafverhaltens [A Typology of Sleeping Behavior]*, S. Roderer Verlag, Regensburg.
- [8] Havránek, T. & Lienert, G.A. (1984). Local and regional versus global contingency testing. *Biometrical Journal* **26**, 483–494.
- [9] Holm, S. (1979). A simple sequentially rejective multiple Bonferroni test procedure, *Biometrics* **43**, 417–423.

8 Configural Frequency Analysis

- [10] Keselman, H.J., Cribbie, R. & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: an alternative approach to familywise and comparison-wise type I error control, *Psychological Methods* **4**, 58–69.
- [11] Krampen, G., von Eye, A. & Brandtstädter, J. (1987). Konfigurationstypen generalisierter Kontrollüberzeugungen. *Zeitschrift für Differentielle und Diagnostische Psychologie* **8**, 111–119.
- [12] Lehman, W. (1981). A more powerful simultaneous test procedure in configural frequency analysis, *Biometrical Journal* **23**, 429–436.
- [13] Lienert, G.A. & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types, *Educational and Psychological Measurement* **35**, 231–238.
- [14] Mahoney, J.L. (2000). School extracurricular activity participation as a moderator in the development of antisocial patterns, *Child Development* **71**, 502–516.
- [15] Spiel, C. & von Eye, A. (2000). Application of configural frequency analysis in educational research. *Psychologische Beiträge* **42**, 515–525.
- [16] von Eye, A. (2002a). *Configural Frequency Analysis – Methods, Models, and Applications*, Lawrence Erlbaum, Mahwah.
- [17] von Eye, A. (2002b). The odds favor antitypes – a comparison of tests for the identification of configural types and antitypes, *Methods of Psychological Research – online* **7**, 1–29.
- [18] von Eye, A. (2003). A comparison of tests used in 2×2 tables and in two-sample CFA. *Psychology Science* **45**, 369–388.
- [19] von Eye, A. & Bergman, L.R. (2003). Research strategies in developmental psychopathology: dimensional identity and the person-oriented approach, *Development and Psychopathology* **15**, 553–580.
- [20] von Eye, A. & Bogat, G.A. (2004). Deviations from multivariate normality, *Psychology Science* **46**, 243–258.
- [21] von Eye, A. & Mun, E.-Y. (2003). Characteristics of measures for 2×2 tables, *Understanding Statistics* **2**, 243–266.
- [22] von Eye, A., Spiel, C. & Rovine, M.J. (1995). Concepts of nonindependence in configural frequency analysis, *Journal of Mathematical Sociology* **20**, 41–54.
- [23] von Weber, S., von Eye, A., & Lautsch, E. (2004). The type II error of measures for the analysis of 2×2 tables, *Understanding Statistics* **3**(4), pp. 259–232.

ALEXANDER VON EYE

Confounding in the Analysis of Variance

RICHARD S. BOGARTZ

Volume 1, pp. 389–391

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Confounding in the Analysis of Variance

Unintentional Confounding

In the analysis of variance (ANOVA), we usually study the effects of categorical independent variables on a numerical-valued dependent variable. Two independent variables are confounded if one is *found with* the other and the two as they stand are inseparable. This means that each time one variable has a given value, the other variable has a matched value. As a consequence, we lose the ability to estimate separately the effects of these two variables or the interaction of these variables with other variables. Conclusions as to which variable had an effect become impossible without making some additional assumption or taking some additional measure.

Consider visual habituation in infants. With repeated presentations of a visual stimulus, infants eventually look at it less and less. The standard interpretation is that the infant forms a representation of the stimulus and that looking decreases as the representation requires less and less correction from the visual input. But habituation requires time in the experiment. So, the number of habituation trials is confounded with time in the experiment. How do we know that the diminished looking is not due to fatigue that might naturally increase with more time in the experiment? One standard tactic is to present a novel stimulus at the conclusion of the habituation trials. Renewed looking to the novel stimulus accompanied by decreased looking to the familiar one argues against fatigue and in favor of the habituation process.

In between-subject designs, we obtain a single measure from a subject. In repeated measures designs (see **Repeated Measures Analysis of Variance; Longitudinal Data Analysis**), we measure the subject repeatedly. Each approach carries its own dangers of confounding and its own methods of remedy.

Between-subjects Designs

Either we run all the subjects at once or some subjects are run before others, either individually or in batches. Almost always, running individual subjects is preferable. With running subjects individually or

in batches, order effects may be confounded with the treatment effects. Drift or refinement in the experimental procedure, change of experimenter, of season, and so on, all may be confounded with the treatment effects. Protection from such confounding can be achieved by randomly assigning the subjects to the conditions and by randomly assigning the order of running the subjects (see **Randomization**). The subjects are still confounded with the experimental conditions since each experimental condition entails scores from some subjects and not from others. The random assignment of subjects to conditions protects against but does not remove such confounding.

Repeated Measures Designs

Different risks of confounding occur in repeated measures designs. A subject is measured on each of a series of occasions or trials. Ideally, to compare the effects of one treatment versus another we would like the subject to be identical on each occasion. But if a subject has been subjected to an earlier treatment, there is always the possibility that a residual effect of the earlier treatment remains at the time of the later one. Thus, the residual effect of the earlier treatment may not be separable from the effect of the later treatment.

Sometimes, it is that very accumulation of effect that is the object of study as in the habituation example. In such cases, the inseparability of the effect of a particular occasion from the accumulation of previous effects ordinarily is not a problem and we do not speak of confounding. Such separation as can be accomplished is usually in terms of a mathematical model of the process that assigns an effect of a given occasion in terms of the present event and the previous effects. Clarification of the separate effects then becomes a matter of the **goodness of fit** of the model.

On the other hand, often there are risks of cumulative effects that are both unintended and undesirable. The amount of time in the experiment can produce trial-to-trial effects such as warm-up, fatigue, and increased sensitivity to distraction. Also, amount of time can result in improved skill at performing the task with a consequent freeing up of mental processing space so that the subject on later trials has more freedom to devote to formulating conjectures about the purpose of the experiment, the motivation of the

2 Confounding in the Analysis of Variance

experimenter, and so on, which in turn may affect performance.

Another form of confounding comes not from a progressive buildup over trials but simply from the carryover of the effect of one trial to the response on the next. We can imagine an experiment on taste sensitivity in which the subject repeatedly tastes various substances. Obviously, it is important to remove traces of the previous taste experience before presenting the next one.

Mixed Designs

Mixed designs combine the features of between-subject designs with those of repeated measures designs (*see Fixed and Random Effects*). Different sets of subjects are given the repeated measures treatments with one and only one value of a between-subjects variable or one combination of between-subject variables. In this type of design, both types of confounding described above can occur. The between-subject variables can be confounded with groups of subjects and the within-subject variables can be confounded with trial-to-trial effects.

Intentional Planned Confounding or Aliasing

If when variables are confounded we cannot separate their effects, it may be surprising that we would ever intentionally confound two or more variables. And yet, there are circumstances when such confounding can be advantageous. In general, we do so when we judge the gains to exceed the losses. Research has costs. The costs involve resources such as time, money, lab space, the other experiments that might have been done instead, and so on. The gains from an experiment are in terms of information about the way the world is and consequent indications of fruitful directions for more research. Every experiment is a trade-off of resources against information.

Intentional confounding saves resources at the cost of information. When the information about some variable or interaction of variables is judged to be not worth the resources it takes to gain the information, confounding becomes a reasonable choice. In some cases, the loss of information is free of cost if we have no interest in that information. We have already seen an example of this in the case of between-subject

designs where subjects are confounded with treatment conditions. We have elected to ignore information about individual subjects in favor of focusing on the treatment effects.

In a design, there are groups of subjects, there are treatments, and there are interactions. Any two of these can be confounded with each other. So, for example, we have the split-plot design in which groups are confounded with treatments, the confounded factorial designs in which groups are confounded with **interactions**, and the fractional replication designs in which treatments are confounded with interactions (*see Balanced Incomplete Block Designs*). In the case of the latter two designs, there may be interactions that are unimportant or in which we have no interest. We can sometimes cut the cost of running subjects in half by confounding such an interaction with groups or with treatments. Kirk [3] provides an extensive treatment of designs with confounding. The classic text by Cochran and Cox [2] is also still an excellent resource.

An Example of Intentional Confounding

An example of confounding groups of subjects with a treatment effect can be seen in the following experimental design for studying an infant looking in response to possible and impossible arithmetic manipulations of objects [1]. Two groups of infants are given four trials in the following situation: An infant is seated in front of a stage on which there are two lowered screens, one on the left, the other on the right. When the screen is lowered, the infant can see what is behind it, and when the screen is raised the infant cannot see. An object is placed behind the lowered left screen, the screen is raised, either one or two additional objects are placed behind the raised left screen, an object is placed behind the lowered right screen, the screen is raised, either one or two additional objects are placed behind the raised right screen. Thus, each screen conceals either two or three objects. One of the two screens is then lowered and either two or three objects are revealed. The duration of the infant looking at the revealed objects is the basic measure. The trials are categorized as possible or impossible, revealing two or three objects, and involving a primacy or recency effect. On possible trials, the same number hidden behind the screen is revealed; on impossible, a different

number is revealed. Primacy refers to revealing what was behind the left screen since it was hidden first; recency refers to revealing the more recently hidden objects behind the right screen. Number revealed of course refers to revealing two or revealing three objects. Thus, with two objects hidden behind the left screen and three objects finally revealed behind the left screen, we have a combination of the three, impossible, and primacy effects. A complete design would require eight trials for each infant: a $2 \times 2 \times 2$ factorial design of number $\times$ possibility $\times$ recency. Suppose previous experience indicates that the full eight trials would probably result in declining performance with 10-month-old infants. So the design in Table 1 is used:

Inspection of the table shows that the primacy effect is perfectly confounded with the possibility $\times$ number revealed $\times$ group interaction. In the present context, a primacy effect seems reasonable in that the infants might have difficulty remembering what was placed behind the first screen after being exposed to the activity at the second screen. On the other

hand, it seems unlikely that a possibility $\times$ number revealed $\times$ group interaction would exist, especially since infants are assigned to groups at random and the order of running infants from the different groups is randomized. In this context, a significant group $\times$ possibility $\times$ number interaction is reasonably interpreted as a recency effect.

The design described above contains a second confound. Primacy-recency is confounded with spatial location of the screen (left or right). The first concealment was always behind the left screen; the second, always behind the right. This means that preferring to look longer to the right than to the left would result in an apparent recency preference that was in fact a position preference. If such a position preference were of concern, the design would have to be elaborated to include primacy on the left for some infants and on the right for others.

References

- [1] Cannon, E.N. & Bogartz, R.S. (2003, April). Infant number knowledge: a test of three theories, in *Poster Session Presented at the Biennial Meeting of the Society for Research in Child Development*, Tampa.
- [2] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Edition, Wiley, New York.
- [3] Kirk, R.E. (1995). *Experimental Design*, 3rd Edition, Brooks/Cole, Pacific Grove.

RICHARD S. BOGARTZ

Table 1 Confounding of the primacy effect with the possibility $\times$ number revealed $\times$ group interaction

	Possible		Impossible	
	2 revealed	3 revealed	2 revealed	3 revealed
Group 1	Primacy	Recency	Recency	Primacy
Group 2	Recency	Primacy	Primacy	Recency

Confounding Variable

PATRICK ONGHENA AND WIM VAN DEN NOORTGATE

Volume 1, pp. 391–392

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Confounding Variable

Consider testing a causal model with an explanatory variable A and a response variable B . In testing such a causal model, a confounding variable C is a variable that is related to the explanatory variable A and at the same time has an effect on the response variable B that is mixed up with the effect of the explanatory variable A [3]. The existence of this confounding variable makes it difficult to assess the size of the effect of the explanatory variable or sometimes even calls the causal link between the explanatory variable and the response variable into question. Confounding variables are able to inflate as well as shrink the observed covariation (*see Covariance/variance/correlation*) between explanatory and response variables, and in some cases, the confounding variable can be held responsible for all of the observed covariation between the explanatory and the response variable (cases of so-called spurious correlation, [1]). Reasons for confounding can be found in the type of research that is being conducted – and that can make it difficult to avoid confounding – or in unanticipated systematic errors in experimental design.

For example, in school effectiveness research, one is interested in the effect of schools and educational practices (A) on student achievement (B). In one Belgian study, 6411 pupils entering one of 57 schools were followed during their secondary schooling career (between ages 12 and 18 yr) using questionnaires related to their language achievement, mathematical skills, attitudes, well-being, and other psychosocial variables [6]. If one wants to establish a causal link between the (type of) school the pupil is attending (A) and his later achievement (B), then student achievement on entrance of the school (C) is an obvious confounding variable. C is related to A because students are not randomly assigned to schools (some schools attract higher-achieving students than others) and at the same time C has a strong link to B because of the relative stability of student achievement during secondary education [6]. It is therefore likely that at least a part of the relation between the school (A) and the later achievement (B) is explained by the initial achievement (C) rather than to a causal relationship between A and B .

Of course, confounding can also be due to more than one variable and can also occur when more complex causal models are considered. In an observational study, like the school effectiveness study just mentioned, a large number of confounding variables could be present. Besides prior achievement, also socioeconomic background, ethnicity, motivation, intelligence, and so on, are potential confounding variables, which would have to be checked before any causal claim can reasonably be made [6].

Although confounding is most evident in the context of **observational studies** and **quasi-experiments** [4, 5], also in randomized experiments one has to be careful for unexpected confounding variables. A striking example can be found in the infamous Pepsi versus Coca-Cola taste comparisons. Alleged Coca-Cola drinkers were given the blind choice between a glass marked Q (containing Coca-Cola) and a glass marked M (containing Pepsi). Much to the surprise of the Coca-Cola officials, significantly more tasters preferred the glass containing Pepsi. However, in a control test it was found that blind tasters preferred the glass marked M even if both glasses contained Coca-Cola. Apparently, in the first study, letters were a confounding variable. People seem to have a preference for high-frequent letters, implying a preference for the drink that was (accidentally?) associated with the most frequent letter [2].

Confounding variables are detrimental to the **internal validity** of any empirical study, and much of the current work on observational and quasi-experimental research deals with methods to avoid or to rule out confounding variables, or to model them as covariates. Excellent technical accounts can be found in [4] and [5].

References

- [1] Cohen, J., Cohen, P. & West, S.G. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Erlbaum, Mahwah.
- [2] Martin, D.W. (2004). *Doing Psychology Experiments*, 6th Edition, Brooks/Cole, Pacific Grove.
- [3] Moore, D.S. & McCabe, G.P. (2003). *Introduction to the Practice of Statistics*, 4th Edition, Freeman, New York.
- [4] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.
- [5] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

2 **Confounding Variable**

- [6] Van Damme, J., De Fraine, B., Van Landeghem, G., Opdenakker, M.-C. & Onghena, P. (2002). A new study on educational effectiveness in secondary schools in Flanders: an introduction, *School Effectiveness and School Improvement* **13**, 383–397.

(See also **Confounding in the Analysis of Variance**)

PATRICK ONGHENA AND
WIM VAN DEN NOORTGATE

Contingency Tables

BRIAN S. EVERITT

Volume 1, pp. 393–397

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Contingency Tables

Contingency tables result from cross-classifying a sample from some population with respect to two or more qualitative (categorical) variables. The cells of such tables contain frequency counts of outcomes. Contingency tables are very common in many areas, for example, all branches of epidemiology, medicine, in particular, psychiatry, and in the social sciences, and psychology. Examples of various types of contingency tables are shown in Tables 1, 2, and 3.

The data in Table 1 come from [6], and arise from classifying 21 accounts of threatened suicide by jumping from high structures, according to the time of year and whether jeering or baiting behavior occurred amongst the onlookers. Interest lies in the question of whether baiting is more likely to occur in warm weather (the data come from the northern hemisphere, so June to September are the warm months).

Table 2 is taken from the 1991 General Social Survey and is also reported in [1]. Here, race and party identification are cross classified.

Table 3 is reported in [5], and involves data from a study of seriously emotionally disturbed (SED) and learning disabled (LD) adolescents.

Tables 1 and 2 are examples of *two-dimensional contingency tables*, that is, data cross-classified with respect to two categorical variables, and Table 3 is a *four-way contingency table*. Table 1, in which each variable has two categories, is also generally known as a **two-by-two contingency table**.

Table 1 Crowds and threatened suicide

Period	Crowd	
	Baiting	Nonbaiting
July – September	8 (<i>a</i>)	4 (<i>b</i>)
October – May	2 (<i>c</i>)	7 (<i>d</i>)

Table 2 Party identification and race (from the 1991 General Social Survey)

Race	Party identification		
	Democrat	Independent	Republican
White	341	105	405
Black	103	15	11

Table 3 Depression in adolescents

Age	Group	Sex	Depression	
			Low	High
12–14	LD	Male	79	18
		Female	34	14
	SED	Male	14	5
		Female	5	8
15–16	LD	Male	63	10
		Female	26	11
	SED	Male	32	3
		Female	15	7
17–18	LD	Male	36	13
		Female	16	1
	SED	Male	36	5
		Female	12	2

SED: Seriously emotionally disturbed. LD: Learning disabled.

Testing for Independence in Two-dimensional Contingency Tables

The hypothesis of primary interest for two-dimensional contingency tables is whether the two variables involved are independent. This may be formulated more formally in terms of p_{ij} , the probability of an observation being in the ij th cell of the table, $p_{i.}$, the probability being in the i th row of the table, and $p_{.j}$, the probability being in the j th column of the table. The null hypothesis of independence can now be written:

$$H_0: p_{ij} = p_{i.} \times p_{.j} \quad (1)$$

Estimated values of $p_{i.}$ and $p_{.j}$ can be found from the relevant marginal totals ($n_{i.}$ and $n_{.j}$) and overall sample size (N) as

$$\hat{p}_{i.} = \frac{n_{i.}}{N}, \hat{p}_{.j} = \frac{n_{.j}}{N} \quad (2)$$

and these can then be combined to give the estimated probability of being in the ij th cell of the table under the hypothesis of independence, namely, $\hat{p}_{i.} \times \hat{p}_{.j}$. The frequencies to be expected under independence, E_{ij} , can then be obtained simply as

$$E_{ij} = N \times \hat{p}_{i.} \times \hat{p}_{.j} = \frac{n_{i.} \times n_{.j}}{N} \quad (3)$$

Independence can now be assessed by comparing the observed (O_{ij}) and estimated expected frequencies

2 Contingency Tables

(E_{ij}) using the familiar test statistic;

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4)$$

where r is the number of rows and c the number of columns in the table. Under H_0 , the distribution of X^2 is, for large N , approximately a *chi-squared distribution* (see **Catalogue of Probability Density Functions**) with $(r-1)(c-1)$ degrees of freedom. Hence, for a significance test of H_0 with approximate significance level α , we reject H_0 if

$$X^2 \geq \chi^2(\alpha)_{(r-1)(c-1)} \quad (5)$$

where $\chi^2(\alpha)_{(r-1)(c-1)}$ is the upper α point of a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

In the case of a two-by-two contingency table with cell frequencies a, b, c , and d , (see Table 1), the X^2 statistic can be written more simply as

$$X^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (6)$$

And in such tables, the independence hypothesis is equivalent to that of the equality of two probabilities, for example, in Table 1, that the probability of crowd baiting in June–September is the same as the probability of crowd baiting in October–May.

Applying the chi-squared test of independence to Table 2 results in a test statistic of 79.43 with two degrees of freedom. The associated P value is very small, and we can conclude with some confidence that party identification and race are not independent. We shall return for a more detailed look at this result later.

For the data in Table 1, the chi-square test gives $X^2 = 4.07$ with a single degree of freedom and a P value of 0.04. This suggests some weak evidence for a difference in the probability of baiting in the different times of the year. But the frequencies in Table 1 are small, and we need to consider how this might affect an asymptotic test statistic such as X^2 .

Small Expected Frequencies

The derivation of the chi-square distribution as an approximation for the distribution of the X^2 statistic is made under the rather vague assumption that the expected values are not ‘too small’. This has, for

many years, been interpreted as implying that all the expected values in the table should be greater than five for the chi-square test to be strictly valid. Since in Table 1 the four expected values are 5.7, 6.2, 4.3, and 4.7, this would appear to shed some doubt on the validity of the chi-squared test for the data and on the conclusion from this test. But as long ago as 1954, Cochran [3] pointed out that such a ‘rule’ is too stringent and suggested that if relatively few expected values are less than five (say one cell in five), a minimum value of one is allowable.

Nevertheless, for small, sparse data sets, the asymptotic inference from the chi-squared test may not be valid, although it is usually difficult to identify *a priori* whether a given data set is likely to give misleading results. This has led to suggestions for alternative test statistics that attempt to make the asymptotic P value more accurate. The best known of these is **Yates’ correction**. But, nowadays, such procedures are largely redundant since *exact* P values can be calculated to assess the hypothesis of independence; for details, see the **exact methods for categorical data** entry.

The availability of such ‘exact methods’ also makes the pooling of categories in contingency tables to increase the frequency in particular cells, unnecessary. The procedure has been used almost routinely in the past, but can be criticized on a number of grounds.

- A considerable amount of information may be lost by the combination of categories, and this may detract greatly from the interest and usefulness of the study.
- The randomness of the sample may be affected; the whole rationale of the chi-squared test rests on the randomness of the sample and in the categories into which the observations may fall being chosen in advance.
- Pooling categories after the data are seen may affect the random nature of the sample with unknown consequences.
- The manner in which categories are pooled can have an effect on the resulting inferences.

As an example, consider the data in Table 4 from [2]. When this table is tested for independence using the chi-squared test, the calculated significance level is 0.086, which agrees with the exact probability

Table 4 Hypothetical data from Baglivo [2]

	Column				
	1	2	3	4	5
Row 1	2	3	4	8	9
Row 2	0	0	11	10	11

to two significant figures, although a standard statistical package issues a warning of the form ‘some of the expected values are less than two, the test may not be appropriate’. If the first two columns of Table 4 are ignored, the P value becomes 0.48, and if the first two columns are combined with the third, the P value becomes one.

The practice of combining categories to increase cell size should be avoided and is nowadays unnecessary.

Residuals

In trying to identify which cells of a contingency table are primarily responsible for a significant overall chi-squared value, it is often useful to look at the differences between the observed values and those values expected under the hypothesis of independence, or some function of these differences. In fact, looking at **residuals** defined as observed value – expected value would be very unsatisfactory since a difference of fixed size is clearly more important for smaller samples. A more appropriate residual would be r_{ij} , given by:

$$r_{ij} = \frac{(n_{ij} - E_{ij})}{\sqrt{E_{ij}}} \quad (7)$$

These terms are usually known as *standardized residuals* and are such that the chi-squared test statistic is given by

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c r_{ij}^2 \quad (8)$$

It is tempting to think that the size of these residuals may be judged by comparison with standard normal percentage points (for example ± 1.96). Unfortunately, it can be shown that the variance of r_{ij} is always less than equal to one, and in some cases, the variance may be *considerably* less

than one. Consequently, the use of standardized residuals in the detailed examination of a contingency table may often give a conservative indication of a cell’s lack of fit to the independence hypothesis.

An improvement over standardized residuals is provided by the *adjusted residuals* (d_{ij}) suggested in [4], and defined as:

$$d_{ij} = \frac{r_{ij}}{\sqrt{[(1 - n_{i.}/N)(1 - n_{.j}/N)]}} \quad (9)$$

When the variables forming the contingency table are independent, the adjusted residuals are approximately normally distributed with mean zero and standard deviation one.

Returning to the data in Table 2, we can now calculate the expected values and then both the standardized and adjusted residuals – see Table 5. Clearly, the lack of independence of race and party identification arises from the excess of blacks who identify with being Democrat and the excess of whites who identify with being Republican.

Table 5 Expected values, standardized residuals, and adjusted residuals for data in Table 2

(1) Expected values

Race	Party identification		
	Democrat	Independent	Republican
White	385.56	104.20	361.24
Black	58.44	15.80	54.76

(2) Standardized residuals

Race	Party identification		
	Democrat	Independent	Republican
White	−2.27	0.08	2.30
Black	5.83	−0.20	−5.91

(3) Adjusted residuals

Race	Party identification		
	Democrat	Independent	Republican
White	−8.46	0.23	8.36
Black	8.46	−0.23	−8.36

In many cases, an informative way of inspecting residuals is to display them graphically using **correspondence analysis** (see **Configural Frequency Analysis**).

Higher-Dimensional Contingency Tables

Three- and higher-dimensional contingency tables arise when a sample of individuals is cross-classified with respect to three (or more) qualitative variables. A four-dimensional example appears in Table 3.

The analysis of such tables presents problems not encountered with two-dimensional tables, where a single question is of interest, namely, that of the independence or otherwise of the two variables involved. In the case of higher-dimensional tables, the investigator may wish to test that some variables are independent of some others, that a particular variable is independent of the remainder or some more complex hypothesis. Again, however, the chi-squared statistic is used to compare observed frequencies with estimates of those to be expected under a particular hypothesis.

The simplest question of interest in a three-dimensional table, for example, is that of the mutual independence of the three variables; this is directly analogous to the hypothesis of independence in a two-way table, and is tested in an essentially equivalent fashion. Other hypotheses that might be of interest are those of the partial independence of a pair of variables, and the conditional independence of two variables for a given level of the third. A more involved hypothesis is that the association between

two of the variables is identical in all levels of the third.

For some hypotheses, expected values can be obtained directly from simple calculations on particular marginal totals. But this is not always the case, and for some hypotheses, the corresponding expected values have to be estimated using some form of iterative procedure – for details see [1].

Three- and higher-dimensional contingency tables are best analyzed using **log-linear models**.

References

- [1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.
- [2] Baglivo, J., Oliver, D. & Pagano, M. (1988). Methods for the analysis of contingency tables with large and small cell counts, *Journal of the American Statistical Association* **3**, 1006–1013.
- [3] Cochran, W.G. (1954). Some methods for strengthening the common chi-squared tests, *Biometrics* **10**, 417–477.
- [4] Haberman, S.J. (1973). The analysis of residuals in cross-classified tables, *Biometrics* **29**, 205–220.
- [5] Maag, J.W. & Behrens, J.T. (1989). Epidemiological data on seriously emotionally disturbed and learning disabled adolescents: reporting extreme depressive symptomatology, *Behavioural Disorders* **15**, 21–27.
- [6] Mann, L. (1981). The baiting crowd in episodes of threatened suicide, *Journal of Personality and Social Psychology*, **41**, 703–709.

(See also **Marginal Independence; Odds and Odds Ratios**)

BRIAN S. EVERITT

Coombs, Clyde Hamilton

JOEL MICHELL

Volume 1, pp. 397–398

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Coombs, Clyde Hamilton

Born: July 22, 1912, in New Jersey, USA.

Died: February 4, 1988, in Michigan, USA.

Clyde H. Coombs contributed to the development of modern **mathematical psychology**. His significance derived from a comprehensive vision of the place of mathematics in social and behavioral science, his ability to integrate diverse influences into that vision, and his willingness to locate novel applications of theoretical insights.

Coombs pioneered a radical approach to psychometrics. Initially struck by L.L. Thurstone's work, he came to believe that the traditional approach is based upon implausibly strong assumptions. His alternative was to derive quantitative conclusions from the qualitative structure of psychological data [1]. The best known example is his unfolding procedure, in which the underlying quantitative structure of a unidimensional stimulus set is (approximately) recovered from peoples' preferences, given the hypothesis of single-peaked preference functions. This hypothesis is that each person has a single point of maximum preference on the relevant dimension around which strength of preference decreases symmetrically. Thus, any person's preference ordering entails an ordering on differences. The idea of basing measurement upon qualitative relations between differences had been explored by mathematicians and economists, but was neglected in psychology. Coombs's original insight was to link this idea to individual differences in preference. Each person's preference ordering implies a distinct partial order on the set of interstimulus midpoints and since different people may have different points of maximum preference, the conjunction of these different partial orders implies an ordering upon the complete set of interstimulus distances.

The analogy between psychology and geometry is the heart of Coombs's general theory of data [2]. He represented wide classes of qualitative data by spatial structures. The combination of this analogy and the hypothesis of single-peaked preference functions gave Coombs a powerful analytical tool, applicable in areas as diverse as choices amongst gambles, political preferences, and conflict resolution. Coombs's thesis that qualitative data

may contain quantitative information proved enormously fruitful. Subsequent work in measurement theory (*see* **Measurement: Overview**), for example, the theories of **multidimensional scaling** and additive conjoint measurement, exploited the insight that qualitative data afford quantitative representations.

Coombs received his A.B. in 1935 and his M.A. in 1937 from the University of California, Berkeley. He then joined Thurstone at the University of Chicago, where he received his Ph.D. in 1940. For the next six years he worked as a research psychologist with the US War Department and subsequently moved to the Psychology Department at the University of Michigan in 1947. He remained at Michigan throughout his career, chairing its Mathematical Psychology Program. He spent visiting terms at the Laboratory of Social Relations at Harvard University, at the Center for Advanced Study in the Behavioral Sciences in Palo Alto, was a Fulbright Fellow at the University of Amsterdam, and following retirement, also taught at the Universities of Hamburg, Calgary, and Santa Barbara. He served as president of the Psychometric Society, was the first head of the Society for Mathematical Psychology, was an honorary fellow of the American Statistical Association, and was elected to the National Academy of Arts and Sciences and to the National Academy of Science.

Because of their fertility and generality, Coombs's contributions to mathematical psychology and psychological theory endure. However, equally important was his contribution to a fundamental change in attitude towards measurement. According to operationism, which had a profound impact upon psychology, measurement is prior to theory in science. According to Coombs, measurement is inextricably theoretical and, importantly, a measurement theory may be false, which means that 'not everything that one would like to measure is measurable' ([3], page 39).

References

- [1] Coombs, C.H. (1950). Psychological scaling without a unit of measurement, *Psychological Review* **57**, 145–158.
- [2] Coombs, C.H. (1964). *A Theory of Data*, Wiley, New York.

2 Coombs, Clyde Hamilton

- [3] Coombs, C.H. (1983). *Psychology and Mathematics: an Essay on Theory*, The University of Michigan Press, Ann Arbor. (See also **Psychophysical Scaling**)

JOEL MICHELL

Correlation

DIANA KORNBROT

Volume 1, pp. 398–400

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Correlation

Correlation is a general term that describes whether two variables are associated [1–3]. The term associated means ‘go together’, that is, knowing the value of one variable, X , enables better prediction of a correlated (associated) variable, Y .

Correlation does *not* imply causality. Height and vocabulary of children are correlated – both increase with age. Clearly, an increase in height does not *cause* an increase in vocabulary, or vice versa. Other examples are less clear. Years of education and income are known to be correlated. Nevertheless, one cannot deduce that more education *causes* higher income. Correlations may be weak or strong, positive or negative, and linear or nonlinear.

Presentation

Correlation can best be seen by a **scatterplot** graph (Figure 1) showing Y as a function of X , together with

a line or curve representing the ‘best fitting’ relation between Y and X . ‘Best fitting’ means that the sum of some measure of deviation of all points from the line or curve is a minimum.

Linear or nonlinear? A linear correlation is one in which the relationship between variables is a straight line. Panels A–C show linear correlations. Panel D shows a nonlinear positive quadratic relation such as the famous Yerkes–Dodson law relating performance to arousal. Panel E shows a nonlinear negative exponential relation, such as the dark adaptation curve relating minimum light energy required to see as a function of time in the dark.

Weak or strong? A strong correlation occurs when knowing one variable strongly predicts the other variable. Panel B shows a weak positive correlation. Panel F shows no correlation for the bulk of points, the horizontal line of best fit; but a substantial correlation, the dashed diagonal line, when the single

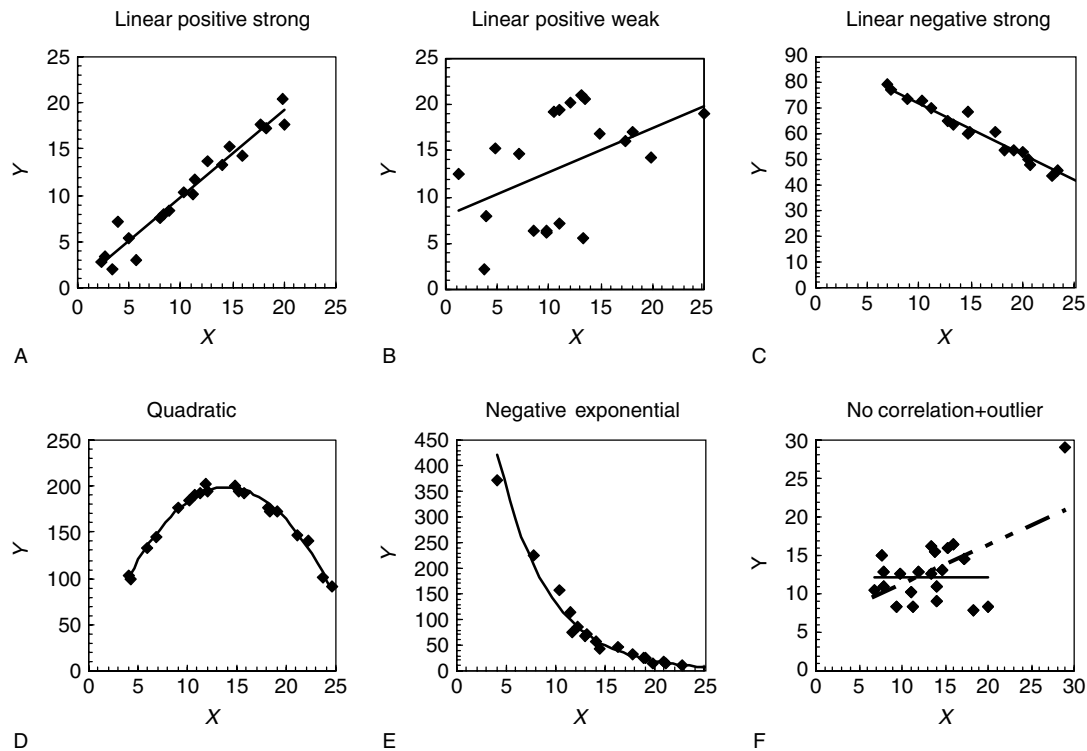


Figure 1 Examples of possible associations between two variables X and Y

2 Correlation

outlier at (29,29) is included. All other panels show some strong relationship.

Positive or negative? Positive correlations arise when both variables increase together, like age and experience, as in Panels A and B. Negative correlations occur when one variable goes down as the other goes up, like age and strength in adults, as in Panel C.

Measures and Uses of Correlation

Measures of correlation range from $+1$, indicating perfect positive agreement to -1 , indicating perfect negative agreement. **Pearson's product moment correlation** r , can be used when both variables are metric (interval or ratio) (see **Scales of Measurement**) and normally distributed. Rank-based measures such as **Spearman's rho** r_s (Pearson's correlation of ranks) or **Kendall's tau** (a measure of how many transpositions are needed to get both variables in the same order) are also useful (sometimes known as nonparametric measures). They are applicable when either variable is ordinal, for example, ranking of candidates; or when metric data has **outliers** or is not bivariate normal. Experts [2] recommend τ over r_s ,

but r_s is widely used and easier to calculate. The **point biserial coefficient** is applicable if one variable is dichotomous and the other metric or ordinal.

Individual correlations are useful in their own right. In addition, correlation matrices, giving all the pairwise correlations of N variables, are useful as input to other procedures such as **Factor Analysis** and **Multidimensional Scaling**.

Acknowledgments

Thanks to Rachel Msetfi and Elena Kulinskaya who made helpful comments on drafts.

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.
- [2] Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th Edition, Edward Arnold, London & Oxford.
- [3] Sheskin, D.J. (2000). *Handbook of Parametric and Non-parametric Statistical Procedures*, 2nd Edition, Chapman & Hall, London.

DIANA KORNROT

Correlation and Covariance Matrices

RANALD R. MACDONALD

Volume 1, pp. 400–402

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Correlation and Covariance Matrices

Covariance and correlation are measures of the degree to which the values on one variable are linearly related to the values on another. In that sense they measure the strength of directional association between two variables. For example, self-esteem is positively correlated with feelings of economic security and negatively correlated with feelings of social isolation [2].

An account of the early history of the term correlation is given in [4]. The term was used during the middle of the nineteenth century, but its statistical use is attributed to **Francis Galton**, who was interested in the relationships between characteristics of related individuals; for example, the correlation between children's and parents' height. The *correlation coefficient* as the central tool for the study of relationships between variables was further developed by F.Y. Edgeworth and **Karl Pearson**, who employed it to look at relationships between physiological and behavioral measurements on the same sets of people.

The population covariance is a measure of the extent to which two variables X and Y are linearly related. It is defined as the expectation (mean)

$$\sigma_{XY} = E[(X - E(X))(Y - E(Y))] \quad (1)$$

and it measures the association between X and Y because whenever a pair of values on X and Y are both above or both below their respective means they add to the covariance whereas if one is above and the other below they reduce it. When Y is a linear function of X (i.e., $Y = a + bX$ where b is a positive constant), then the covariance of X and Y is the product of the standard deviations of X and Y ($\sigma_X \sigma_Y = b \sigma_X^2$) and it is minus this value when $Y = a - bX$. Thus, the covariance ranges from $+\sigma_X \sigma_Y$ to $-\sigma_X \sigma_Y$ as the extent of the linear relationship goes from perfectly positive to perfectly negative.

A population correlation is defined to be the covariance divided by the product of the standard deviations $\sigma_X \sigma_Y$ and it ranges from 1 to -1 . This measure is sometimes called the **Pearson product-moment correlation** and it is useful because it does

not depend on the scale of X and Y assuming they are measured on interval scales. Another interpretation of a correlation is that its square is the proportion of the variance of any one of the variables that is explained in a linear regression (see **Multiple Linear Regression**) equation predicting it from the other variable. It should be remembered that when a correlation or covariance is zero, the variables are not necessarily independent. It only means that there is no linear relationship between the variables. Other nonlinear relationships, for example, a U-shaped relationship, may be present. More information about the interpretation of correlations can be found in [1].

The sample covariance is computed as

$$s_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}), \quad (2)$$

where N is the sample size, $\bar{x}$ is the mean of the sample $x_1, \dots, x_N$, and $\bar{y}$ is the mean of $y_1, \dots, y_N$. As with the variance, the covariance when calculated from a sample is divided by the factor $N-1$ to provide an unbiased estimator of the population covariance. The **Pearson product-moment correlation coefficient** [3] can be used to estimate the population correlation and is calculated from the sample as

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2 \right]^{0.5}} \quad (3)$$

Figure 1 illustrates correlations of a number of different sizes where the only relationship between the variables is linear.

When covariances are used to characterize the relationships between a set of random variables $Z_1, \dots, Z_n$, they are typically presented in the form of a square and symmetric *variance-covariance matrix*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix} \quad (4)$$

where $\sigma_{ij} = \sigma_{ji}$ is the covariance between Z_i and Z_j and the diagonal elements are all equal to the

2 Correlation and Covariance Matrices

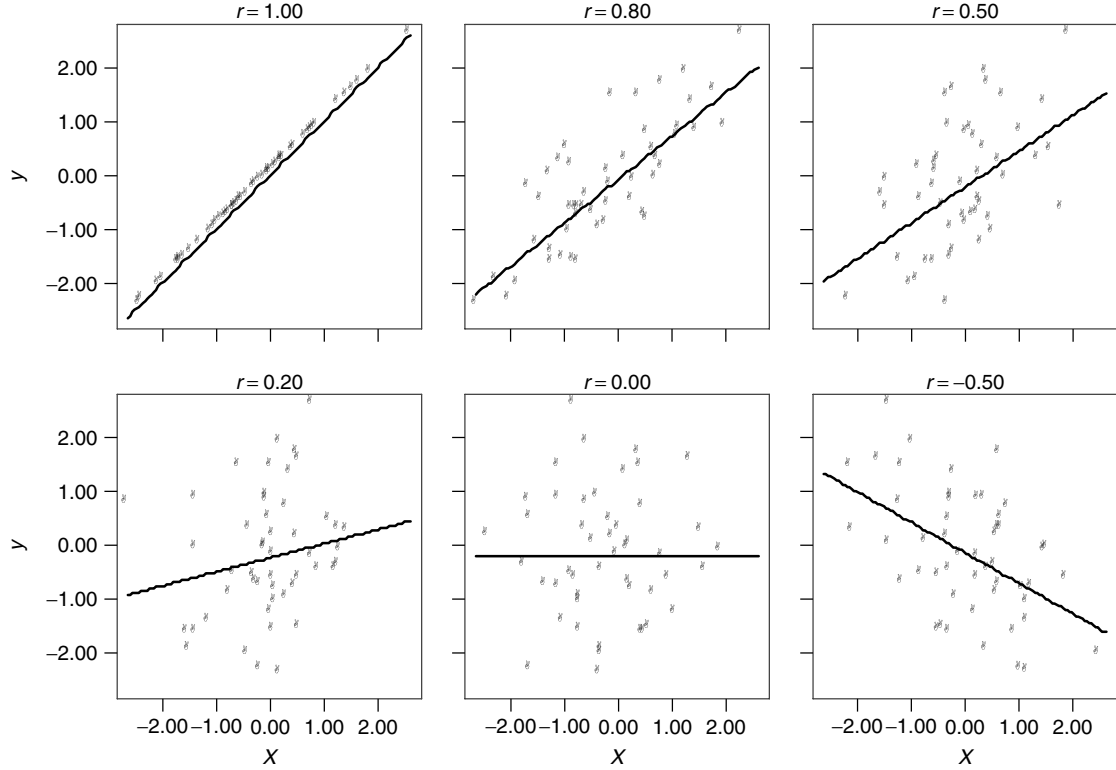


Figure 1 Scatter plots with regression lines illustrating Pearson correlations of 1.0, 0.8, 0.5, 0.2, 0, and -0.5 . X and Y are approximately normally distributed variables with means of 0 and standard deviations of 1

variances because the covariance of a variable with itself is its variance. When correlations are used they are presented in the form of a square correlation matrix

$$\Phi = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{bmatrix} \quad (5)$$

where ρ_{ij} is the correlation between Z_i and Z_j . The diagonal elements of the matrix are unity since by definition a variable can be fully predicted by itself. Also note that in the special case where variables are standardized to unit standard deviation the covariance matrix becomes the correlation matrix.

The variance-covariance matrix is important in multivariate modeling because relationships between variables are often taken to be linear and the multivariate **central limit theorem** suggests that a multivariate normal distribution (see **Catalogue of**

Probability Density Functions), which can be fully characterized by its means and covariance matrix, is a suitable assumption for many multivariate models (see **Multivariate Analysis: Overview**). For example, the commonly employed multivariate techniques **principal component analysis** and **factor analysis** operate on the basis of a covariance or correlation matrix, and the decision whether to start with the covariance or the correlation matrix is essentially one of choosing appropriate scales of measurement.

References

- [1] Howell, D.C. (2002). *Statistical Methods in Psychology*, Duxbury Press, Belmont.
- [2] Owens, T.J. & King, A.B. (2001). Measuring self-esteem: race, ethnicity & gender considered, in *Extending Self-esteem Theory*, T.J. Owens, S. Stryker & N. Goodman, eds, Cambridge University Press, Cambridge.

-
- [3] Pearson, K. (1896). Mathematical contributions to the theory of evolution, III: regression, heredity and panmixia, *Philosophical Transactions of the Society of London, Series A* **187**, 253–318.
- [4] Stigler, S.M. (1986). *The History of Statistics*, Belknap Press, Cambridge.

(See also **Covariance/variance/correlation**; **Kendall's Tau – τ** ; **Partial Least Squares**; **Tetrachoric Correlation**)

RANALD R. MACDONALD

Correlation Issues in Genetics Research

STACEY S. CHERNY

Volume 1, pp. 402–403

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Correlation Issues in Genetics Research

Correlation is of central importance in behavioral genetic research. A basic goal of behavioral genetic research is to partition variance in a trait into variance due to genetic differences among people (commonly referred to as **heritability** (or h^2), variance because of environmental influences shared among family members (shared environmental variance, or c^2), and non-shared environmental variance as a result of unique individual experiences (e^2) (see **ACE Model**). This is accomplished by analyzing correlational structure within different familial relationships. Although in practice we perform **maximum-likelihood** model-fitting analyses on raw data and partition (unstandardized) covariance structure, examining familial correlations directly is the best way of quickly getting an idea of what is going on in the data.

There are two general classes of correlation to estimate the extent of similarity between two (or more) variables. The standard Pearson (interclass) correlation (see **Pearson Product Moment Correlation**) is appropriate when we have two variables measured on a set of individuals. For example, if we have a set of N people on whom we measure height (x) and weight (y), we can compute the interclass correlation between height and weight by first computing the sums of squares of x (SS_x) and y (SS_y) and the sum of cross products of x and y (SS_{xy}):

$$SS_x = \sum_{i=1}^N (x_i - \bar{x})^2 \quad (1)$$

$$SS_y = \sum_{i=1}^N (y_i - \bar{y})^2 \quad (2)$$

$$SS_{xy} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \quad (3)$$

where $\bar{x}$ and $\bar{y}$ are the mean of x and y , respectively. The **intraclass correlation**, r , is then given by:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} \quad (4)$$

In univariate behavioral genetic analyses, we deal with a single variable, for example, the personality trait neuroticism, and the unit of measurement is

not the individual, as above, but rather the family. If, for example, we measure neuroticism on a set of opposite-sex dizygotic twin pairs, the Pearson correlation is the appropriate measure of correlation to estimate twin similarity in that sample. We now have N twin pairs and variable x can be the female member of the pair's neuroticism score and y can be the male member's neuroticism score. The interclass correlation is appropriate whenever we have familial relationships where there are two classes of people, in this case males and females. Similarly, if we wanted to estimate (opposite-sex) spouse similarity, we would also use the Pearson correlation.

However, many relationships do not involve two distinct classes of people and so there is no way to determine who will be put in the x column and who will be in the y column. The most common behavioral genetic design involves the analysis of monozygotic (MZ) and same-sex dizygotic (DZ) twin pairs, and so the appropriate measure of twin similarity is the intraclass correlation. This correlation is computed from quantities obtained from an **analysis of variance**:

$$r = \frac{MS_b - MS_w}{MS_b + (s - 1)MS_w} \quad (5)$$

where MS_b is the mean-square between groups and MS_w is the mean-square within groups, with family (or pair) being the grouping factor. s is the average number of individuals per group, which is simply equal to 2 in the case of a sample of MZ or DZ twin pairs. In practice, use of the Pearson correlation instead of the intraclass correlation will not lead one very much astray, and if one computes a Pearson correlation on double-entered data, where each pair's data is entered twice, once with the first twin in the x column and once with the first twin in the y column, a convenient approximation to the intraclass correlation is obtained.

Both the above forms of correlation deal with continuous data. However, if our data are discrete, such as simple binary data like depressed versus not depressed, or have multiple categories, such as not depressed, moderately depressed, or severely depressed, we would need to employ the polychoric correlation, or the tetrachoric correlation in the special binary case, to obtain measures of familial relationship, the computation of which is beyond the scope of this short article.

STACEY S. CHERNY

Correlation Studies

PATRICIA L. BUSK

Volume 1, pp. 403–404

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Correlation Studies

Correlational research is a form of descriptive research. Because correlational designs are used extensively in educational research, they are considered separately from descriptive research and are used to assess relationships between two or more variables. These methods vary from the simple relationship between two variables to the complex interrelationships among several variables. The advantage of correlational research is that a variety of relationships can be investigated in the same study. Each of these types of studies will be illustrated.

Unlike experimental research, there is no manipulated independent variable, and only the association between variables is studied (*see* **Experimental Design; observational study**). Causal inferences cannot be made on the basis of correlational research [1]. The researcher is seeking to discover the direction and degree of relationship among variables. Test developers use correlational research when assessing the validity of an instrument. Test scores from a new instrument will often be correlated with those from an existing one with established validity evidence. A single sample is obtained and two or more variables measured, and correlational statistics are used to address the research questions. For example, a researcher may be interested in the extent of the relationship between amount of exercise and weight loss. Or a researcher may be interested in the relationship between health beliefs (measured by a questionnaire), social support (measured by a questionnaire), and adherence to medication regime (measured by periodic checkup). For these examples, the variables are continuous.

When data from more than two variables are collected on the same sample, there are two possible ways to handle the information. Correlations could be obtained between all possible pairs of variables, as might be done if the researcher is interested in the intercorrelations between scales on an instrument. A 50-item self-esteem measure was constructed to assess three aspects of self-esteem: global, social, and internal. After obtaining data on a sample of individuals, the researcher would correlate the global scale with the social and internal scales and correlate the social and internal scales. Three correlation coefficients would result from this analysis. If the researcher was not interested in simple correlation

coefficients, then to assess the relationship among the three variables a multiple correlation coefficient (*see* **Multiple Linear Regression**) could be used. This correlation is the relationship between a single dependent variable and the scores derived from a linear combination of independent variables. For example, the adherence to medication regime might be considered the dependent variable and the independent variables would be health-care beliefs and social support.

If a researcher had more than one dependent variable and more than one independent variable when investigating the relationship between these variables, then canonical correlation analysis could be used. Two linear combination or composites are formed – one for the independent variables and one for the dependent variables – and the correlation between the two composites is the canonical correlation [2] (*see* **Canonical Correlation Analysis**).

Correlational research may involve qualitative and quantitative variables. The relationships between variables that are based on the (*see* **Pearson Product Moment Correlation**) would be considered correlational, but those relationships between variables that are not based on the Pearson product-moment correlation are called **measures of association** [3]. The correlational measures involve continuous variables, ranked variables (with few ties), and dichotomous variables. Other forms of categorical variables are involved with the measures of association. Research involving one or two categorical variables (with a small number of categories) is called *correlational-comparative research* or *differential studies*. For the correlational measures, good estimates of the relationships will result when the sample size is adequate (at least 30 individuals) and when the full range of values are obtained for the participants in the research. Otherwise, the researcher is faced with a restriction of range [2]. Although a researcher will have **power** to detect statistical significance with small samples [1], the size of the **confidence interval** would be large, which indicates that the population correlation coefficient is not estimated with a great deal of accuracy. Therefore, sample sizes of at least 30 are recommended. Another consideration with continuous variables is that the range or the variability of the variables is similar. Otherwise, the magnitude of the correlation coefficient will be affected. For example, if the measure of adherence to medication regime is assessed on a scale of 1 to 6 and the other measures

2 Correlation Studies

have a range of 30 or 40 points, then there is unequal variability that will result in a lower value for the correlation coefficient. The variables could be standardized and the correlation coefficient obtained on the standardized variables, which should result in a more accurate estimate of the relationship between the variables. The reliability of the measuring instruments also affects the magnitude of the correlation coefficient. Because instruments are not perfectly reliable, there is attenuation of the correlation coefficient, which can be corrected [4]. If the self-esteem measures are not perfectly reliable, then the correlation coefficient will be smaller than if the measures had perfect reliability. The more reliable the measures, the closer the estimate of the correlation to the corrected one.

References

- [1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Hillsdale.
- [2] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.
- [3] Goodman, L.A. & Kruskal, W.H. (1979). *Measures of Association for Cross Classification*, Springer-Verlag, New York.
- [4] Pedhazur, E.J. & Schmelkin, L.P. (1991). *Measurement, Design, and Analysis: An Integrated Approach*, Lawrence Erlbaum, Mahwah.

PATRICIA L. BUSK

Correspondence Analysis

MICHAEL GREENACRE

Volume 1, pp. 404–415

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Correspondence Analysis

Introduction

Categorical variables are ubiquitous in the behavioral sciences and there are many statistical models to analyze their interrelationships (*see*, for example, **Contingency Tables; Log-linear Models; Exact Methods for Categorical Data**). Correspondence analysis (CA), is a method of **multivariate analysis** applicable to categorical data observed on two or more variables. Its objective is to obtain sets of numerical values for the categories of the variables, where these values reflect the associations between the variables and may be interpreted in a number of ways, especially visually in the form of a spatial map of the categories. Here we shall restrict attention to the correlational and graphical interpretations of CA.

In its simple form, CA analyzes the association between two categorical variables. The method can be extended in different ways to analyze three or more categorical variables, the most common variant being called multiple correspondence analysis (MCA) or *homogeneity analysis*.

Simple Correspondence Analysis: Correlational Definition

To illustrate a simple CA, consider the data on two categorical variables given in Table 1, obtained from the International Social Survey Program (ISSP) on Family and Changing Gender Roles in 1994: the country of origin of the respondent (24 different countries) and the question ‘Do you think it is wrong or not wrong if a man and a woman have sexual relations before marriage?’ with four possible responses: ‘always wrong’ (recorded as 1), ‘almost always wrong’ (2), ‘sometimes wrong’ (3), ‘never wrong’ (4), and a missing category (5), making five categories in total. The data are shown schematically in Table 1 in three different but equivalent forms:

1. The original response pattern matrix, the form typically stored in the database.
2. The individual responses coded as 29 dummy variables with zero-one coding. The resulting *indicator matrix* is subdivided into two submatrices, the first with 24 dummy variables coding

the country variable, the second with 5 dummy variables coding the responses to the substantive question about premarital sex.

3. The 24×5 contingency table which cross-tabulates the responses of all 33 590 cases in the survey.

The chi-square statistic for the contingency table is extremely high ($\chi^2 = 7101$, $df = 92$), due mainly to the large sample sizes in each country, thus indicating significant statistical association. Cramer’s V coefficient, a **measure of association** that has values between 0 (zero association) and 1 (perfect association), is equal to 0.230, indicating a fair association between countries and question responses. But what is the nature of this association? Which countries are similar in terms of their responses to this question, and which are the most different? CA attempts to answer these questions as follows.

Consider two sets of scale values $\mathbf{a} = [a_1, \dots, a_I]^T$ and $\mathbf{b} = [b_1, \dots, b_J]^T$ assigned to the $I = 24$ categories of the first variable and the $J = 5$ categories of the second variable (we write vectors as column vectors, superfix T stands for transpose). If we denote by $\mathbf{Z}_1$ and $\mathbf{Z}_2$ the two submatrices of the indicator matrix in Figure 1, then these scale values imply a pair of scores for each of the respondents in the survey, a ‘country score’ and a ‘question response score’, with all respondent scores in the vectors $\mathbf{Z}_1\mathbf{a}$ and $\mathbf{Z}_2\mathbf{b}$ respectively. Which scale values will maximize the correlation between $\mathbf{Z}_1\mathbf{a}$ and $\mathbf{Z}_2\mathbf{b}$? Since correlations are invariant with respect to linear transformations, we add the identification conditions that $\mathbf{Z}_1\mathbf{a}$ and $\mathbf{Z}_2\mathbf{b}$ are standardized to have mean 0 and variance 1, in which case the correlation to be maximized is equal to $\mathbf{a}^T\mathbf{Z}_1^T\mathbf{Z}_2\mathbf{b}$. Notice that the matrix $\mathbf{Z}_1^T\mathbf{Z}_2$ in this function is exactly the contingency table $\mathbf{N}$ in Table 1(c).

This problem is identical to **canonical correlation analysis** applied to the two sets of dummy variables. The solution is provided by the singular-value decomposition (SVD) (*see Principal Component Analysis*) of the matrix $\mathbf{N}$, where the SVD has been generalized to take into account the identification conditions imposed on the solution. It can be shown that the solution is obtained as follows, where we use the notation n_{i+} , n_{+j} and n for the row sums, column sums, and grand total respectively of the contingency table $\mathbf{N}$.

2 Correspondence Analysis

Table 1 Three forms of the same data (a) responses to the questions, coded according to the response categories; (b) coding of the same data as dummy variables, zero-one data in an indicator matrix $\mathbf{Z}$ with $24 + 5 = 29$ columns; (c) 24×5 contingency table $\mathbf{N}$ cross-tabulating the two variables ($\mathbf{N}$ has a grand total equal to 33 590, the number of cases)

		29 dummy variables		5 categories	
33590 cases	1 2	33590 cases	10000000000000000000000000 01000	24 countries	222 142 315 996 104
	1 4		10000000000000000000000000 00010		97 52 278 1608 289
	1 3		10000000000000000000000000 00100		17 13 95 841 131
	1 5		10000000000000000000000000 00001		107 42 121 623 91
	1 4		10000000000000000000000000 00010		183 48 78 280 58
	.				361 144 228 509 205
	.				36 75 180 636 50
	.				271 147 248 781 53
	2 4		01000000000000000000000000 00010		186 67 153 585 27
	2 2		01000000000000000000000000 01000		298 69 128 357 86
	2 3		01000000000000000000000000 00100		130 51 252 1434 101
	.				139 57 249 1508 134
	.				50 19 59 1086 58
	.				49 62 200 672 41
	24 3		00000000000000000000000001 00100		30 23 111 752 116
	24 2		00000000000000000000000001 01000		256 123 189 856 173
	24 2		00000000000000000000000001 01000		235 91 107 577 116
			00000000000000000000000001 01000		231 227 329 1043 168
					176 53 136 584 98
					152 60 198 893 137
					715 194 151 134 6
					237 88 107 793 62
					234 258 534 175 106
					458 179 207 1445 205

(a) Calculate the matrix $\mathbf{S} = [s_{ij}]$, where

$$s_{ij} = \frac{1}{\sqrt{n}} \left(\frac{n_{ij} - \frac{n_{i+}n_{+j}}{n}}{\sqrt{\frac{n_{i+}n_{+j}}{n}}} \right). \quad (1)$$

(b) Calculate the SVD of $\mathbf{S}$

$$\mathbf{S} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T, \quad (2)$$

where the left and right singular vectors in the columns of $\mathbf{U}$ and $\mathbf{V}$ respectively satisfy $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{\Gamma}$ is the diagonal matrix of positive singular values in descending order down the diagonal: $\gamma_1 \geq \gamma_2 \geq \dots \geq 0$.

(c) Calculate the two sets of optimal scale values from the first singular vectors as follows:

$$\begin{aligned} a_i &= \sqrt{n/n_{i+}}u_{i1}, \quad i = 1, \dots, I \\ b_j &= \sqrt{n/n_{+j}}v_{j1}, \quad j = 1, \dots, J. \end{aligned} \quad (3)$$

The following results can easily be verified and are standard results in CA theory:

1. The maximum correlation achieved by the solution is equal to γ_1 , the largest singular value of $\mathbf{S}$.

2. From the form (1) of the elements of $\mathbf{S}$, the total sum of squares of the matrix $\mathbf{S}$ is equal to χ^2/n , the Pearson **chi-square statistic** χ^2 for the contingency table divided by the sample size n . This quantity, also known as Pearson's mean-square contingency coefficient and denoted by ϕ^2 , is called the *(total) inertia* in CA. Notice that Cramer's $V = \sqrt{(\phi^2/d)}$, where d is equal to the smaller of $I - 1$ or $J - 1$.
3. If we continue our search for scale values, giving scores uncorrelated with the optimal ones found above but again with maximum correlation, the solution is exactly as in (3) but for the second left and right singular vectors of $\mathbf{S}$, with maximum correlation equal to γ_2 , and so on for successive optimal solutions. There are exactly d solutions ($d = 4$ in our example).

Simple Correspondence Analysis: Geometric Definition

A geometric interpretation can be given to the above results: in fact, it is the visualization aspects of CA that have made it popular as a method of

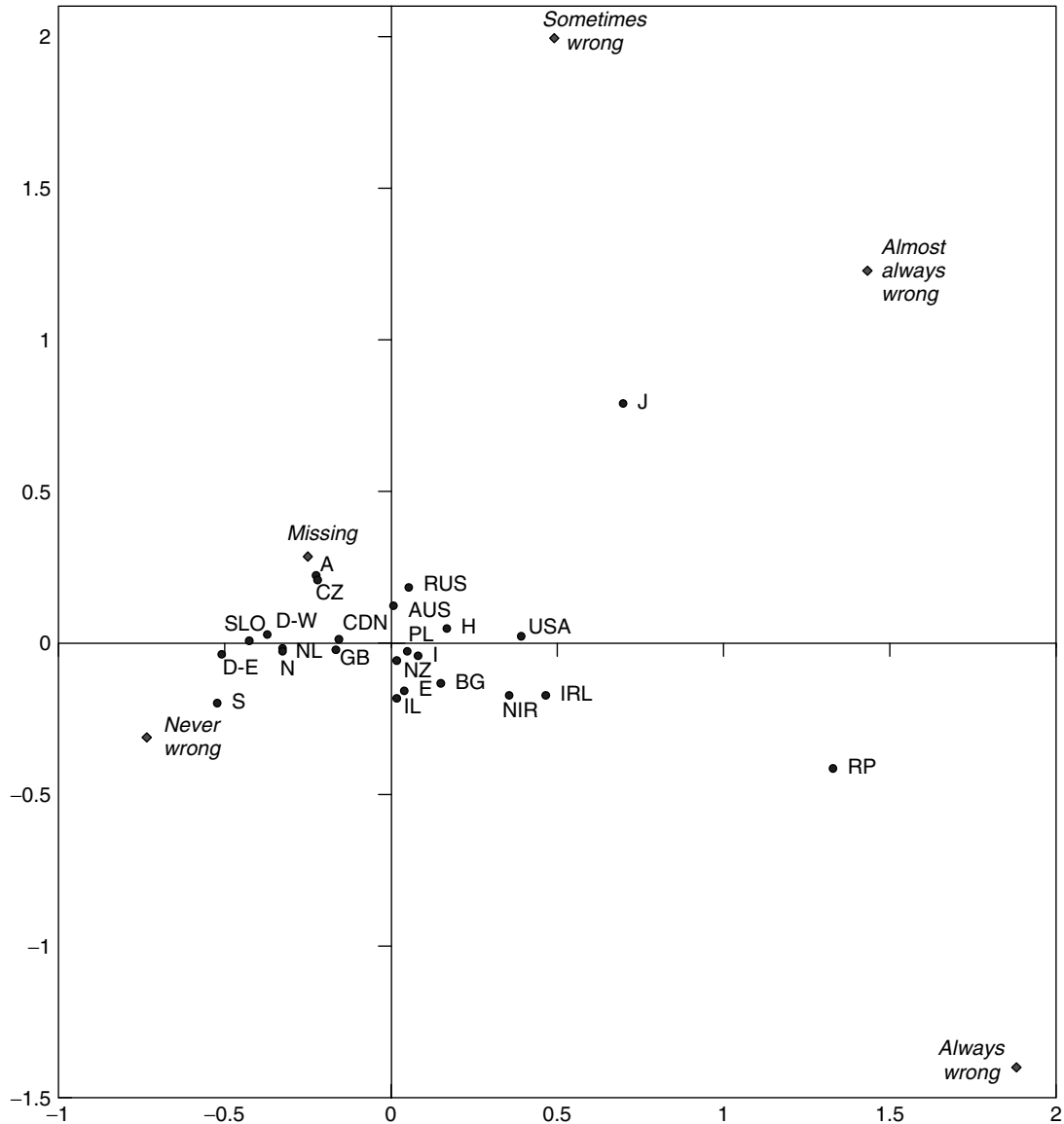


Figure 1 Asymmetric CA map of countries by response categories, showing rows in principal coordinates and columns in standard coordinates. Inertia on first axis: 0.1516 (71.7%), on second axis: 0.0428 (20.2%)

data analysis. Consider the same contingency table $\mathbf{N}$ in Table 1 and calculate, for example, the table of row proportions for each country, across the response categories (Table 2). Each of these five-component vectors, called *profiles*, defines a point in multidimensional space for the corresponding country. In fact, the *dimensionality* d of these row profiles is equal to 4 because each profile's set

of five components sums to a constant 1 (d is the same quantity calculated in previous section). Now assign a weight to each of the row profiles equal to its relative frequency in the survey, called the *mass*, given in the last column of Table 2; thus weighting the profile is proportional to its sample size. Next, define distances between the row profile vectors by the *chi-square distance*:

4 Correspondence Analysis

Table 2 Row (country) profiles based on the contingency table in Table 1(c), showing the row masses and the average row profile used in the chi-square metric

	Country	Abbr.	Always wrong	Almost always wrong	Sometimes wrong	Never wrong	Missing	Mass
1	Australia	AUS	0.125	0.080	0.177	0.560	0.058	0.053
2	Germany (former west)	D-W	0.042	0.022	0.120	0.692	0.124	0.069
3	Germany (former east)	D-E	0.015	0.012	0.087	0.767	0.119	0.033
4	Great Britain	GB	0.109	0.043	0.123	0.633	0.092	0.029
5	Northern Ireland	NIR	0.283	0.074	0.121	0.433	0.090	0.019
6	United States	USA	0.249	0.100	0.158	0.352	0.142	0.043
7	Austria	A	0.037	0.077	0.184	0.651	0.051	0.029
8	Hungary	H	0.181	0.098	0.165	0.521	0.035	0.045
9	Italy	I	0.183	0.066	0.150	0.575	0.027	0.030
10	Ireland	IRL	0.318	0.074	0.136	0.381	0.092	0.028
11	Netherlands	NL	0.066	0.026	0.128	0.729	0.051	0.059
12	Norway	N	0.067	0.027	0.119	0.723	0.064	0.062
13	Sweden	S	0.039	0.015	0.046	0.854	0.046	0.038
14	Czechoslovakia	CZ	0.048	0.061	0.195	0.656	0.040	0.030
15	Slovenia	SLO	0.029	0.022	0.108	0.729	0.112	0.031
16	Poland	PL	0.160	0.077	0.118	0.536	0.108	0.048
17	Bulgaria	BG	0.209	0.081	0.095	0.512	0.103	0.034
18	Russia	RUS	0.116	0.114	0.165	0.522	0.084	0.059
19	New Zealand	NZ	0.168	0.051	0.130	0.558	0.094	0.031
20	Canada	CDN	0.106	0.042	0.138	0.620	0.095	0.043
21	Philippines	RP	0.596	0.162	0.126	0.112	0.005	0.036
22	Israel	IL	0.184	0.068	0.083	0.616	0.048	0.038
23	Japan	J	0.179	0.197	0.409	0.134	0.081	0.039
24	Spain	E	0.184	0.072	0.083	0.579	0.082	0.074
	Weighted average		0.145	0.068	0.139	0.571	0.078	

normalize each column of the profile matrix by dividing it by the square root of the marginal profile element (e.g., divide the first column by the square root of 0.145 and so on), and then use Euclidean distances between the transformed row profiles. Finally, look for a low-dimensional subspace that approximates the row profiles optimally in a weighted least-squares sense; for example, find the two-dimensional plane in the four-dimensional space of the row profiles, which is closest to the profile points, where closeness is measured by (mass-) weighted sum-of-squared (chi-square) distances from the points to the plane. The profile points are projected onto this best-fitting plane in order to interpret the relative positions of the countries (Figure 1).

Up to now, the geometric description of CA is practically identical to classical metric **multidimensional scaling** applied to the chi-square distances between the countries, with the additional feature of weighting each point proportional to its frequency.

But CA also displays the categories of responses (columns) in the map. The simplest way to incorporate the columns is to define fictitious unit row profiles, called *vertices*, as the most extreme rows possible; for example, the vertex profile [1 0 0 0] is totally concentrated into the response ‘always wrong’, as if there were a country that was unanimously against premarital sex. This vertex is projected onto the optimal plane, as well as the vertex points representing the other response categories, including ‘missing’. These vertex points are used as reference points for the interpretation of the country profiles.

It can be shown that this geometric version of the problem has exactly the same mathematical solution as the correlational one described previously. In fact, the following results are standard in the geometric interpretation of simple CA:

1. The column points, that is the projected vertices in Figure 1, have coordinates equal to the scale values obtained in previous section; that is,

we use the b_j 's calculated in (3) for the first (horizontal) dimension, and the similar quantities calculated from the second singular vector for the second (vertical) dimension. The b_j 's are called the *standard coordinates* (of the columns in this case).

2. The positions of the row points in Figure 1 are obtained by multiplying the optimal scale values for the rows by the corresponding correlation; that is, we use the $\gamma_1 a_i$'s for the first dimension, where the a_i 's are calculated as in (3) from the first singular vector and value, and the corresponding values for the second dimension calculated in the same way from the second singular value and vector. These coordinates of the profiles are called *principal coordinates* (of the rows in this case).
3. In the full space of the row profiles (a four-dimensional space in our example) as well as in the (two-dimensional) reduced space of Figure 1, the row profiles lie at weighted averages of the column points. For example, Sweden (S) is on the left of the display because its profile is equal to [0.039 0.015 0.046 0.854 0.046], highly concentrated into the fourth category 'never wrong' (compare this with the average profile at the center, which is [0.145 0.068 0.139 0.571 0.078]). If the response points are assigned weights equal to 0.039, 0.015, and so on, (with high weight of 0.854 on "never wrong") the weighted average position is exactly where the point S is lying. This is called the *barycentric principle* in CA and is an alternative way of thinking about the joint mapping of the points.
4. The joint display can also be interpreted as a **biplot** of the matrix of row profiles; that is, one can imagine oblique axes drawn through the origin of Figure 1 through the category points, and then project the countries onto these axes to obtain approximate profile values. On this biplot axis, the origin will correspond exactly to the average profile element (given in the last row of Table 2), and it is possible to calibrate the axis in profile units, that is, in units of proportions or percentages.
5. The inertia ϕ^2 is a measure of the total variation of the profile points in multidimensional space, and the parts of inertia $\lambda_1 = \gamma_1^2$, $\lambda_2 = \gamma_2^2$, $\dots$, called the *principal inertias*, are those parts of variation displayed with respect to each

dimension, or *principal axis*, usually expressed as percentages. The λ_k 's are the **eigenvalues** of the matrix $\mathbf{S}^T \mathbf{S}$ or $\mathbf{S} \mathbf{S}^T$. This is analogous to the decomposition of total variance in **principal component analysis**.

6. The map in Figure 1 is called the *asymmetric map* of the rows or the 'row-principal' map. An alternative asymmetric map displays column profiles in principal coordinates and row vertices in standard coordinates, called the 'column-principal' map. However, it is common practice in CA to plot the row and column points jointly as profiles, that is, using their principal coordinates in both cases, giving what is called a *symmetric map* (Figure 2). All points in both asymmetric and symmetric maps have the same relative positions along individual principal axes. The symmetric map has the advantage of spreading out the two sets of profile points by the same amounts in the horizontal and vertical directions – the principal inertias, which are the weighted sum-of-squared distances along the principal axes, are identical for the set of row profiles and the set of column profiles.

Notice in Figures 1 and 2 the curve, or *horseshoe*, traced out by the ordinal scale from 'never wrong' on the left, up to 'sometimes wrong' and 'almost always wrong' and then down to 'always wrong' on the right (see **Horseshoe Pattern**). This is a typical result of CA for ordinally scaled variables, showing the ordinal scale on one dimension and the contrast between extreme points and intermediate points on the other. Thus, countries have positions according to two features: first, their overall strength of attitude on the issue, with more liberal countries (with respect to the issue of premarital sex) on the left, and more conservative countries on the right, and second, their polarization of attitude, with countries giving higher than average in-between responses higher up, and countries giving higher than average extreme responses lower down. For example, whereas Spain (S) and Russia (RUS) have the same average attitude on the issue, slightly to the conservative side of average (horizontal dimension), the Russian responses contain relatively more intermediate responses compared to those of the Spanish responses that are more extreme in both directions (see Table 2 to corroborate this finding).

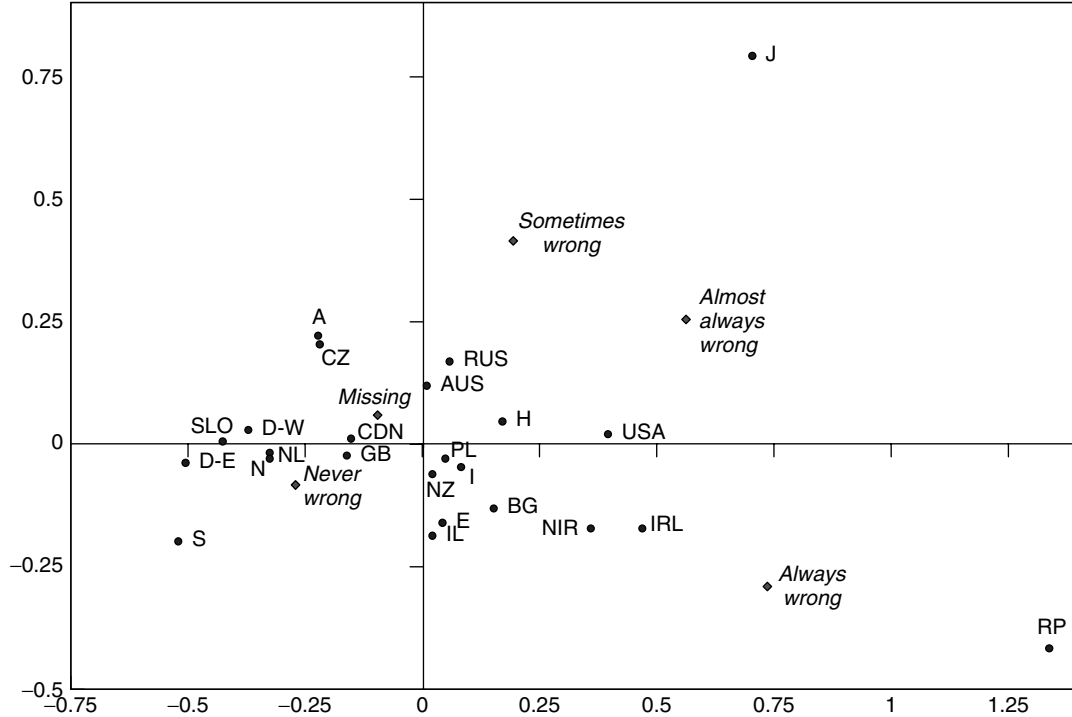


Figure 2 Symmetric CA map of countries by response categories, showing rows and columns in principal coordinates. Inertias and percentages as in Figure 1. Notice how the column points have been pulled in compared to Figure 1

Multiple Correspondence Analysis: Correlational Definition

To illustrate MCA, consider the data from four questions in the same ISSP survey, including the question studied above (A) ‘Do you think it is wrong or not wrong if a man and a woman have sexual relations before marriage?’ (B) ‘Do you think it is wrong or not wrong if a man and a woman in their early teens have sexual relations?’ (C) ‘Do you think it is wrong or not wrong if a man or a woman has sexual relations with someone other than his or her spouse?’ (D) ‘Do you think it is wrong or not wrong if adults of the same sex have sexual relations?’ All the variables have the same categories of response. We have intentionally chosen a set of variables revolving around one issue. The data structures of interest are shown in Table 3:

1. The original response categories for the whole data set of 33 590 cases, with each variable q having four response categories and a missing

category ($J_q = 5, q = 1, \dots, 4$); in general there are Q variables.

2. The $33\,590 \times 20$ indicator matrix $\mathbf{Z} = [\mathbf{Z}_1 \mathbf{Z}_2 \mathbf{Z}_3 \mathbf{Z}_4]$, with four sets of five dummy variables as columns; in general $\mathbf{Z}$ is a $n \times J$ matrix with $J = \sum_q J_q$.
3. The 20×20 block matrix $\mathbf{B}$ of all pairwise contingency tables between the four variables; $\mathbf{B}$ is called the *Burt matrix* and is square ($J \times J$) symmetric, with $\mathbf{B} = \mathbf{Z}^T \mathbf{Z}$. On the block diagonal of $\mathbf{B}$ are the diagonal matrices cross-tabulating each variable with itself, that is, with the marginal frequencies down the diagonal.

Instead of a single association we now have six associations (in general, $\frac{1}{2}Q(Q-1)$ associations) between distinct pairs of variables. To generalize the correlational definition, we use a generalization of canonical correlation analysis to more than two categorical variables. For Q variables, let $\mathbf{a}$ denote the $J \times 1$ vector of unknown scale values, subdivided as $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_Q$ for each variable. The Q scores

Table 3 Three forms of the same multivariate data: (a) responses to the four questions, coded according to the response categories; (b) coding of the same data as dummy variables, zero-one data in an indicator matrix $\mathbf{Z}$ with $4 \times 5 = 20$ columns; (c) 20×20 Burt matrix $\mathbf{B}$ of all 5×5 contingency tables $\mathbf{N}_{qq'}$ cross-tabulating pairs of variables, including each variable with itself on the block diagonal

		20 dummy variables					20 categories									
33590 cases	(a)	2	1	1	1		33590 cases	0	1	0	0	0	0	0	0	0
		4	4	4	3			0	0	0	1	0	0	0	0	0
		3	4	3	2			0	0	1	0	0	0	0	0	0
		5	4	3	3			0	0	0	0	1	0	0	0	0
		4	4	3	2			0	0	1	0	0	0	0	0	0
		...	...	...	...			...	...	...	...	...	...	...	...	...
		...	...	...	...			...	...	...	...	...	...	...	...	...
		3	4	3	1			0	0	1	0	0	0	0	0	0
		4	4	4	4			0	0	0	1	0	0	0	0	0
		5	4	4	3			0	0	0	0	1	0	0	0	0
33590 cases	(b)	0	1	0	0	0	33590 cases	4	8	7	0	0	0	0	4	3
		0	0	0	1	0		0	2	2	8	4	0	0	1	7
		0	0	1	0	0		0	0	4	6	5	3	0	0	0
		0	0	0	0	1		0	0	0	0	1	9	1	6	8
		0	0	1	0	0		0	0	0	0	0	2	6	1	5
		...	...	...	...	...		...	...	...	...	...	...	...	...	...
		...	...	...	...	...		...	...	...	...	...	...	...	...	...
		0	0	1	0	0		4	3	3	1	7	5	4	2	6
		0	0	0	1	0		0	4	0	4	4	2	3	0	0
		0	0	0	0	1		0	0	0	0	1	0	0	0	0
20 categories	(c)	0	0	0	0	0	20 categories	4	3	3	1	7	5	4	2	6
		0	0	0	0	0		0	4	0	4	4	2	3	0	0
		0	0	1	0	0		0	0	4	6	5	3	0	0	0
		0	0	0	1	0		0	0	0	0	1	9	1	6	8
		0	0	1	0	0		0	0	0	0	0	2	6	1	5
		...	...	...	...	...		...	...	...	...	...	...	...	...	...
		...	...	...	...	...		...	...	...	...	...	...	...	...	...
		0	0	1	0	0		4	3	3	1	7	5	4	2	6
		0	0	0	1	0		0	4	0	4	4	2	3	0	0
		0	0	0	0	1		0	0	0	0	1	0	0	0	0

for each case are in $\mathbf{Z}_1\mathbf{a}_1, \mathbf{Z}_2\mathbf{a}_2, \dots, \mathbf{Z}_Q\mathbf{a}_Q$, and the averages of these scores in $(1/Q)\mathbf{Z}\mathbf{a}$. It is equivalent to maximize the sum-of-squared correlations between the Q scores or to maximize the sum-of (or average) -squared correlations between the Q scores and the average:

$$\underset{\mathbf{a}}{\text{maximize}} \frac{1}{Q} \sum_{q=1}^Q \left[\text{cor} \left(\mathbf{Z}_q\mathbf{a}_q, \left(\frac{1}{Q} \right) \mathbf{Z}\mathbf{a} \right) \right]^2. \quad (4)$$

As before, we need an identification condition on $\mathbf{a}$, conventionally the average score $(1/Q)\mathbf{Z}\mathbf{a}$ is standardized to have mean 0 and variance 1. The solution is provided by the same CA algorithm described in formulae (1)–(3), applied either to $\mathbf{Z}$ or to $\mathbf{B}$. The standard coordinates of the columns of $\mathbf{Z}$, corresponding to the maximum singular value γ_1 , provide the optimal solution, and $\lambda_1 = \gamma_1^2$ is the attained maximum of (4). Subsequent solutions are provided by the following singular values and vectors as before. If $\mathbf{B}$ is analyzed rather than $\mathbf{Z}$, the standard coordinates of the columns of $\mathbf{B}$ (or rows, since $\mathbf{B}$ is symmetric) are identical to those of the columns of $\mathbf{Z}$, but because $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$, the singular values and eigenvalues of $\mathbf{B}$ are the squares of those of $\mathbf{Z}$, that is the eigenvalues of $\mathbf{B}$ are $\gamma_1^4, \gamma_2^4, \dots$ and so on.

In homogeneity analysis, an approach equivalent to MCA, the squared correlations $[\text{cor}(\mathbf{Z}_q\mathbf{a}_q, (1/Q)\mathbf{Z}\mathbf{a})]^2$

are called ‘discrimination measures’ and are analogous to squared (and thus unsigned) factor loadings (see **Factor Analysis: Exploratory**). Moreover, in homogeneity analysis the objective of the analysis is defined in a different but equivalent form, namely, as the minimization of the *loss function*:

$$\underset{\mathbf{a}}{\text{minimize}} \frac{1}{nQ} \sum_{q=1}^Q \left\| \mathbf{Z}_q\mathbf{a}_q - \left(\frac{1}{Q} \right) \mathbf{Z}\mathbf{a} \right\|^2, \quad (5)$$

where $\|\dots\|^2$ denotes ‘sum-of-squares’ of the elements of the vector argument. The solution is the same as (4) and the minima are 1 minus the eigenvalues maximized previously. The eigenvalues are interpreted individually and not as parts of variation – if the correlation amongst the variables is high, then the loss is low and there is high homogeneity, or *internal consistency*, amongst the variables; that is, the optimal scale successfully summarizes the association amongst the variables.

Using the ‘optimal scaling’ option in SPSS module Categories, Table 4 gives the eigenvalues and discrimination measures for the four variables in the first five dimensions of the solution (we shall interpret the solutions in the next section) (see **Software for Statistical Analyses**).

8 Correspondence Analysis

Table 4 Eigenvalues and discrimination measures (squared correlations) for first five dimensions in the homogeneity analysis (MCA) of the four questions in Table 3

Dimension	Eigenvalue	Discrimination measures			
		Variable A	Variable B	Variable C	Variable D
1	0.5177	0.530	0.564	0.463	0.514
2	0.4409	0.405	0.486	0.492	0.380
3	0.3535	0.307	0.412	0.351	0.344
4	0.2881	0.166	0.370	0.344	0.256
5	0.2608	0.392	0.201	0.171	0.280

Multiple Correspondence Analysis: Geometric Definition

The geometric paradigm described in section ‘Simple Correspondence Analysis: Geometric Definition’ for simple CA, where rows and columns are projected from the full space to the reduced space, is now applied to the indicator matrix **Z** or the Burt matrix **B**. Some problems occur when attempting to justify the chi-square distances between profiles and the notion of total and explained inertia. There are two geometric interpretations that make more sense: one originates in the so-called Gifi system of homogeneity analysis, the other from a different generalization of CA called *joint correspondence analysis* (JCA).

Geometry of Joint Display of Cases and Categories

As mentioned in property 3 of section ‘Simple Correspondence Analysis: Geometric Definition’, in the asymmetric CA map, each profile (in principal coordinates) is at the weighted average, or centroid, of the set of vertices (in standard coordinates). In the MCA of the indicator matrix with the categories (columns), say, in standard coordinates, and the cases (rows) in principal coordinates, each case lies at the ordinary average of its corresponding category points, since the profile of a case is simply the constant value $1/Q$ for each category of response, and zero otherwise. From (5), the optimal map is the one that minimizes the sum-of-squared distances between the cases and their response categories. It is equivalent to think of the rows in standard coordinates and the columns in principal coordinates, so that each response category is at the average of the case points who have given that response. Again the optimal display is the one that minimizes the sum-of-squared case-to-category

distances. The loss, equal to 1 minus the eigenvalue, is equal to the minimum sum-of-squared distances with respect to individual dimensions in either asymmetric map. The losses can thus be added for the first two dimensions, for example, to give the minimum for the planar display.

It is clearly not useful to think of the joint display of the cases and categories as a biplot, as we are not trying to reconstruct the zeros and ones of the indicator matrix. Nor is it appropriate to think of the CA map of the Burt matrix as a biplot, as explained in the following section.

Geometry of all Bivariate Contingency Tables (Joint CA)

In applying CA to the Burt matrix, the diagonal submatrices on the ‘diagonal’ of the block matrix **B** inflate both the chi-square distances between profiles and the total inertia by artificial amounts. In an attempt to generalize simple CA more naturally to more than two categorical variables, JCA accounts for the variation in the ‘off-diagonal’ tables of **B** only, ignoring the matrices on the block diagonal. Hence, in the two-variable case ($Q = 2$) when there is only one off-diagonal table, JCA is identical to simple CA. The weighted least-squares solution can no longer be obtained by a single application of the SVD and various algorithms have been proposed. Most of the properties of simple CA carry over to JCA, most importantly, the reconstruction of profiles with respect to biplot axes, which is not possible in regular MCA of **B**. The percentages of inertia are now correctly measured, quantifying the success of approximating the off-diagonal matrices relative to the total inertia of these matrices only.

Adjustment of Eigenvalues in MCA

It is possible to remedy partially the percentage of inertia problem in a regular MCA by a compromise between the MCA solution and the JCA objective. The total inertia is measured (as in JCA) by the average inertia of all off-diagonal subtables of **B**, calculated either directly from the tables themselves or by reducing the total inertia of **B** as follows:

$$\begin{aligned} &\text{average off-diagonal inertia} \\ &= \frac{Q}{Q-1} \left(\text{inertia}(\mathbf{B}) - \frac{J-Q}{Q^2} \right). \quad (6) \end{aligned}$$

Parts of inertia are then calculated from the eigenvalues λ_s^2 of **B** (or λ_s of **Z**) as follows: for each

$\lambda_s \geq 1/Q$ calculate the adjusted inertias

$$\left(\frac{Q}{Q-1} \right)^2 \left(\lambda_s - \frac{1}{Q} \right)^2$$

and express these as percentages of (6). Although these percentages underestimate those of JCA, they

dramatically improve the results of MCA and are recommended in all applications of MCA.

In our example, the total inertia of **B** is equal to 1.1659 and the first five principal inertias are such that $\lambda_s \geq 1/Q$, that is, $\lambda_s^2 \geq 1/Q^2 = 1/16$. The different possibilities for inertias and percentages of inertia are given in Table 5. Thus, what appears to be a percentage explained in two dimensions of 29.9% (= 12.9 + 11.0) in the analysis of the indicator matrix **Z**, or 39.7% (= 23.0 + 16.7) in the analysis of the Burt matrix **B**, is shown to be at least 86.9% (= 57.6 + 29.3) when the solution is rescaled. The JCA solution gives only a slight extra benefit in this case, with an optimal percentage explained of 87.4%, so that the adjusted solution is practically optimal.

Figure 3 shows the adjusted MCA solution, that is, in adjusted principal coordinates. The first dimension lines up the four ordinal categories of each question in their expected order (accounting for 57.6% of the inertia), whereas the second dimension opposes all the missing categories against the rest (29.3% of inertia). The positions of the missing values on the

Table 5 Eigenvalues (principal inertias) of the indicator matrix and the Burt matrix, their percentages of inertia, the adjusted inertias, and their lower bound estimates of the percentages of inertia of the off-diagonal tables of the Burt matrix. The average off-diagonal inertia on which these last percentages are based is equal to $4/3(1.1659 - 16/16) = 0.2212$

Dimension	Eigenvalue of Z	Percentage explained	Eigenvalue of B	Percentage explained	Adjusted eigenvalue	Percentage explained
1	0.5177	12.9	0.2680	23.0	0.1274	57.6
2	0.4409	11.0	0.1944	16.7	0.0648	29.3
3	0.3535	8.8	0.1249	10.7	0.0190	8.6
4	0.2881	7.2	0.0830	7.1	0.0026	1.2
5	0.2608	6.5	0.0681	5.8	0.0002	0.0

Table 6 Category scale values (standard coordinates) on first dimension (see Figure 3), and their linearly transformed values to have 'not wrong at all' categories equal to 0 and highest possible score (sum of underlined scale values) equal to 100

	Always wrong	Almost always wrong	Only sometimes wrong	Not wrong at all	Missing
(a) Sex before marriage	<u>2.032</u> 27.9	1.167 19.2	0.194 9.5	-0.748 0.0	0.333 10.9
(b) Sex teens under 16	<u>0.976</u> 25.4	-0.813 7.4	-1.463 0.9	-1.554 0.0	-0.529 10.3
(c) Sex other than spouse	<u>0.747</u> 23.5	-1.060 5.3	-1.414 1.8	-1.591 0.0	-0.895 7.0
(d) Sex two people of same sex	<u>1.001</u> 23.2	-0.649 6.6	-0.993 3.2	-1.310 0.0	-0.471 8.4

10 Correspondence Analysis

first axis will thus provide estimated scale values for missing data that can be used in establishing a general attitude scale of attitudes to sex. The scale values on the first dimension can be transformed to have

more interpretable values, for example, each category 'always wrong' can be set to 0 and the upper limit of the case score equalized to 100 (Table 6). The process of redefining the scale in this way is invariant

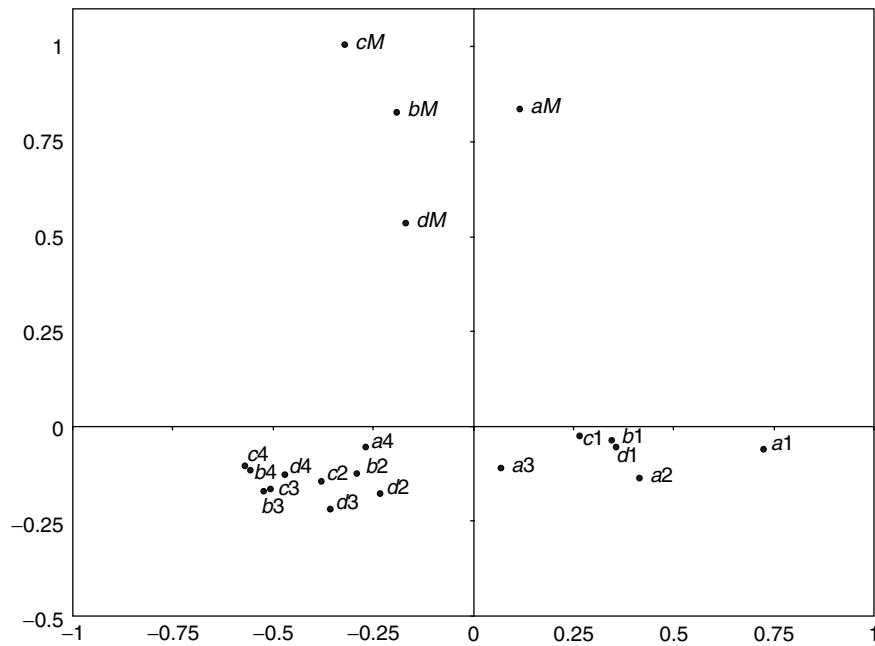


Figure 3 MCA map of response categories (analysis of Burt matrix): first (horizontal) and second (vertical) dimensions, using adjusted principal inertias. Inertia on first axis: 0.1274 (57.6%), on second axis: 0.0648 (29.3%). Labels refer to variable A to D, with response categories 1 to 4 and missing (M)

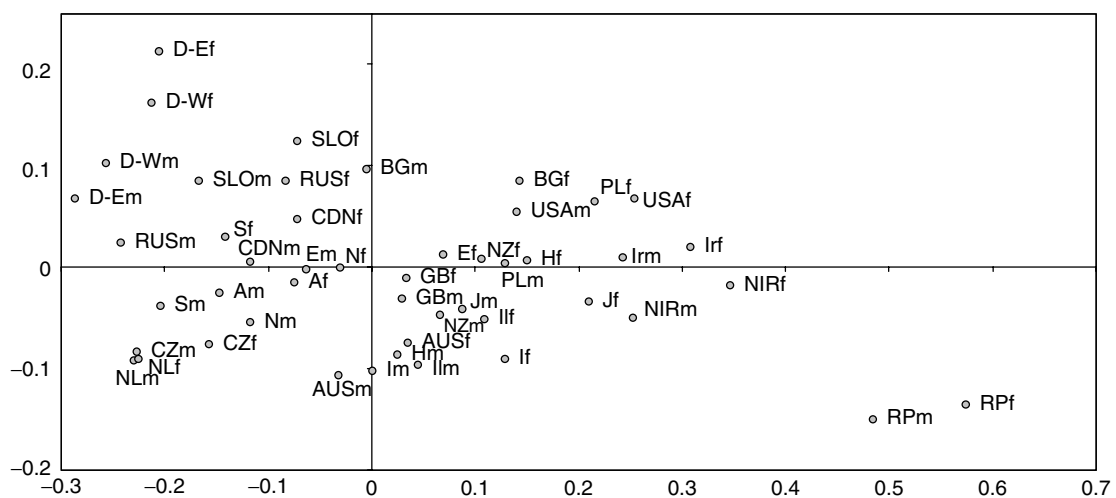


Figure 4 Positions of supplementary country-gender points in the map of Figure 3. Country abbreviations are followed by m for male or f for female

with respect to the particular scaling used: standard, principal, or adjusted principal.

Having identified the optimal scales, subgroups of points may be plotted in the two-dimensional map; for example, a point for each country or a point for a subgroup within a country such as Italian females or Canadian males. This is achieved using the barycentric principle, namely, that the profile position is at the weighted average of the vertex points, using the profile elements as weights. This is identical to declaring this subgroup a supplementary point, that is, a profile that is projected onto the solution subspace. In Figure 4 the positions of males and females in each country are shown. Apart from the general spread of the countries with conservative countries (e.g., Philippines) more to the right and liberal countries more to the left (e.g., Germany), it can be seen that the female groups are consistently to the conservative side of their male counterparts and also almost always higher up on the map, that is, there are also more nonresponses amongst the females.

Further Reading

Benzécri [1] – in French – represents the original material on the subject. Greenacre [4] and [7] both give introductory and advanced treatments of the French approach, the former being a more encyclopedic treatment of CA and the latter including other methods of analyzing large sets of categorical data. Gifi [3] includes CA and MCA in

the framework of nonlinear multivariate analysis, a predominantly Dutch approach to data analysis. Greenacre [5] is a practical user-oriented introduction. Nishisato [8] provides another view of the same methodology from the viewpoint of quantification of categorical data. The edited volumes by Greenacre and Blasius [6] and Blasius and Greenacre [2] are self-contained state-of-the-art books on the subject, the latter volume including related methods for visualizing categorical data.

References

- [1] Benzécri, J.-P. (1973). *Analyse des Données. Tôme 2: Analyse des Correspondances*, Dunod, Paris.
- [2] Blasius, J. & Greenacre, M.J., (eds) (1998). *Visualization of Categorical Data*, Academic Press, San Diego.
- [3] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester.
- [4] Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.
- [5] Greenacre, M.J. (1993). *Correspondence Analysis in Practice*, Academic Press, London.
- [6] Greenacre, M.J. & Blasius, J., (eds) (1994). *Correspondence Analysis in the Social Sciences*, Academic Press, London.
- [7] Lebart, L., Morineau, A. & Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis*, Wiley, Chichester.
- [8] Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*, Lawrence Erlbaum, Hillsdale.

MICHAEL GREENACRE

Co-twin Control Methods

JACK GOLDBERG AND MARY FISCHER

Volume 1, pp. 415–418

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Co-twin Control Methods

Overview

The co-twin control method is one of the most simple and elegant research designs available in the behavioral sciences. The method uses differences within twin pairs to examine the association between a putative environmental risk factor and an outcome variable. The unique value of the design is that numerous potential confounding factors, such as age and parental socioeconomic status, are matched and cannot confound the risk factor–outcome association. This design can be extremely efficient for examining risk factor–outcome associations compared to unmatched designs [12]. When the within-pair comparisons are restricted to monozygotic (MZ) twins, who share 100% of their genetic material, the risk factor–outcome association is completely controlled for confounding due to genetic factors.

The central feature of the design is a focus on within-pair differences among either risk factor or outcome discordant pairs. In most co-twin control studies, within-pair discordance is defined as a dichotomous variable where one twin has the risk factor (i.e., smokes cigarettes) while the other twin does not (i.e., nonsmoker). Co-twin control studies that use an experimental approach create the discordant pairs by randomly assigning one member of a pair to the active treatment group and the other member of the pair to the control group. Experimental co-twin control studies then compare within-pair differences in an outcome variable among the treatment and control groups. Feeding studies have used this design extensively to examine the effects of altering diet on weight and body fat distribution [2].

Observational co-twin control studies can be formulated in two different ways, depending on how the twin sample is assembled. A cohort formulation begins by identifying pairs who are discordant for an environmental risk factor and then compares within-pair differences in the outcome variable. An example of the cohort co-twin control approach is the Finnish study of lumbar disc degeneration in 45 twin pairs discordant for lifetime driving exposure [1]. The so-called case-control co-twin formulation starts by defining pairs discordant for an outcome variable and then examines within-pair differences

for an environmental risk factor. The case-control co-twin design is especially valuable when the outcome being investigated is rare and the twins are derived from treatment seeking samples or volunteers. The case-control co-twin design was used in a recent study of regional cerebral blood flow in 21 MZ twin pairs discordant for chronic fatigue syndrome [10].

Statistical Methods

The statistical analysis of co-twin control studies uses well-established methods for matched data to account for the lack of independence between members of a twin pair. We present the basic analytic methods that can be applied to co-twin control studies with a discordant environmental risk factor and a continuous or dichotomous outcome variable.

Dichotomous Environmental Risk Factor and a Continuous Outcome Variable

The analysis begins by calculating the mean value of the outcome variable among all discordant twin pairs, separately in those exposed and not exposed to the risk factor. These means are useful descriptive statistics in helping to understand the pattern of association between the risk factor and the outcome. Formal statistical testing is conducted on the basis of the within-pair mean difference in those exposed and not exposed. The standard error for the within-pair mean difference is used to calculate the matched pair *t* Test and 95% **confidence intervals**. When information on twin zygosity is available, it is informative to repeat these analyses separately in MZ and dizygotic (DZ) pairs. Zygosity-specific patterns of within-pair mean differences can be helpful in understanding the risk factor–outcome association. If the within-pair mean differences are not equal to zero and similar in MZ and DZ pairs, this would suggest that the risk factor–outcome association is not confounded by familial or genetic influences. However, if after zygosity stratification a sizable within-pair difference is present in DZ pairs but completely absent in MZ pairs, this would strongly suggest genetic confounding [4]. Intermediate patterns in which the size of the MZ within-pair difference is reduced, but is still present, imply a modest degree of genetic confounding. Differences in the MZ and DZ within-pair mean

difference can be formally tested for heterogeneity using a two-sample t Test.

Dichotomous Environmental Risk Factor and a Dichotomous Outcome Variable

The analysis of dichotomous outcomes in the co-twin control study starts by estimating the prevalence of the outcome in twins exposed and unexposed to the risk factor. This descriptive information is valuable for interpreting study findings since it is often difficult to present the details of a within-pair analysis of dichotomous outcomes in a simple table. Estimation of the strength of the association between the risk factor and outcome is based on the matched pair **odds ratio (OR)**. The matched pair odds ratio uses the count of the number of discordant twin pairs according to the pair configuration on the risk factor and outcome. With a dichotomous risk factor and outcome, there are four possible pair configurations to count: n_{11} = the number of risk factor discordant pairs in which both members have the outcome; n_{10} = the number of risk factor discordant pairs in which the twin exposed to the risk factor has the outcome and the twin not exposed does not have the outcome; n_{01} = the number of risk factor discordant pairs in which the twin exposed to the risk factor does not have the outcome and the twin not exposed has the outcome; n_{00} = the number of risk factor discordant pairs in which neither twin has the outcome. The odds ratio is simply n_{10}/n_{01} . McNemar's X^2 test for matched pairs (*see Paired Observations, Distribution Free Methods*) can be used to test the significance of the odds ratio, and the standard error for the matched pair odds ratio can be used to derive 95% confidence intervals.

If zygosity information is available, the matched pair statistical analysis can be repeated after stratifying by zygosity. This analysis obtains separate odds ratio estimates of the risk factor–outcome association in MZ and DZ pairs. Interpretation of these zygosity-specific matched pair odds ratios is directly analogous to that used when examining within-pair mean differences [9]. When both the MZ and DZ risk factor–outcome odds ratios are greater than 1 and of similar magnitude, this suggests the absence of genetic confounding. Conversely, when the DZ odds ratio is elevated while the MZ odds ratio approaches 1, this strongly suggests that the risk factor–outcome association is due to the confounding influence of

genetic factors. Testing the difference between the stratum-specific matched pair odds ratios involves using the risk factor–outcome discordant pairs in a Pearson's X^2 test with 1 degree of freedom (*see Contingency Tables*).

An Illustration: A Co-twin Control Analysis of Military Service in Vietnam and Posttraumatic Stress Disorder (PTSD)

The co-twin control analysis is illustrated using data derived from the Vietnam Era Twin Registry [6]. The Registry was established in the 1980s and contains male–male veteran twin pairs born between 1939 and 1957 who served on active duty military service during the Vietnam era (1965–1975). In 1987, a mail/telephone survey collected data on a number of demographic, behavioral, and health characteristics including zygosity, service in Vietnam, and PTSD symptoms.

Zygosity was determined using a questionnaire similarity algorithm supplemented with limited blood group typing from the military records [5]. Service in Vietnam was based on a single yes/no question item. Fifteen question items inquired about the frequency of PTSD symptomology in the past 6 months; twins could rank frequency according to a five-level ordinal response ranging from never = 1 to very often = 5. The items were summed to create a symptom scale with higher values representing a greater likelihood of PTSD symptoms. The statistical analysis first examines the relationship of Vietnam service to the PTSD symptom scale where the scale is considered a continuous outcome variable; the analysis is then repeated after making the PTSD scale a dichotomy, with the upper quartile defined as those more likely to have PTSD.

Table 1 presents the within-pair mean differences for the PTSD symptom scale according to service in Vietnam. There were a total of 1679 twin pairs in the Registry, where one member of the pair served in Vietnam but their twin sibling did not. In all twins, the mean PTSD symptom scale was higher in those who served in Vietnam compared to those who served elsewhere; the within-pair mean difference was 5 (95% CI 4.4,5.6) with a matched pair t Test (*see Catalogue of Parametric Tests*) of 16.6 ($p < .001$). Analysis stratified by zygosity found similar within-pair mean differences in MZ

Table 1 Co-twin control analysis of Vietnam service discordant pairs and mean levels of PTSD symptoms

Sample	Number of Vietnam discordant pairs	Mean level of PTSD symptoms for Vietnam service		Within-pair mean difference in PTSD symptoms	95% CI	Matched pair <i>t</i> Test		Heterogeneity	
		Vietnam service	No service in Vietnam			<i>t</i>	<i>P</i> value	<i>t</i>	<i>P</i> value
All pairs	1679	29.6	24.6	5.0	4.4,5.6	16.6	<0.001		
MZ	846	29.7	24.5	5.2	4.4,6.0	13.0	<0.001	0.67	0.50
DZ	833	29.5	24.7	4.8	3.9,5.7	10.6	<0.001		

Table 2 Co-twin control analysis of Vietnam service discordant pairs and the presence of high levels of PTSD symptoms

Sample	Number of Vietnam discordant pairs	PTSD prevalence ^a		Pair configuration ^b						McNemar's test		Heterogeneity	
		Vietnam service (%)	No service in Vietnam (%)	<i>n</i> ₁₁	<i>n</i> ₁₀	<i>n</i> ₀₁	<i>n</i> ₀₀	OR	95% CI	<i>X</i> ²	<i>P</i> value	<i>X</i> ²	<i>P</i> value
All pairs	1679	34.8	15.5	148	437	113	981	3.9	3.1,4.8	191.0	<0.001	0.37	0.54
MZ	846	34.0	14.4	69	219	53	505	4.1	3.1,5.6	101.4	<0.001		
DZ	833	35.7	16.7	79	218	60	476	3.6	2.7,4.8	89.9	<0.001		

^aThe prevalence of PTSD was defined as a PTSD score greater than 32, which represents the upper quartile of the PTSD scale distribution.

^b*n*₁₁ = number of Vietnam discordant pairs in which both twins have PTSD; *n*₁₀ = number of Vietnam discordant pairs in which the twin with service in Vietnam has PTSD and the twin without Vietnam service does not have PTSD; *n*₀₁ = number of Vietnam discordant pairs in which the twin with service in Vietnam does not have PTSD and the twin without Vietnam service has PTSD; *n*₀₀ = number of Vietnam discordant pairs in which neither twin has PTSD.

and DZ pairs. The comparison of heterogeneity of the MZ and DZ within-pair mean differences was not significant ($p = .503$), providing little evidence of genetic confounding.

Table 2 presents the matched pairs analysis of service in Vietnam and the dichotomous indicator of PTSD symptoms. Overall, 35% of those who served in Vietnam were in the upper quartile of the PTSD symptom scale, while only 16% of those who did not serve in Vietnam had similarly high levels of PTSD symptoms. The matched pair odds ratio indicates that twins who served in Vietnam were nearly 4 times more likely to have high levels of PTSD symptoms compared to their twin who did not serve. The within-pair difference in PTSD was highly significant based on the McNemar test ($X^2 = 191.00$; $p < .001$). Stratification by zygosity demonstrated that the effects were similar in both MZ and DZ pairs; the difference in the matched pair odds ratios

in MZ and DZ pairs was not statistically significant ($X^2 = 0.37$, $p = .543$).

More Advanced Methods

For co-twin control designs using the experimental or cohort approach with risk factor discordant twin pairs, there is now a wide range of methods for clustered data that can be used to analyze twins [7, 11]. Duffy [3] has described how co-twin control studies can be analyzed using **structural equation models**. Methods such as random effects regression (see **Linear Multilevel Models; Generalized Linear Mixed Models**) models and **generalized estimating equations** are readily adapted to the analysis of twins [14]. These methods are extremely flexible and allow additional covariates to be included in the regression model as both main-effects and interaction terms.

Options are available to examine outcome variables that can take on virtually any structure, including continuous, dichotomous, ordinal, and censored. Further extensions allow the simultaneous analysis of multiple outcomes as well as longitudinal analysis of repeated measures over time (*see Longitudinal Data Analysis; Repeated Measures Analysis of Variance*). These more complex applications can also go beyond the discordant co-twin control method to incorporate all types of exposure patterns within twin pairs [8]. However, application of these more sophisticated analyses should be done with great care to make sure that the appropriate statistical model is used [13].

References

- [1] Battie, M.C., Videman, T., Gibbons, L.E., Manninen, H., Gill, K., Pope, M. & Kaprio, J. (2002). Occupational driving and lumbar disc degeneration: a case-control study, *Lancet* **360**, 1369–1374.
- [2] Bouchard, C. & Tremblay, A. (1997). Genetic influences on the response of body fat and fat distribution to positive and negative energy balances in human identical twins, *The Journal of Nutrition* **127**, (Suppl. 5), 943S–947S.
- [3] Duffy, D.L. (1994). Biometrical genetic analysis of the cotwin control design, *Behavior Genetics* **24**, 341–344.
- [4] Duffy, D.L., Mitchell, C.A. & Martin, N.G. (1998). Genetic and environmental risk factors for asthma: a cotwin-control study, *American Journal of Respiratory and Critical Care Medicine* **157**, 840–845.
- [5] Eisen, S.A., Neuman, R., Goldberg, J., Rice, J. & True, W.R. (1989). Determining zygosity in the Vietnam era twin (VET) registry: an approach using questionnaires, *Clinical Genetics* **35**, 423–432.
- [6] Goldberg, J., Curran, B., Vitek, M.E., Henderson, W.G. & Boyko, E.J. (2002). The Vietnam Era Twin (VET) registry, *Twin Research* **5**, 476–481.
- [7] Goldstein, H. (2003). *Multilevel Statistical Models*, Hodder Arnold, London.
- [8] Hu, F.B., Goldberg, J., Hedeker, D. & Henderson, W.G. (1999). Modeling ordinal responses from co-twin control studies, *Statistics in Medicine* **17**, 957–970.
- [9] Kendler, K.S., Neale, M.C., MacLean, C.J., Heath, A.C., Eaves, L.J. & Kessler, R.C. (1993). Smoking and major depression: a causal analysis, *Archives of General Psychiatry* **50**, 36–43.
- [10] Lewis, D.H., Mayberg, H.S., Fischer, M.E., Goldberg, J., Ashton, S., Graham, M.M. & Buchwald, D. (2001). Monozygotic twins discordant for chronic fatigue syndrome: regional cerebral blood flow SPECT, *Radiology* **219**, 766–773.
- [11] Liang, K.Y. & Zeger, S.L. (1993). Regression analysis for correlated data, *Annual Review of Public Health* **14**, 43–68.
- [12] MacGregor, A.J., Snieder, H., Schork, N.J. & Spector, T.D. (2000). Twins: novel uses to study complex traits and genetic diseases, *Trends in Genetics* **16**, 131–134.
- [13] Neuhaus, J.M. & Kalbfleisch, J.D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data, *Biometrics* **54**, 638–645.
- [14] Quirk, J.T., Berg, S., Chinchilli, V.M., Johansson, B., McClearn, G.E. & Vogler, G.P. (2001). Modelling blood pressure as a continuous outcome variable in a co-twin control study, *Journal of Epidemiology and Community Health* **55**, 746–747.

JACK GOLDBERG AND MARY FISCHER

Counterbalancing

VENITA DEPUY AND VANCE W. BERGER

Volume 1, pp. 418–420

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Counterbalancing

Counterbalancing, in its simplest form, is the procedure of varying the order in which subjects receive treatments in order to minimize the bias potentially associated with the treatment order. When each participant receives more than one treatment (as in a **crossover design** [6]), counterbalancing is necessary to ensure that treatment effects are not confounded with **carryover** effects. One well-known example is the Pepsi–Coke taste test, common in the 1980s. Half the participants were given Pepsi first, and half the participants were given Coke first. Participants were also offered a cracker before trying the second cola. The random ordering neutralized the order effect, while the cracker minimized the carryover effect.

When and Why Counterbalancing is Needed

In some experiments, it may be suspected that the outcome of a given treatment may be affected by the number of treatments preceding it, and/or the treatment immediately preceding it [2]. Counterbalancing is used to control these effects. If a subject is given three treatments, then the results may be affected by the order of the treatments, carryover from the prior treatments, and in some cases where a task is repeated, subjects may learn to complete the task better over time. This learning effect, known more formally as asymmetric skill transfer, is discussed further in [7].

In the example of the cola taste test mentioned above, thirsty participants may be more likely to choose the first drink given. Similarly, if the first drink satisfies the thirst, then the lack of thirst may affect how favorably the participant views the second drink. A devoted cola fan, after taking the taste test multiple times, might even learn to distinguish between the two colas and pick the one identified as his or her favorite each time, even if that cola would not have been the one selected in the absence of prior knowledge.

While this is a simplistic example, these effects can drastically alter the outcome of experiments if ignored. There are, however, some limitations to when counterbalancing can be used. Manipulations

must be reversible. For instance, suppose that the two available treatments for a certain medical condition are surgery and no surgery. Patients receiving surgery first would likely experience enormous carryover effect, while those receiving no surgery, then surgery, would be expected to have little, if any, carryover. Counterbalancing would not balance out these effects. Instead, a between-subjects design should be considered.

Complete Counterbalancing

Complete counterbalancing requires that every set of combinations occurs equally often (in the same number of subjects). This is often considered the optimal method. With only two treatments, A and B, half the participants would receive treatments in the order AB, and the other half would receive BA. With three treatments, there are six orders, specifically ABC, ACB, BAC, BCA, CAB, and CBA. With complete counterbalancing, each of these six orders would be represented in a given number of subjects, thereby requiring equal numbers of subjects in all six groups. With more treatments, the number of permutations grows dramatically. In fact, for k treatments, there are $k!$ different combinations. This potentially large number of permutations can be one of the primary drawbacks to complete counterbalancing.

Incomplete Counterbalancing

Incomplete, or partial, counterbalancing is often used when complete counterbalancing is not feasible. The most prevalent type is a **Latin square**. Further details of the Latin Square construction are given in [2]. It should be noted that Latin squares of order n are extremely numerous for $n > 3$. Indeed, the first row can be any one of $n!$ permutations. After the first row is chosen, there are approximately $n!/e$ choices for the second row (e is the base for the natural logarithm). A simple Latin square provides a set of treatment orders in which every treatment is given first once, second once, and so on. In a simple example with four treatments and four subjects, treatments A, B, C, and D would be given in order as follows:

Subject 1:	A	B	C	D
Subject 2:	B	C	D	A
Subject 3:	C	D	A	B
Subject 4:	D	A	B	C

2 Counterbalancing

While every treatment appears in every time point exactly once, it should be noted that treatment order does not vary in a simple Latin square. Every treatment C is followed by treatment D, and so on. While this method can account for both the order effect and the learning effect, it does not counteract the carryover effect. A better method is a balanced Latin square. In this type of Latin square, each treatment immediately follows and immediately precedes each other treatment exactly once, as shown here:

Subject 1:	A	B	D	C
Subject 2:	B	C	A	D
Subject 3:	C	D	B	A
Subject 4:	D	A	C	B

It should be noted that balanced squares can be constructed only when the number of treatments is even. For odd numbers of treatments, a mirror image of the square must be constructed. Further details are given in [3]. Latin squares are used when the number of subjects, or blocks of subjects, equals the number of treatments to be administered to each. When the number of sampling units exceeds the number of treatments, multiple squares or rectangular arrays should be considered, as discussed in [8]. In the event that participants receive each treatment more than once, reverse counterbalancing may be used. In this method, treatments are given in a certain order and then in the reverse order. For example, a subject would receive treatments ABCCBA or CBAABC. A variation on this theme, in which subjects were randomized to be alternated between treatments A and B in the order ABABAB... or BABABA..., was recently used to evaluate itraconazole to prevent

fungal infections [5] and critiqued [1]. It is also possible to randomize treatments to measurement times, for a single individual or for a group [4].

References

- [1] Berger, V.W. (2003). Preventing fungal infections in chronic granulomatous disease, *New England Journal of Medicine* **349**(12), 1190.
- [2] Bradley, J.V. (1958). Complete counterbalancing of immediate sequential effects in a Latin square design (Corr: V53 p1030-31), *Journal of the American Statistical Association* **53**, 525–528.
- [3] Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*, 2nd Edition, Wiley, New York. (First corrected printing, 1968).
- [4] Ferron, J. & Onghena, P. (1996). The power of randomization tests for single-case phase designs, *Journal of Experimental Education* **64**(3), 231–239.
- [5] Gallin, J.I., Alling, D.W., Malech, H.L., wesley, R., Koziol, D., Marciano, B., Eisenstein, E.M., Turner, M.L., DeCarlo, E.S., Starling, J.M. & Holland, S.M. (2003). Itraconazole to prevent fungal infections in chronic granulomatous disease, *New England Journal of Medicine* **348**, 2416–2422.
- [6] Maxwell, S.E. & Delaney, H.D. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Brooks/Cole, Pacific Grove.
- [7] Poulton, E.C. & Freeman, P.R. (1966). Unwanted asymmetrical transfer effects with balanced experimental designs, *Psychological Bulletin* **66**, 1–8.
- [8] Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*, 2nd Edition, McGraw-Hill, Boston.

VENITA DEPUY AND VANCE W. BERGER

Counterfactual Reasoning

PAUL W. HOLLAND

Volume 1, pp. 420–422

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Counterfactual Reasoning

The term ‘counterfactual conditional’ is used in logical analysis to refer to any expression of the general form: ‘If A were the case, then B would be the case’, and in order to be counterfactual or contrary to fact, A must be false or untrue in the world. Examples abound. ‘If kangaroos had no tails, they would topple over’ [7]. ‘If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone’ [14]. Perhaps the most obnoxious counterfactuals in any language are those of the form: ‘If I were you, I would. . .’

Lewis [6] observed the connection between counterfactual conditionals and references to causation. He finds these logical constructions in the language used by Hume in his famous discussion of causation.

Hume defined causation twice over. He wrote, ‘We may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*’ ([6], italics are Lewis’s).

Lewis draws attention to the comparison between the factual first definition where one object is followed by another and the counterfactual second definition where, counterfactually, it is supposed that if the first object ‘had not been’, then the second object would not have been either.

It is the connection between counterfactuals and causation that makes them relevant to behavioral science research. From the point of view of some authors, it is difficult, if not impossible, to give causal interpretations to the results of statistical calculations without using counterfactual language, for example [3, 10, 11, 14]. Other writers are concerned that using such language gives an emphasis to unobservable entities that is inappropriate in the analysis and interpretation of empirical data, for example [2, 15]. The discussion here accepts the former view and regards the use of counterfactuals as a key element in the causal interpretation of statistical calculations.

A current informal use of the term counterfactual is as a synonym for control group or comparison. What is the counterfactual? This usage usually just means, ‘To what is the treatment being compared?’ It is a sensible question because the effect of a treatment is always relative to some other treatment. In behavioral science research, this simple observation can be

quite important, and ‘What is the counterfactual?’ is always worth asking and answering.

I begin with a simple observation. Suppose that we find that a student’s test performance changes from a score of X to a score of Y after some educational intervention. We might then be tempted to attribute the pretest–posttest change, $Y - X$ to the intervening educational experience – for example, to use the gain score as a measure of the improvement due to the intervention. However, this is behavioral science and not the tightly controlled ‘before-after’ measurements made in a physics laboratory. There are many other possible explanations of the gain, $Y - X$. Some of the more obvious are: simple maturation, other educational experiences occurring during the relevant time period, and differences in either the tests or the testing conditions at pre- and posttests. Cook and Campbell [1] provide a classic list of ‘threats to internal validity’ that address many of the types of alternative explanations for apparent causal effects of interventions (see **Quasi-experimental Designs**).

For this reason, it is important to think about the real meaning of the attribution of cause. In this regard, Lewis’s discussion of Hume serves us well. From it we see that what is important is what the value of Y *would have been* had the student *not had* the educational experiences that the intervention entailed. Call this score value Y^* . Thus, enter counterfactuals. Y^* is not directly observed for the student, for example, he or she *did have* the educational intervention of interest, so asking for what his or her posttest score *would have been* had he or she *not had it* is asking for information collected under conditions that are contrary to fact. Hence, it is not the difference $Y - X$ that is of causal interest, but the difference $Y - Y^*$, and the gain score has a causal significance only if X can serve as a substitute for the counterfactual Y^* . In physical-science laboratory experiments, this *counterfactual substitution* is often easy to make, but it is rarely believable in many behavioral science applications of any consequence. In fact, justifying the substitution of data observed on a control or comparison group for what the outcomes would have been in the treatment group had they not had the treatment, that is, justifying the counterfactual substitution, is the key issue in all of causal inference.

A formal statistical model for discussing the problem of estimating causal effects (of the form $Y - Y^*$ rather than $Y - X$) was developed by Neyman [8, 9] (for randomized experiments), and

2 Counterfactual Reasoning

by Rubin [13, 14] (for a wider variety of causal studies). This formal model is described in [3] and compared to the more usual statistical models that are appropriate for descriptive rather than causal inference. This approach to defining and estimating causal effects has been applied to a variety of research designs by many authors including [4, 5, 12] and the references therein.

References

- [1] Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston.
- [2] Dawid, A.P. (2000). Causal Inference without counterfactuals (with discussion), *Journal of the American Statistical Association* **95**, 407–448.
- [3] Holland, P.W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–970.
- [4] Holland, P.W. (1988). Causal inference, path analysis and recursive structural equations models, in *Sociological Methodology*, C. Clogg, ed., American Sociological Association, Washington, pp. 449–484.
- [5] Holland, P.W. & Rubin, D.B. (1988). Causal inference in retrospective studies, *Evaluation Review* **12**, 203–231.
- [6] Lewis, D. (1973a). Causation, *Journal of Philosophy*. **70**, 556–567.
- [7] Lewis, D. (1973b). *Counterfactuals*, Harvard University Press, Cambridge.
- [8] Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes, *Roczniki Nauk Rolniczki* **10**, 1–51; in Polish: English translation by Dabrowska, D. & Speed, T. (1991). *Statistical Science* **5**, 463–480.
- [9] Neyman, J. (1935). Statistical problems in agricultural experimentation, *Supplement of the Journal of the Royal Statistical Society* **2**, 107–180.
- [10] Robins, J.M. (1985). A new theory of causality in observational survival studies-application of the healthy worker effect, *Biometrics* **41**, 311.
- [11] Robins, J.M. (1986). A New approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect, *Mathematical Modelling* **7**, 1393–1512.
- [12] Robins, J.M. (1997). Causal inference from complex longitudinal data, in *Latent Variable Modeling with Applications to Causality*, M. Berkane, ed., Springer-Verlag, New York, pp. 69–117.
- [13] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.
- [14] Rubin, D.B. (1978). Bayesian inference for casual effects: the role of randomization, *Annals of Statistics* **6**, 34–58.
- [15] Shafer, G. (1996). *The Art of Causal Conjecture*, MIT Press, Cambridge.

PAUL W. HOLLAND

Counternull Value of an Effect Size

ROBERT ROSENTHAL

Volume 1, pp. 422–423

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Counternull Value of an Effect Size

The counternull value of any obtained **effect size** was introduced as a new statistic in 1994 [5]. Its purpose was to aid our understanding and reporting of the results of research. One widespread error in the understanding and reporting of data in the behavioral sciences has been the tendency to equate failure to reject the null hypothesis with evidence for the truth of the null hypothesis. For example, suppose an investigator compares the results of a new intervention with the results of the standard intervention in a randomized experiment. Suppose further that the mean benefit scores for the participants were 5 units and 2 units for the new versus old interventions, respectively, with a within-group standard deviation for each condition of 6.0. A commonly used index of effect size is the difference between the means of the two groups being compared divided by the within-group standard deviation or $(\text{Mean}_1 - \text{Mean}_2)/SD$ [1, 2], and [3]. For our present example then, our effect size, Hedges's g , would be $(5 - 2)/6.0 = 0.50$ or a one-half unit of a standard deviation, a quite substantial effect size. Further, suppose that each of our conditions had a dozen participants, that is, $n_1 = n_2 = 12$ so the value of the t Test comparing the two conditions would be 1.22 with a one-tailed p of .12, clearly 'not significant' by our usual criteria. At this point, many investigators would conclude that, given a null hypothesis that could not be rejected, 'there is no difference between the two conditions.' That conclusion would be incorrect.

For the frequent situation in which the value of the condition differences is 0.00 under the null hypothesis, the counternull value of the obtained effect size is simply $2 \times$ effect size or $2g = 2(0.50) = 1.00$ in our example. The interpretation is that, on the basis of our obtained effect size, the null hypothesis that $g = 0.00$ in the population from which we have sampled is no more likely to be true than that the population value of $g = 1.00$. Indeed, a population value of $g = 0.99$ is *more* likely to be true than that the population value of $g = 0.00$.

The obtained effect size estimate always falls between the null value of the effect size and the

counternull value. In addition, for any effect size estimate that is based on symmetric distributions such as the normal or t distributions (e.g., g , d , Δ) (see **Catalogue of Probability Density Functions**), the obtained effect size falls exactly halfway between the null value of the effect size and the counternull value. For these effect size estimates, the null value is typically 0.00, and in such cases, the counternull value is simply twice the obtained effect size. In general, when dealing with effect sizes with asymmetric distributions (e.g., Pearson's correlation r), it is best to transform the effect size to have a symmetric distribution (e.g., r should be transformed to Fisher's Z_r). After calculating the counternull on the symmetric scale of Z_r , we then transform back to obtain the counternull on the original scale of r .

In this framework, when we note that 'the obtained effect size (0.50 in our example) is not significantly different from the null value' (0.00 in our example), the counternull value forces us to confront the fact that though the assertion is true, it is no more true than the assertion that 'the obtained effect size is not significantly different from the counternull value' (1.00 in this example).

(1) shows a general procedure for finding the counternull value of the effect size for any effect size with a symmetric reference distribution (e.g., the normal or t distribution) no matter what the effect size (ES) magnitude is under the null:

$$ES_{\text{counternull}} = 2ES_{\text{obtained}} - ES_{\text{null}}. \quad (1)$$

Since the effect size expected under the null is zero in so many applications, the value of the counternull is often simply twice the obtained effect size or $2ES_{\text{obtained}}$.

An example in which the null effect size is not zero might be the case of a study testing a new treatment against a placebo control. The currently standard treatment is known to have an average effect size d of 0.40. The null hypothesis is that the new treatment is not different from the old. The d obtained for the new treatment is 0.60 but, with the size of study employed, that d was not significantly greater than the standard d of 0.40. The counternull value of the effect size in this example, however, is $2ES_{\text{obtained}} - ES_{\text{null}} = 2(0.60) - 0.40 = 0.80$. The evidence for a d of 0.80, therefore, is as strong as it is for a d of 0.40.

The counternull value of an effect size can be employed in other contexts as well. For example,

2 Counternull Value of an Effect Size

sometimes when sample sizes are very large, as in some **clinical trials**, highly significant results may be associated with very small effect sizes. In such situations, when even the counternull value of the obtained effect size is seen to be of no practical import, clinicians may decide there is insufficient benefit to warrant introducing a new and possibly very expensive intervention. Finally, it should be noted that the counternull values of effect sizes can be useful in multivariate cases, as well as in contrast analyses [4] and [5].

References

- [1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.
- [2] Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators, *Journal of Educational Statistics* **6**, 107–128.
- [3] Rosenthal, R. (1994). Parametric measures of effect size, in *Handbook of Research Synthesis*, H. Cooper & L.V. Hedges eds, Russell Sage Foundation, New York, pp. 231–244.
- [4] Rosenthal, R., Rosnow, R.L. & Rubin, D.B. (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*, Cambridge University Press, New York.
- [5] Rosenthal, R. & Rubin, D.B. (1994). The counternull value of an effect size: a new statistic, *Psychological Science* **5**, 329–334.

ROBERT ROSENTHAL

Covariance

DIANA KORNBROT

Volume 1, pp. 423–424

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Covariance

Covariance is a descriptive measure of the extent to which two variables vary together. Figure 1 shows an example from UK women's athletic records from 1968 to 2003 [2]. Shorter 1500 m times go together with longer high-jump distances (Figure 1a), but shorter 200 m times (Figure 1b). There is no direct causal link – covariance does not imply causality. However, performance in all events tends to improve with time, thus leading to associations, such that shorter times in running events go with longer distances in jump events.

The covariance unit of measurement is the product of the units of measurement of the two variables. Thus, covariance changes if the unit of measurement is changed, say from seconds to minutes. Covariances in different units (a) cannot be meaningfully compared and (b) give no idea of how well a straight line fits the data. For example, the data in Figure 1(a) gives covariance (high jump, 1500 m) = $-0.0028 \text{ m min} = -0.17 \text{ m sec}$, with Pearson's $r = -0.53$. The data in Figure 1(b) gives covariance (200 m, 1500 m) = $+0.017 \text{ sec min} = 1.03 \text{ sec}^2$, with Pearson's $r = +0.59$.

Calculation

The covariance of N pairs of variables X and Y , cov_{XY} , is defined by (1) [1]:

$$\text{cov}_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}. \quad (1)$$

However, it is most directly and accurately computed using (2):

$$\text{cov}_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N - 1}. \quad (2)$$

Uses of Covariance and Relation to Pearson's r

The main use of covariance is as a building block for **Pearson's product moment correlation coefficient**, r , which measures the covariance relative to the **geometric mean** of the standard deviations of X and Y , s_X , s_Y , as shown in (3),

$$r = \frac{\text{cov}_{XY}}{s_X s_Y} \text{ or } r^2 = \frac{\text{cov}_{XY}^2}{\text{var}_X \text{var}_Y}. \quad (3)$$

Pearson's r has the advantage of being independent of the unit of measurement of X and Y . It is also the basis of hypothesis tests of whether the relation between Y and X is statistically significant. Consequently, Pearson's r is used when comparing the extent to which pairs of variables are associated. A higher magnitude of r (positive or negative) always indicates a stronger association. Equally useful, r^2 , the proportion of variance accounted for by the linear relation between X and Y , is a measure of covariance

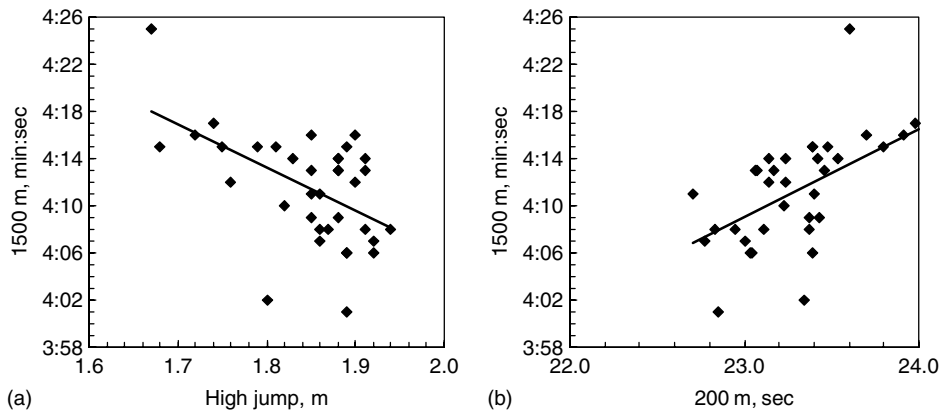


Figure 1 UK women's athletic performance from 1968 to 2003 to demonstrate covariance. (a) shows negative relation between 1500 m and high-jump performance. (b) shows positive relation between 1500 m and 200 m performance

2 Covariance

relative to the variance of X and Y , var_X , var_Y (as shown in the alternative form of (3)).

When there are more than two variables, the covariance matrix is important. It is a square matrix with variance of i th variable on the diagonal and covariance of i th and j th variable in column i row j . The covariance matrix underpins multivariate analyses, such as **multiple linear regression** and **factor analysis**.

Covariance is only useful as a stepping stone to correlation or multivariate analyses.

Acknowledgment

Thanks to Martin Rix who maintains pages for United Kingdom Track and Field at gbrathletics.com.

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.
- [2] Rix, M. (2004). *AAA championships (Women)*, United Kingdom Track and Field, [gbrathletics.com](http://www.gbrathletics.com/bc/waaa.htm).
<http://www.gbrathletics.com/bc/waaa.htm>

DIANA KORNBROT

Covariance Matrices: Testing Equality of

WOJTEK J. KRZANOWSKI

Volume 1, pp. 424–426

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Covariance Matrices: Testing Equality of

Parametric statistical inference requires distributional assumptions to be made about the observed data, and most multivariate inferential procedures assume that the data have come from either one or more multivariate normal populations (*see Catalogue of Probability Density Functions*). Moreover, in hypothesis-testing situations that involve more than one population, a further assumption on which the tests depend is that the dispersion matrices are equal in all populations. This assumption is made, for example, in Hotelling's T^2 test for equality of mean vectors in two populations and in all **multivariate analysis of variance** (MANOVA) situations, including the one-way case in which equality of mean vectors in $g(>2)$ populations is tested. Furthermore, the same assumption is made in various descriptive multivariate techniques such as **discriminant analysis** and **canonical discriminant analysis**. In any practical situation, of course, the sample **covariance matrices** will only be estimates of the corresponding population dispersion matrices and will exhibit sampling variability. So it is important to have a reliable test that determines whether a set of sample covariance matrices could have come from multivariate normal populations with a common dispersion matrix.

Suppose that we have p -variate samples of sizes $n_1, n_2, \dots, n_g$ from g multivariate normal populations, and that $\mathbf{S}_i$ is the unbiased (i.e., divisor $n_i - 1$) sample covariance matrix for population i ($i = 1, \dots, g$). Let $\mathbf{S}$ be the pooled within-sample covariance matrix, that is, $(N - g)\mathbf{S} = \sum_{i=1}^g (n_i - 1)\mathbf{S}_i$ where $N = \sum_{i=1}^g n_i$. Wilks [4] was the first to give a test of the null hypothesis that all the population dispersion matrices are equal against the alternative that at least one is different from the rest, and his test is essentially the one obtained using the likelihood ratio principle. The test statistic is $M = (N - g) \log_e |\mathbf{S}| - \sum_{i=1}^g (n_i - 1) \log_e |\mathbf{S}_i|$ where $|\mathbf{D}|$ denotes the determinant of matrix $\mathbf{D}$, and if the null hypothesis of equality of all population matrices is true then for large samples M has an approximate Chi-squared distribution on $f_1 = p(p + 1)(g - 1)/2$ degrees of freedom. Box [1] showed that a more accurate approximation to this distribution is obtained on multiplying M

by the factor $1 - (2p^2 + 3p - 1)/(6(p + 1)(g - 1))(\sum_{i=1}^g 1/(n_i - 1) - 1/(N - g))$. Values of the test statistic greater than the upper $100\alpha\%$ critical point of the Chi-squared distribution on f_1 degrees of freedom imply that the null hypothesis should be rejected in favor of the alternative at the $100\alpha\%$ significance level. Commonly α is chosen to be 0.05 (for a test at the 5% significance level), in which case the critical point is the value above which 5% of the Chi-squared (f_1) distribution lies. This is the most common test used for this situation, but an alternative approximation to the null distribution, and references to associated tables of critical values, are given in Appendix A8 of [2].

To illustrate the calculations, consider the example given by Morrison [3] relating to reaction times of 32 male and 32 female subjects when given certain visual stimuli. For two of the variables in the study, the unbiased covariance matrices for the males and females were $\mathbf{S}_1 = \begin{pmatrix} 4.32 & 1.88 \\ 1.88 & 9.18 \end{pmatrix}$ and $\mathbf{S}_2 = \begin{pmatrix} 2.52 & 1.90 \\ 1.90 & 10.06 \end{pmatrix}$, respectively. This gives the pooled covariance matrix $\mathbf{S} = \begin{pmatrix} 3.42 & 1.89 \\ 1.89 & 9.62 \end{pmatrix}$, and the determinants of the three matrices are found to be $|\mathbf{S}_1| = 36.123$, $|\mathbf{S}_2| = 21.741$, $|\mathbf{S}| = 29.328$. Here, $p = 2$, $g = 2$ and $N = 64$, so $M = 62 \log_e 29.328 - 31(\log_e 36.123 + \log_e 21.741) = 2.82$ and the Box correction factor is $1 - (2 \times 4 + 3 \times 2 - 1)/(6 \times 3 \times 1) \times (1/31 + 1/31 - 1/62) = 1 - 13/18 \times 3/62 = 359/372 = 0.965$. The adjusted value of the test statistic is thus $2.82 \times 0.965 = 2.72$. We refer to a Chi-squared distribution on $(1/2) \times 2 \times 3 \times 1 = 3$ degrees of freedom, the upper 5% critical value of which is 7.81. The adjusted test statistic value is much less than this critical value, so we conclude that the two sample covariance matrices could indeed have come from populations with a common dispersion matrix.

In conclusion, however, it is appropriate to make some cautionary comments. A formal significance test has limited value, and a nonsignificant result does not justify the blind acceptance of the null hypothesis. The data should still be examined critically and if necessary transformed or edited by removal of outliers or other dubious points. It should also be remembered that the above test makes the vital assumption of normality of data, so a significant result may be just as much an indication of nonnormality as

2 Covariance Matrices

of heterogeneity of dispersion matrices. The bottom line is that all significance tests should be interpreted critically and with caution.

References

- [1] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.
- [2] Krzanowski, W.J. & Marriott, F.H.C. (1994). *Multivariate Analysis Part 1: Distributions, Ordination and Inference*, Edward Arnold, London.

- [3] Morrison, D.F. (1990). *Multivariate Statistical Methods*, 3rd Edition, McGraw-Hill, New York.
- [4] Wilks, S.S. (1932). Certain generalisations in the analysis of variance, *Biometrika* **24**, 471–494.

(See also **Multivariate Analysis: Overview**)

WOJTEK J. KRZANOWSKI

Covariance Structure Models

Y.M. THUM

Volume 1, pp. 426–430

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Covariance Structure Models

Problems with Measuring Constructs

Constructs, represented by **latent variables** in statistical models, abound in scientific theories but they are especially commonplace in social research. Unlike fundamental properties such as length, mass, or time, measuring a construct is, however, difficult. One reason is that constructs, such as ‘mathematics ability’ in educational research or ‘consumer confidence’ in economics, are not directly observable, and only their indicators are available in practice. Consequently, a teacher may gauge a pupil’s mathematics ability only indirectly through, for example, the pupil’s performance on a pencil and paper test of mathematics, or from the pupil’s returned homework assignments. In economics, consumer confidence is discussed frequently in terms of what is observable about spending behavior such as actual tallies of sales receipts, or plans for major purchases in the near term.

Furthermore, indicators are imperfect measures of a construct. First, different indicators tap nonoverlapping aspects of the hypothesized construct. Researchers thus need to consider more than a few indicators to represent, by triangulation, the construct more completely. The teacher in our example should therefore thoughtfully balance test results with homework grades, and perhaps other pertinent sources of information as well, when evaluating a pupil’s ability. The economist surveys a broad mix of consumer spending attitudes and behaviors in tracking consumer confidence. Second, an indicator may contain information specific to itself but which is unrelated to the construct in question. Returning to our classroom assessment example above, while both the test score and performance on homework assignments provide corroborative evidence of mathematics ability, they are clearly two different forms of assessments and thus may each exert their own influence on the outcome. It is quite plausible, for example, that test scores were depressed for some students because the test was completed under time pressure, while the same students were not impacted in a similar way in their homework assignments.

The issues become considerably more challenging when we move from measuring a single construct to understanding the relationships among several constructs. Consider the enduring interest in psychology and education of measuring individual traits and behaviors using multiple items, or multiple tests, in a test battery for the purposes of validating constructs. If individual tests, subtests, or test-lets assembled in a test battery intended for gauging subject matter mastery are affected by incidental features such as presentation or response format, it is clearly critical for establishing the veracity of the test battery that we separate out information that is peculiar to presentation and response formats from information pertaining to subject matter. Elsewhere, research on halo effects, test form factors, and response sets reflects concern for essentially the same set of analytic issues.

Covariance Structure Model

A systematic assessment of the relative impact of known or hypothesized item or test features on the subject’s performance may be facilitated, first, by comparing within the respondent item or test effects using experimental design principles (e.g., [10]), a suggestion which Bock [1] attributes to **Sir Cyril Burt** [4]. Guilford’s [7] study of the structure of the intellect analyzed subject responses to an incomplete cross-classification of test items. Another prominent example of this approach to understanding measures is the classic attempt of Campbell and Fiske [5] to detect method bias in a test validation study by examining the sample correlation matrix for a set of tests of several traits assessed by alternative methods in a crossed factorial design. Second, an adequately justified statistical evaluation of the data from such designs is needed in place of the rudimentary comparisons of variances and covariances typified, for example, by the suggestions offered in [5] for exploring their ‘multitrait–multimethod’ matrix (*see also* [24] and **Multitrait–Multimethod Analyses**).

Covariance structure analysis, originally introduced by Bock [1] and elaborated in Bock and Bargmann [2], is a statistical method for the structural analysis of the sample variance-covariance matrix of respondent scores on a test battery aimed at

2 Covariance Structure Models

understanding the make-up of the battery. Covariance structure analysis is based on the familiar linear model

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{A}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

for a set of p continuous measures, $\mathbf{y}_i$, observed for a sample of subjects $i = 1, 2, \dots, N$ drawn randomly from a single population.

Considering each term in model (1) in turn, $\boldsymbol{\mu}$ is the population mean, a quantity that is typically ignored for analyses with arbitrary measurement scales. A notable exception arises of course in studies involving repeated measurements, in which case it is meaningful to compare scores from p different administrations of the same test. $\boldsymbol{\xi}_i$ is the $m \times 1$ vector of latent scores for subject i , which is assumed to be distributed multinormal as $N_m(\mathbf{0}, \boldsymbol{\Phi})$. They locate each subject in the latent space spanned presumably by the m constructs. $\mathbf{A}$ is a $p \times m$, for $m \leq p$, matrix of known constants of rank $\ell \leq m$. Each element, a_{jk} , indicates the contribution of the k th latent component ($k = 1, 2, \dots, m$) to the j th observed measure ($j = 1, 2, \dots, p$). $\boldsymbol{\varepsilon}_i$ is the $p \times 1$ vector for subject i representing measurement error and is also assumed to be distributed multinormal as $N_p(\mathbf{0}, \boldsymbol{\Psi})$. Under these assumptions, and conditional on $\boldsymbol{\xi}_i$, $\mathbf{y}_i$ is multinormal with population mean $\boldsymbol{\mu}$ and population variance-covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi}. \quad (2)$$

In cases where $\mathbf{A}$ is rank-deficient (i.e., $\ell < m$), only ℓ linearly independent combinations of the m latent variables, $\boldsymbol{\xi}_i$, are estimable. Bock and Bargmann [2] suggested a reparameterization that takes $\boldsymbol{\xi}_i$ into $\boldsymbol{\theta}_i$ ($\ell \times 1$), by choosing a set of ℓ linear contrasts $\mathbf{L}$ ($\ell \times m$) such that $\boldsymbol{\theta}_i = \mathbf{L}\boldsymbol{\xi}_i$. Given any choice for $\mathbf{L}$, model (1) is then

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{K}\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \quad (3)$$

from setting $\mathbf{A}\boldsymbol{\xi}_i = \mathbf{K}\mathbf{L}\boldsymbol{\xi}_i = \mathbf{K}\boldsymbol{\theta}_i$ and solving $\mathbf{K} = \mathbf{A}\mathbf{L}'(\mathbf{L}\mathbf{L}')^{-1}$. Writing the variance-covariance matrix of $\boldsymbol{\theta}_i$ as $\boldsymbol{\Phi}^* = \mathbf{L}\boldsymbol{\Phi}\mathbf{L}'$, the population variance-covariance matrix for a given $\mathbf{L}$ is $\boldsymbol{\Sigma} = \mathbf{K}\boldsymbol{\Phi}^*\mathbf{K}' + \boldsymbol{\Psi}$. Jöreskog [9] noted that reparameterization in rank-deficient cases is an alternative to imposing equality constraints on $\boldsymbol{\Phi}$ where reasonable.

Specific Covariance Models

From (1) and (2), it is clear that the covariance structure model is a **confirmatory factory model** with known factory loadings [2]. Covariance structure analysis focuses on estimating and testing *a priori* hypotheses induced by $\mathbf{A}$ on the variance-covariance matrix of the latent variables, $\boldsymbol{\Phi}$, under alternative specifications of both $\boldsymbol{\Phi}$ and the measurement error components represented by $\boldsymbol{\Psi}$. In Bock [1] and Bock and Bargmann [2], these covariance matrices are uncorrelated, but they may be either homogeneous or heterogeneous. In a specification designated as Case I in Bock and Bargmann [2] the latent variables are uncorrelated and the error are uncorrelated and homogeneous, that is, $\boldsymbol{\Phi} = (\varphi_1^2, \varphi_2^2, \dots, \varphi_k^2, \dots, \varphi_m^2)$ and $\boldsymbol{\Psi} = \psi^2\mathbf{I}$. A special Case I model occurs when all the items in a battery tap a single latent variable, leading to the so-called ‘true-score’ model with compound symmetry ($\mathbf{A} = \mathbf{1}$, $\boldsymbol{\Phi} = \phi^2\mathbf{I}$, $\boldsymbol{\Psi} = \psi^2\mathbf{I}$) with population variance-covariance matrix $\boldsymbol{\Sigma} = \phi^2\mathbf{1}\mathbf{1}' + \psi^2\mathbf{I}$. A Case II covariance structure model covers applications in which the latent variables are uncorrelated and the errors are uncorrelated but are heterogeneous, that is, $\boldsymbol{\Psi} = (\psi_1^2, \psi_2^2, \dots, \psi_j^2, \dots, \psi_p^2)$.

In a Case III model, the latent variables are uncorrelated and the errors are also uncorrelated and homogeneous as in Case I, but only after scale differences are removed. Case III assumptions lead to a generalization of model (1) in which the population variance-covariance matrix takes the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})\boldsymbol{\Gamma}, \quad (4)$$

for the diagonal matrix of unknown scaling constants, $\boldsymbol{\Gamma}$. Models with scale factors may be important in social research because, as noted above, measurement scales are most often arbitrary. In practice, however, achieving identifiability requires fixing one of the unknown scaling coefficients to an arbitrary constant, or by setting $\boldsymbol{\Psi} = \psi^2\mathbf{I}$ [22].

Note that other forms for $\boldsymbol{\Phi}$ may be plausible but its identification will depend on the available data. Wiley et al. [22] considered a model with correlated factors $\boldsymbol{\xi}_i$ in a study of teaching practices which we will reexamine below.

Estimation and Testing

Given the population variance-covariance matrix (4) and its sample estimate $\mathbf{S}$, Bock and Bargmann [2]

provided a **maximum likelihood** solution based on the log-likelihood

$$\ln L = -\frac{Np}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| + N \ln |\Gamma^{-1}| - \frac{N}{2} \text{tr}(\Sigma^{-1} \Gamma^{-1} S \Gamma^{-1}). \quad (5)$$

It is clear from our discussion above that a covariance structure model belongs to the broader class of **structural equation models** (SEM). Consequently, widely available programs for SEM analyses, such as AMOS, EQS, LISREL, MPLUS, and SAS PROC CALIS (*see Structural Equation Modeling: Software*), will routinely provide the necessary estimates, standard errors, and a host of fit statistics for fitting covariance structure models.

A Study of Teaching Practices

Wiley, Schmidt, and Bramble [22] examined data from [17], consisting of responses from 51 students to a test with items sharing combinations of three factors thought to influence classroom learning situations and teaching practices (see Table 1). The study sought to compare conditions in first and sixth grade classrooms, teaching styles that were deemed teacher-centered as opposed to pupil-centered, and teaching methods that were focused on drill as opposed to promoting discovery. The eight subtests comprised a 2^3 factorial design.

Following Wiley, Schmidt, and Bramble [22], we parameterize an overall latent component, a contrast between grade levels, a contrast between teaching styles, and teaching methods represented seriatim by

the four columns in design matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \quad (6)$$

Results from SAS PROC CALIS support [22] conclusion that a model with correlated latent components estimated by

$$\hat{\Phi} = \begin{bmatrix} 9.14 (1.92) & & & \\ 0.73 (0.48) & 0.68 (0.23) & & \\ 0.63 (0.42) & -0.06(0.15) & 0.43 (0.19) & \\ -0.61(1.05) & -0.51(0.37) & 1.13 (0.35) & 5.25 (1.14) \end{bmatrix} \quad (7)$$

and heterogeneous error variances estimated by $\hat{\Psi} = \text{diag}[1.63 (0.90), 5.10 (1.40), 8.17 (1.90), 5.50 (1.56), 1.93 (0.91), 2.33 (0.84), 5.79 (1.44), 2.55 (0.93)]$ indicated an acceptable model fit ($\chi^2_{18} = 25.24$, $p = 0.12$; CFI = 0.98; RMSEA = 0.09) superior to several combinations of alternative forms for Φ and Ψ .

In comparing the relative impact of the various factors in teachers' judgments on conditions that facilitated student learning, Wiley et al. [22, p. 322], however, concluded erroneously that teaching style did not affect teacher evaluations due, most likely, to a clerical error in reporting 0.91 as the standard error estimate for $\hat{\phi}_3^2 = 0.43$ instead of 0.19. The results therefore suggested that, on the contrary, both teaching practices influenced the performance of subjects. Of particular interest to research on teaching, the high

Table 1 Miller–Lutz sample variance–covariance matrix

Grade	Style	Method								
1	Teacher	Drill	18.74							
1	Teacher	Discovery	9.28	18.80						
1	Pupil	Drill	15.51	7.32	21.93					
1	Pupil	Discovery	3.98	15.27	4.10	26.62				
6	Teacher	Drill	15.94	4.58	13.79	−2.63	19.82			
6	Teacher	Discovery	7.15	13.63	3.86	15.33	3.65	16.81		
6	Pupil	Drill	11.69	6.05	10.18	1.13	13.55	5.72	16.58	
6	Pupil	Discovery	2.49	12.35	0.03	16.93	−0.86	14.33	2.99	18.26

4 Covariance Structure Models

positive correlation between teacher approach and teacher method ($\hat{\rho}_{34} = 0.75$) suggested that combining a teacher-centered approach with drill produced dramatically different responses among subjects when compared with subjects engaged in pupil-centered discovery.

Relationships with Other Models

Covariance structure analysis generalizes the conventional mixed-effects model for a design with one random way of classification (subjects or respondents) and a possibly incomplete fixed classification (tests) by providing tests for a more flexible variance-covariance structure. As a statistical model, its application extends beyond the immediate context of tests and testing to include all problems in which the design on the observed measures is known.

For some applications, the matrix of coefficients $\mathbf{A}$ may be structured in a way to produce a hypothesized covariance matrix for the latent variables. For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (8)$$

induces the Guttman's [8] simplex pattern with correlation $\rho^{|j-j'|}$ among a set of $p = 4$ ordered tests ($j = 1, 2, \dots, p; j' = 1, 2, \dots, p$).

Agreement between results from a covariance structure analysis and Generalizability theory should not be surprising when the models considered are equivalent (see, for example, the excellent review by Linn and Werts [11]). As an example, Marcoulides [14] considered the flexibility of a covariance structure analysis for Generalizability analysis (see **Generalizability**), with an illustration involving four job analysts who each provided job satisfaction ratings on 27 described occupations on two occasions. This application employed a design with a random way of classification (job descriptions) and a crossed-classification of raters and occasions. The proposed

model with

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}, \quad \Phi = \begin{bmatrix} \varphi_1^2 & & & & & & \\ & \varphi_2^2 & & & & & \\ & & \varphi_2^2 & & & & \\ & & & \varphi_2^2 & & & \\ & & & & \varphi_2^2 & & \\ & & & & & \varphi_3^2 & \\ & & & & & & \varphi_3^2 \end{bmatrix} \quad (9)$$

and $\Psi = \psi^2 \mathbf{I}$, estimated an overall variance among job categories (φ_1^2), constrained variability to be equal among raters (φ_2^2), constrained variability between the two occasions (φ_3^2) to be equal, and assumed that error variances to be homogeneous (ψ^2). Estimation and model comparison results via maximum likelihood using many of the software available for fitting SEMs gave results comparable to the more usual Analysis of Variance (ANOVA) solution (see [3, 19]).

Willett and Sayer [23], reviewing the work of McArdle and Epstein [15] and Meredith and Tisak [16], noted that mixed-effects models for growth are implicit in covariance structure analysis of complete repeated measurement designs. SEM software provided the overarching modeling framework for balanced growth data: A conventional growth model is estimated when the design on time is known, but when some coefficients of the design matrix for growth are unknown, a latent curve model is considered.

This reading is of course consistent with earlier attempts within the SEM literature to extend the applications to multilevel multivariate data (e.g., [6, 13, 18] and **Structural Equation Modeling: Multilevel**). Note that the models considered in Bock [1], Bock and Bargmann [2], Wiley, Schmidt, and Bramble [22] dealt with the covariance structure for multivariate outcomes of unreplicated sampled units. If replications for subjects are present and the replications are small and equal (balanced), covariance structure models need no modification because replications may be treated as part of a subject's

single observation vector. However, when replications within subjects are unbalanced, that is, for each subject i we observed a $n_i \times p$ matrix $\mathbf{Y}_i = [\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \dots, \mathbf{y}'_{ir}, \dots, \mathbf{y}'_{in_i}]$, the covariance structure model (1) takes the extended form of a multivariate mixed-effects model

$$\mathbf{Y}_i = \mathbf{\Xi}_i \mathbf{A} + \mathbf{E}_i \quad (10)$$

treated, for example, in Thum [20, 21] in Littell, Milliken, Stroup, and Wolfinger [12], and in more recent revisions of SEM and multilevel frameworks (see **Generalized Linear Mixed Models**).

References

- [1] Bock, R.D. (1960). Components of variance analysis as a structural and discriminant analysis for psychological tests, *British Journal of Statistical Psychology* **13**, 151–163.
- [2] Bock, R.D. & Bargmann, R.E. (1966). Analysis of covariance structures, *Psychometrika* **31**, 507–533.
- [3] Brennan, R.L. (2002). *Generalizability Theory*, Springer-Verlag, New York.
- [4] Burt, C. (1947). Factor analysis and analysis of variance, *British Journal of Psychology Statistical Section* **1**, 3–26.
- [5] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
- [6] Goldstein, H. & McDonald, R.P. (1988). A general model for the analysis of multilevel data, *Psychometrika* **53**, 455–467.
- [7] Guilford, J.P. (1956). The structure of intellect, *Psychological Bulletin* **53**, 276–293.
- [8] Guttman, L.A. (1954). A new approach to factor analysis: the radex, in *Mathematical Thinking in the Social Sciences*, P.F. Lazarsfeld, ed., Columbia University Press, New York, pp. 258–348.
- [9] Jöreskog, K.G. (1979). Analyzing psychological data by structural analysis of covariance matrices, in *Advances in Factor Analysis and Structural Equation Models*, J. Madgison, ed., University Press of America, Lanham.
- [10] Kirk, R.E. (1995). *Experimental Design*, Brooks/Cole, Pacific Grove.
- [11] Linn, R.L. & Wert, C.E. (1979). Covariance structures and their analysis, in *New Directions for Testing and Measurement: Methodological Developments*, Vol. 4, R.E. Traub, ed., Jossey-Bass, San Francisco, pp. 53–73.
- [12] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*, SAS Institute, Cary.
- [13] Longford, N. & Muthén, B.O. (1992). Factor analysis for clustered observations, *Psychometrika* **57**, 581–597.
- [14] Marcoulides, G.A. (1996). Estimating variance components in generalizability theory: the covariance structure approach, *Structural Equation Modeling* **3**(3), 290–299.
- [15] McArdle, J.J. & Epstein, D.B. (1987). Latent growth curves within developmental structural equation models, *Child Development* **58**(1), 110–133.
- [16] Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.
- [17] Miller, D.M. & Lutz, M.V. (1966). Item design for an inventory of teaching practices and learning situations, *Journal of Educational Measurement* **3**, 53–61.
- [18] Muthén, B. & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models, in *Multilevel Analysis of Educational Data*, D.R. Bock, ed., Academic Press, San Diego, pp. 87–99.
- [19] Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*, Sage Publications, Newbury Park.
- [20] Thum, Y.M. (1994). *Analysis of individual variation: a multivariate hierarchical linear model for behavioral data*, Doctoral dissertation, University of Chicago.
- [21] Thum, Y.M. (1997). Hierarchical linear models for multivariate outcomes, *Journal of Educational and Behavioral Statistics* **22**, 77–108.
- [22] Wiley, D.E., Schmidt, W.H. & Bramble, W.J. (1973). Studies of a class of covariance structure models, *Journal of American Statistical Association* **68**, 317–323.
- [23] Willett, J.B. & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of change, *Psychological Bulletin* **116**, 363–381.
- [24] Wothke, W. (1996). Models for multitrait-multimethod matrix analysis, in *Advanced Structural Equation Modelling*, G.A. Marcoulides & R.E. Schumacher, eds, Lawrence Erlbaum, Mahwah.

(See also **Linear Statistical Models for Causation: A Critical Review; Structural Equation Modeling: Nontraditional Alternatives**)

Y.M. THUM

Covariance/variance/correlation

JOSEPH LEE RODGERS

Volume 1, pp. 431–432

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Covariance/variance/correlation

The covariance is a ‘uniting concept’ in statistics, yet is often neglected in undergraduate and graduate statistics courses. The variance and correlation, which are much more popular statistical indices, are each a special case of the covariance. **Structural equation modeling** (SEM) is the statistical method in which these three measures are most completely respected and accounted for. The current entry will describe each index in a common framework that allows their interrelationships to be apparent. Following separate treatment, a formula will be presented that explicitly shows the relationships between these three indices. Finally, the role of each in SEM will be described.

Hays [1] defines the covariance, an indicator of the relationship between two variables, as a measure of ‘departure from independence of X and Y ’. When X and Y are independent, the covariance is zero. To the extent that they are not independent, the covariance will be different from zero. If the two variables are positively related, the covariance will be positive; an inverse relationship between X and Y will result in a negative covariance. However, the covariance is sensitive to the scale of measurement of the two variables – that is, variables with large variance will have more extreme covariance values than variables with small variance – which makes the magnitude of the covariance difficult to interpret across different measurement settings. The covariance is, however, defined on centered variable values; that is, the means are subtracted from each score as they enter the covariance formula, and the resulting transformed scores are guaranteed to each have a mean of zero. In this sense, the covariance equates the means of the two variables.

It is exactly this feature of the covariance – the fact that its bounds depend on the variance of X and Y – that can be used to transform the covariance into the correlation. The correlation is often referred to informally as a ‘standardized covariance’. One of the many ways it can be defined is by computing the covariance, then dividing by the product of the two standard deviations. This adjustment rescales the covariance into an index – Pearson’s product-moment correlation coefficient – that is guaranteed to range between -1 and $+1$. The transformation results

in new variables that have their variability equated to one another, and equal to one (which means the variance and standard deviation of each variable become identical). But it is important to note that the more well-behaved measure – the correlation – no longer contains any information about either the mean or the variance of the original variables.

It is of conceptual value to note that an even broader measure than the covariance exists, in which the raw moments are defined by simply multiplying the raw scores for each X and Y value, summing them, and dividing by a function of the sample size. This measure is sensitive to both the mean and the variance of each of the two variables. The covariance, on the other hand, is sensitive only to the variance of each of the two variables. The correlation is not sensitive to either the mean or the variance. This gives the correlation the advantage of being invariant under linear transformation of either of the raw variables, because linear transformation will always return the variable to the same standardized value (often called the *z-score*). It also gives it the disadvantage that no information about either the means or the variances is accounted for within the correlational formula.

The variance is the simplest of these concepts, because it applies to a single variable (whereas the other two concepts are bivariate by definition). Other entries in this encyclopedia describe the details of how the variance is defined, so we will not develop computational details but rather will discuss it conceptually. The variance – referred to by mathematical statisticians as a function of the ‘second moment about the mean’ – measures the average deviation of each score from the mean, in squared units of the variable’s scale of measurement. The variance is often unsquared to rescale it into units interpretable in relation to the original scale, and this unsquared variance measure is called the *standard deviation*. Certain properties of the variance are of particular importance within statistics. First, the variance has an important least squares property. Because the squared deviations are defined about the mean, this computation is guaranteed to give the minimum value compared to using other measures (constants) within the formula. This optimality feature helps place the variance within normal theory, maximum likelihood estimation, and other topics in statistics. However, the variance is also highly sensitive to outliers, because the deviations from the mean are squared (and therefore magnified) within the variance formula. This concern

often leads applied statisticians to use more robust measures of variability, such as the **median absolute deviation** (MAD) statistic. Whereas the mean minimizes the sum of squared deviations compared to any other constant, the median has the same optimality property in the context of absolute deviations.

The variance, like the correlation, is a special case of the covariance – it is the covariance of a variable with itself. The variance is zero if and only if all the scores are the same; in this case, the variable can be viewed as ‘independent of itself’ in a sense. Although covariances may be positive or negative – depending on the relation between the two variables – a variable can only have a positive relationship with itself, implying that negative variances do not exist, at least computationally. Occasionally, estimation routines can estimate negative variances (in **factor analysis**, these have a special name – Heywood Cases). These can result from missing data patterns or from lack of fit of the model to the data.

The formula given in Hays to compute the covariance is $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$, which shows that the covariance measures the departure from independence of X and Y . A straightforward computational formula for the covariance shows the relationships between the three measures, the covariance, the correlation, and variance:

$$\text{cov}(X, Y) = r_{XY}s_Xs_Y. \quad (1)$$

Note that this computational formula may take slightly different forms, depending on whether the unbiased estimates or the sample statistics are used – the formula above gives the most straightforward conceptual statement of the relationship between

these three statistics. Dividing this formula through by the product of the standard deviations shows how the correlation can be obtained by standardizing the covariance (i.e., by dividing the covariance by the product of the standard deviations).

In structural equations (SEM) models (and factor analysis to a lesser extent), appreciating all three of these important statistical indices is prerequisite to understanding the theory and to being able to apply the method. Most software packages that estimate SEM models can fit a model to either covariances or to correlations – that is, these packages define predicted covariance or correlation matrices between all pairs of variables, and compare them to the observed values from real data (*see **Structural Equation Modeling: Software***). If the model is a good one, one or more of the several popular fit statistics will indicate a good match between the observed and predicted values. With a structural equation model, observed or latent variables can be linked to another observed or latent variable either through a correlation or covariance, and can also be linked back to itself through a variance. These variances may be constrained to be equal to one – implying standardized variables – or unconstrained – implying unstandardized variables. The covariances/correlations can be estimated, constrained, or fixed.

Reference

- [1] Hays, W.L. (1988). *Statistics*, Holt, Rinehart, & Winston, Chicago.

JOSEPH LEE RODGERS

Cox, Gertrude Mary

DAVID C. HOWELL

Volume 1, pp. 432–433

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cox, Gertrude Mary

Born: January 13, 1900, in Dayton, USA.

Died: October 17, 1978, in Durham, USA.

Gertrude Cox was born in the American Midwest and grew up with a strong commitment to social services. She intended to become a deaconess in the Methodist Episcopal Church, but then decided that she would prefer to follow academic pursuits. She graduated from Iowa State College in 1929 with a degree in mathematics.

After graduation, Cox stayed on at Iowa State College (now Iowa State University) to earn a master's degree in statistics under the direction of **George Snedecor**. This was the first master's degree in statistics at that institution, but far from the last.

Cox went on to the University of California at Berkeley to work on her Ph. D., but left without her degree to return to Iowa State to direct the Computing Laboratory under George Snedecor. In fact, she never did finish her Ph.D., although she was appointed an Assistant Professor of Statistics at Iowa.

In 1940, George Snedecor was asked to recommend names to chair the new Department of Experimental Statistics at North Carolina State University at Raleigh. Gertrude Cox was appointed to this position with the title of Professor of Statistics. From that time until her retirement, her efforts were devoted to building and strengthening the role of statistics in the major university centers in North Carolina.

In 1944, North Carolina established an all-University Institute of Statistics, and Gertrude Cox was selected to head it. The next year, she obtained funds to establish a graduate program in statistics at North Carolina State, and the following year, she obtained additional funds to establish a Department of Mathematical Statistics at the University of North Carolina at Chapel Hill. Not content with that, she went on to find further funding to establish the Department of Biostatistics at UNC, Chapel Hill in 1949. These departments now offer some of the best-known programs in statistics in the United States.

One of her major interests was experimental design, and she had been collecting material for a book on design since her early days at Iowa. As R. L. Anderson [1] has pointed out, Cox believed very strongly in the role of **randomization in experimental design** and in the need to estimate **power**.

In 1950, Gertrude Cox and **William Cochran** published the classic Cochran and Cox: *Experimental Design* [3].

Though she never completed her Ph. D., Iowa State bestowed on her an honorary Doctorate of Science in 1958, in recognition of the major contributions she made to her field.

In this same year, Cox and others from North Carolina State began work to create a Statistics Division within what is now Research Triangle Park, a cooperative arrangement between the universities in Raleigh, Durham, and Chapel Hill. Not surprisingly, Gertrude Cox was asked to head that division, and she retired from North Carolina State in 1960 for that purpose. She retired from the division position in 1964.

Aside from her academic responsibilities, Gertrude Cox was President of the American Statistical Association in 1956. In addition, she served as the first editor of *Biometrics* and was the president of the International Biometric Society in 1968–69. In 1975, she was elected to the National Academy of Sciences. It is worth noting that in addition to herself, four of the faculty that she hired at North Carolina were also elected to the National Academy of Sciences (Bose, Cochran, Hoeffding, and **Hotelling**) [2].

Following her retirement in 1965, Cox served as a consultant to promote statistical activities in Thailand and Egypt. In 1989, the American Statistical Association Committee on Women in Statistics and the Women's Caucus in Statistics established the Gertrude M. Cox Scholarship. It is presented annually to encourage more women to enter statistics, and has become a prestigious award.

Gertrude Cox contracted leukemia and died in 1978.

References

- [1] Anderson, R.L. (1979). Gertrude M. Cox – a modern pioneer in statistics, *Biometrics* **33**, 3–7.
- [2] Cochran, W.G. (1979). Some reflections, *Biometrics* **35**, 1–2.
- [3] Cochran, W.G. & Cox, G.M. (1950). *Experimental Design*, Wiley, New York.

DAVID C. HOWELL

Cramér–von Mises Test

CLIFFORD E. LUNNEBORG

Volume 1, pp. 434–434

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cramér–von Mises Test

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_m)$ be independent random samples from distributions with cumulative distribution functions $F(z)$ and $G(z)$, respectively. The existence of cumulative distribution functions (cdfs) implies that the scale of measurement of the random variables X and Y is at least ordinal: $F(z)$ is the probability that the random variable X takes a value less than or equal to z and thus increases from zero to one as the value of z increases.

The hypothesis to be nullified is that $F(z) = G(z)$ for all values of z . The alternative hypothesis is that the two cdfs differ for one or more values of z . The Cramér–von Mises test is not a test of equivalence of means or variances, but a test of equivalence of distributions. The test is inherently nondirectional.

Let $F^*(z)$ and $G^*(z)$ be the empirical (cumulative) distribution functions based on the samples $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_m)$ respectively. That is, $F^*(z)$ is the proportion of values in the sample $(x_1, x_2, \dots, x_n)$ that are less than or equal to z and $G^*(z)$ is the proportion of values in the sample $(y_1, y_2, \dots, y_m)$ that are less than or equal to z .

The test statistic is

$$T = k\{\sum_{j=1,2,\dots,n}[F^*(x_j) - G^*(x_j)]^2 + \sum_{j=1,2,\dots,m}[F^*(y_j) - G^*(y_j)]^2\} \quad (1)$$

where

$$k = \frac{mn}{(m+n)^2}. \quad (2)$$

The squared difference between the two empirical cdfs is evaluated at each of the $(m+n)$ values in the

combined samples. If the sum of these squared differences is large, the hypothesis of equal distributions can be nullified.

The exact null distribution of T can be found by permuting the purported sources of the observed $(m+n)$ values in all possible ways, n from random variable X and m from random variable Y , and then computing $F^*(z)$, $G^*(z)$, and T from each permutation. Where m and n are too large to make this feasible, an adequate Monte Carlo (*see Monte Carlo Simulation*) approximation to the null distribution can be created by randomly sampling the potential permutations a very large number of times, for example, 5000. An asymptotic null distribution, valid for large samples from continuous random variables, has been developed [1] and relevant tail probabilities are provided in [2].

There is a one-sample goodness-of-fit version of the Cramér–von Mises test in which a known probability distribution, for example, the normal, replaces $G^*(z)$. Both Cramér–von Mises tests are dominated in usage by the one- and two-sample **Kolmogorov–Smirnov** or Smirnov tests.

References

- [1] Anderson, T.W. & Darling, D.A. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes, *Annals of Mathematical Statistics* **23**, 193–212.
- [2] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.

CLIFFORD E. LUNNEBORG

Criterion-Referenced Assessment

RONALD K. HAMBLETON AND SHUHONG LI

Volume 1, pp. 435–440

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Criterion-Referenced Assessment

Criterion-referenced testing (CRT) was introduced in the United States in the 1960s as a response to the need for assessments that could determine what persons knew and could do in relation to a well-defined domain of knowledge and skills, rather than in relation to other persons [3, 15]. With the CRT score information, the level of proficiency of candidates can be determined, and in many cases, diagnostic information can be provided that will be helpful to candidates in working on their weaknesses. Today, the uses of CRTs are widespread in education, the military, and industry [9]. What follows first in the entry, is a brief description of the differences between norm-referenced tests (NRTs) and CRTs. It is because of the fundamental differences that a number of challenges have arisen regarding CRTs – standard-setting and estimating reliability, to name two, and it is these technical challenges that are the focus of this entry. NRTs, on the other hand, have received extensive research and development over the years, and from a technical perspective, there are few remaining challenges to overcome for their effective use.

Differences Between Norm-referenced and Criterion-referenced Tests

Criterion-referenced tests and norm-referenced tests serve different purposes and these differences have implications for test development and evaluation. Norm-referenced tests are primarily intended to distinguish or compare examinees on the construct measured by the test. Examinees are basically rank-ordered based on their test scores. For the rank ordering to be reliable, the test itself needs to spread out the examinees so that the always-present measurement errors do not distort too much the ranking that would be obtained if true scores had been used. This means that a good norm-referenced test will spread out the examinee scores, and to do that, items of middle difficulty and high discriminating power are usually best – test score variability needs to be maximized to the extent possible, given constraints on such things as test content and test length. Test score

reliability is judged by the stability of the examinee rankings or scores over parallel-form administrations or test-retest administrations of the test. Proxies for the double administration of the test or administration of parallel forms of the test come from single-administration reliability estimates such as corrected split-half and internal consistency estimates (i.e., correlation between scores derived from two halves of a test, and then adjusting the correlation upward by the Spearman–Brown formula to predict the reliability of the full-length test; or coefficient alpha for polytomous response data, and the KR-20 and KR-21 formulas for binary data). Validity is established by how well the scores serve their intended purpose. The evidence might come from criterion-related or construct validity studies (*see Validity Theory and Applications*).

CRTs, on the other hand, are intended to indicate an examinee's level of proficiency in relation to a well-defined domain of content. Usually scores are interpreted in relation to a set of performance standards that are set on the test score reporting scale. Primary focus in item selection is not on the item statistics as it is when building an NRT, though they are of concern (for example, items with negative point biserial correlations would never be selected), but rather primary focus in item selection is on the content match of items to the content domain being measured by the test. Test items are needed that insure the content validity of the test and so content is a primary consideration in item selection. That there may be limited score variability in the population of examinees is not of any significance, since examinee scores, independent of other examinees, are compared to the content domain covered by the test, and the performance standards in place for test score interpretation and test score uses.

Today with many state criterion-referenced tests, examinees, based upon their test scores, are assigned to one of four performance levels: Failing, Basic, Proficient, and Advanced. Performance standards are the points on the reporting scale that are used to sort examinees into the performance levels. For criterion-referenced credentialing exams, normally only two performance levels are used: passing and failing. Reliability is established, not by correlational statistics as is the case with NRTs, but rather by assessing the consistency of performance classifications of examinees over retests and parallel forms. Proxies for the concept of decision consistency estimated

from single administrations are also possible and will be discussed later in this entry. Validity is typically assessed by how well the test items measure the content domain to which the test scores are referenced. Validity also depends on the performance standards that are set for sorting candidates into performance categories. If they are set improperly (perhaps set too high or too low because of a political agenda of those panelists who set them), then examinees will be misclassified (relative to how they would be classified if true scores were available, and a valid set of performance standards were in place), and the validity of the resulting performance classifications is reduced.

What is unique about CRTs is the central focus on the content measured by the test, and subsequently, on how the performance standards are set, and the levels of decision consistency and accuracy of the resulting examinee classifications. These technical problems will be addressed next.

Setting Performance Standards

Setting performance standards on CRTs has always been problematic (see [1]) because substantial judgment is involved in preparing a process for setting them, and no agreement exists in the field about the best choice of methods (see **Setting Performance Standards: Issues, Methods**). One instructor may be acceptable to set performance standards on a classroom test (consequences are usually low for students, and the instructor is normally the most qualified person to set the performance standards), but when the stakes for the testing get higher (e.g., deciding who will receive a high school diploma, or a certificate to practice in a profession), multiple judges or panelists will be needed to defend the resulting performance standards. Of course, with multiple panelists and each with their own opinion, the challenge is to put them through a process that will converge on a defensible set of performance standards. In some cases, even two or more randomly equivalent panels are set up so that the replicability of the performance standards can be checked. Even multiple panels may not appease the critics: The composition of the panel or panels, and the number of panel members can become a basis for criticism.

Setting valid performance standards involves many steps (see [4]): Choosing the composition of the panel or panels and selecting a representative sample of

panel members, preparing clear descriptions of the performance levels, developing clear and straightforward materials for panels to use in the process, choosing a standard-setting method that is appropriate for the characteristics of the test itself and the panel itself (for example, some methods can only be used with multiple-choice test items, and other methods require item statistics), insuring effective training (normally, this is best accomplished with field testing in advance of the actual standard-setting process), allowing sufficient time for panels to complete their ratings and participate in discussions and revising their ratings (this activity is not always part of a standard-setting process), compiling the panelists' ratings and deriving the performance standards, collecting validity data from the panelists, analyzing the available data, and documenting the process itself.

Counting variations, there are probably over 100 methods for setting performance standards [1]. Most of the methods involve panelists making judgments about the items in the test. For example, with the Angoff method, panelists predict the expected performance of borderline candidates at the Basic cut score, the Proficient cut score, and at the Advanced cut score, on all of the items on the test. These expected item scores at a cut score are summed to arrive at a panelist cut score, and then averaged across panelists to arrive at an initial cut score for the panel. This process is repeated to arrive at each of the cut scores. Normally, discussion follows, and then panelists have an opportunity to revise their ratings, and then the cut scores are recalculated. Sometime during the process panelists may be given some item statistics, or consequences of particular cut scores that they have set (e.g., with a particular cut score, 20% of the candidates will fail). This is known as the Angoff method.

In another approach to setting performance standards, persons who know the candidates (called 'reviewers') and who know the purpose of the test might be asked to sort candidates into four performance categories: Failing, Basic, Proficient, and Advanced. A cut score to distinguish Failing from Basic on the test is determined by looking at the actual test score distributions of candidates who were assigned to either the Failing or Basic categories by reviewers. A cut score is chosen to maximize the consistency of the classifications between candidates based on the test and the reviewers. The process is then repeated for the other cut scores. This is known

as the contrasting groups method. Sometimes, other criteria for placing cut scores might be used, such as doubling the importance of minimizing one type of classification error (e.g., false positive errors) over another (e.g., false negative errors).

Many more methods exist in the measurement literature: Angoff, Ebel, Nedelsky, contrasting groups, borderline group, book-mark, booklet classification, and so on. See [1] and [5] for complete descriptions of many of the current methods.

Assessing Decision Consistency and Accuracy

Reliability of test scores refers to the consistency of test scores over time, over parallel forms, or over items within the test. It follows naturally from this definition that calculation of reliability indices would require a single group of examinees taking two forms of a test or even a single test a second time, but this is often not realistic in practice. Thus, it is routine to report single-administration reliability estimates such as corrected split-half reliability estimates and/or coefficient alpha. Accuracy of test scores is another important concern that is often checked by comparing test scores against a criterion score, and this constitutes a main aspect of validity [8].

With CRTs, examinee performance is typically reported in performance categories and so reliability and the validity of the examinee classifications are of greater importance than the reliability and validity associated with test scores. That is, the consistency and accuracy of the decisions based on the test scores outweighs the consistency and the accuracy of test scores with CRTs.

As noted by Hambleton and Slater [7], before 1973, it was common to report a KR-20 or a corrected split-half reliability estimate to support the use of a credentialing examination. Since these two indices only provide estimates of the internal consistency of examination scores, Hambleton and Novick [6] introduced the concept of the consistency of decisions based on test scores, and suggested that the reliability of classification decisions should be defined in terms of the consistency of examinee decisions resulting from two administrations of the same test or parallel forms of the test, that is, an index of reliability which reflects the consistency of classifications across repeated testing. As compared with the definition of

decision consistency (DC) given by Hambleton and Novick [6], decision accuracy (DA) is the ‘extent to which the actual classifications of the test takers agree with those that would be made on the basis of their true scores, if their true scores could somehow be known’ [12].

Methods of Estimating DC and DA

The introduction of the definition of DC by Hambleton and Novick [6] pointed to a new direction for evaluating the reliability of CRT scores. The focus was to be on the reliability of the classifications or decisions rather than on the scores themselves. Swaminathan, Hambleton, and Algina [19] extended the Hambleton–Novick concept of decision consistency to the case where there were not just two performance categories:

$$p_0 = \sum_{i=1}^k p_{ii} \quad (1)$$

where p_{ii} is the proportion of examinees consistently assigned to the i -th performance category across two administrations, and k is the number of performance categories. In order to correct for chance agreement, based on the kappa coefficient (*see Rater Agreement – Kappa*) by Cohen [2], which is a generalized proportion agreement index frequently used to estimate inter-judge agreement, Swaminathan, Hambleton, and Algina [20] put forward the kappa statistic which is defined by:

$$\kappa = \frac{p - p_c}{1 - p_c} \quad (2)$$

where p is the proportion of examinees classified in the same categories across administrations, and p_c is the agreement expected by chance factors alone.

The concepts of decision consistency and kappa were quickly accepted by the measurement field for use with CRTs, but the restriction of a double administration was impractical. A number of researchers introduced single-administration estimates of decision consistency and kappa, analogous to the corrected split-half reliability that was often the choice of researchers working with NRTs. Huynh [11] put forward his two-parameter ‘bivariate beta-binomial model’. His model relies on the assumption that a group of examinees’ ability scores follow the beta

4 Criterion-Referenced Assessment

distribution with parameters α and β , and the frequency of the observed test scores x follow the beta-binomial (or negative hypergeometric) distribution with parameters α and β . The model is defined by the following:

$$f(x) = \frac{n!}{x!(n-x)!} \frac{B(\alpha+x, \alpha+\beta-x)}{B(\alpha, \beta)} \quad (3)$$

where n is the total number of items in the test, and B is the beta function with parameters α and β , which can be estimated either with the moment method – making use of the first two moments of the observed test scores – or with the **maximum likelihood (ML)** method described in his paper. The probability that an examinee has been consistently classified into a particular category can then be calculated by using the beta-binomial density function. Hanson and Brennan [10] extended Huynh's approach by using the four-parameter beta distribution for true scores.

Subkoviak's method [18] is based on the assumptions that observed scores are independent and distributed binomially, with two parameters – the number of items and the examinee's proportion-correct true score. His procedure estimates the true score for each individual examinee without making any distributional assumptions for true scores. When combined with the binomial or compound binomial error model, the estimated true score will provide a consistency index for each examinee, and averaging this index over all examinees gives the DC index.

Since the previous methods all deal with binary data, Livingston and Lewis [12] came up with a method that can be used with data including either dichotomous, polytomous, or the combination of the two. It involves estimating the distribution of the proportional true scores T_p using strong true score theory [13]. This theory assumes that the proportional true score distribution has the form of a four-parameter beta distribution with density

$$g\left(\frac{T_p}{\alpha, \beta, a, b}\right) = \frac{1}{Beta(\alpha+1, \beta+1)} \times \frac{(T_p - a)^\alpha (b - T_p)^\beta}{(b - a)^{\alpha+\beta+1}} \quad (4)$$

where $Beta$ is the beta function, and the four parameters of the function can be estimated by using the first four moments of the observed scores for

the group of examinees. Then the conditional distribution of scores on an alternate form (given true score) is estimated using a binomial distribution.

All of the previously described methods operate in the framework of **classical test theory (CTT)**. With the popularization of **item response theory (IRT)**, the evaluation of decision consistency and accuracy under IRT has attracted the interest of researchers. For example, Rudner ([16], [17]) introduced his method for evaluating decision accuracy in the framework of IRT.

Rudner [16] proposed a procedure for computing expected classification accuracy for tests consisting of dichotomous items and later extended the method to tests including polytomous items [17]. It should be noted that Rudner referred to θ and $\hat{\theta}$ as 'true score' and 'observed score' respectively in his papers. He pointed out that because of the fact that for any given true score θ , the corresponding observed score $\hat{\theta}$ is expected to be normally distributed, with a mean θ and a standard deviation of $se(\theta)$, the probability of an examinee with a given true score θ of having an observed score in the interval $[a, b]$ on the theta scale is then given by

$$p(a < \hat{\theta} < b|\theta) = \Phi\left[\frac{b-\theta}{se(\theta)}\right] - \Phi\left[\frac{a-\theta}{se(\theta)}\right], \quad (5)$$

where $\Phi(Z)$ is the cumulative normal distribution function. He noted further that multiplying (5) by the expected proportion of examinees whose true score is θ yields the expected proportion of examinees whose true score is expected to be in the interval $[a, b]$, and summing or integrating over all examinees in interval $[c, d]$ gives us the expected proportion of all examinees that have a true score in $[c, d]$ and an observed score in $[a, b]$. If we are willing to make the assumption that the examinees' true scores (θ) are normally distributed, the expected proportions of all examinees that have a true score in the interval $[c, d]$ and an observed score in the interval $[a, b]$ are given by

$$\sum_{\theta=c}^d P(a < \hat{\theta} < b|\theta) f(\theta) = \sum_{\theta=c}^d \left[\Phi\left[\frac{b-\theta}{se(\theta)}\right] - \Phi\left[\frac{a-\theta}{se(\theta)}\right] \right] \Phi\left(\frac{\theta-\mu}{\delta}\right), \quad (6)$$

where $se(\theta)$ is the reciprocal of the square root of the test information function at θ which is the sum of the item information functions in the test, and $f(\theta)$ is the standard normal density function $\Phi(Z)$ [16]. The problem with this method, of course, is that the normality assumption is usually problematic.

Reporting of DC and DA

Table 1 represents a typical example of how DC of performance classifications is being reported. Each of the diagonal elements represents the proportion of examinees in the total sample who were consistently classified into a certain category on both administrations (with the second one being hypothetical), and summing up all the diagonal elements yields the total DC index.

It is a common practice now to report ‘kappa’ in test manuals to provide information on the degree of agreement in performance classifications after correcting for the agreement due to chance.

Table 1 Grade 4 English language arts decision consistency results

Status on parallel form					
Status on form taken	Failing	Basic	Proficient	Advanced	Total
Failing	0.083	0.030	0.000	0.000	0.113
Basic	0.030	0.262	0.077	0.001	0.369
Proficient	0.000	0.077	0.339	0.042	0.458
Advanced	0.000	0.001	0.042	0.018	0.060
Total	1.00	0.369	0.458	0.060	1.00

Note: From Massachusetts Department of Education [14].

Also reported is the ‘conditional error’, which is the measurement error associated with test scores at each of the performance standards. It is helpful because it indicates the size of the measurement error for examinees close to each performance standard.

The values of DA are usually reported in the same way as in Table 1, only that the cross-tabulation is between ‘true score status’ and the ‘test score status’. Of course, it is highly desirable that test manuals also report other evidence to support the score inferences from a CRT, for example, the evidence of content, criterion-related, and construct and consequential validity.

Appropriate Levels of DC and DA

A complete set of approaches for estimating decision consistency and accuracy are contained in Table 2. Note that the value of DA is higher than that of DC because the calculation of DA involves one set of observed scores and one set of true scores which are supposed to be without any measurement error due to improper sampling test questions, flawed test items, problems with the test administration and so on, while the calculation of DC involves two sets of observed scores.

The levels of DC and DA required in practice will depend on the intended uses of the CRT and the number of performance categories. There have not been any established rules to help determine the levels of decision consistency and accuracy needed for different kinds of educational and psychological assessments. In general, the more important the educational decision to be made, the higher consistency and accuracy need to be.

Table 2 Summary of decision consistency and decision accuracy estimation methods

Method	One-admin.	Two-admin.	0-1 Data	0-m Data	CTT-Based	IRT-Based
Hambleton & Novick (1973)		✓	✓		✓	
Swaminathan, Hambleton & Algina (1974)		✓	✓		✓	
Swaminathan, Hambleton & Algina (1975)		✓	✓		✓	
Huynh (1976)	✓		✓		✓	
Livingston & Lewis (1995)	✓		✓	✓	✓	
Rudner (2001)*	✓		✓			✓
Rudner (2004)*	✓		✓	✓		✓

Note: Rudner methods are for decision accuracy estimates only.

References

- [1] Cizek, G. ed. (2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*. Lawrence Erlbaum, Mahwah.
- [2] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.
- [3] Glaser, R. (1963). Instructional technology and the measurement of learning outcomes, *American Psychologist* **18**, 519–521.
- [4] Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process, in *Setting Performance Standards: Concepts, Methods, and Perspectives*, G. Cizek, ed., Lawrence Erlbaum, Mahwah, pp. 89–116.
- [5] Hambleton, R.K., Jaeger, R.M., Plake, B.S. & Mills, C.N. (2000). Setting performance standards on complex performance assessments, *Applied Measurement in Education* **24**(4), 355–366.
- [6] Hambleton, R.K. & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests, *Journal of Educational Measurement* **10**(3), 159–170.
- [7] Hambleton, R.K. & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies, *Applied Measurement in Education* **10**(1), 19–38.
- [8] Hambleton, R.K. & Traub, R. (1973). Analysis of empirical data using two logistic latent trait models, *British Journal of Mathematical and Statistical Psychology* **26**, 195–211.
- [9] Hambleton, R.K. & Zenisky, A. (2003). Advances in criterion-referenced testing methods and practices, in *Handbook of Psychological and Educational Assessment of Children*, 2nd Edition, C.R. Reynolds & R.W. Kamphaus, eds, Guilford, New York, pp. 377–404.
- [10] Hanson, B.A. & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models, *Journal of Educational Measurement* **27**, 345–359.
- [11] Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement* **13**, 253–264.
- [12] Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores, *Journal of Educational Measurement* **32**, 179–197.
- [13] Lord, F.M. (1965). A strong true score theory, with applications, *Psychometrika* **30**, 239–270.
- [14] Massachusetts Department of Education. (2001). *2001 Massachusetts MCAS Technical Manual*, Author, Malden.
- [15] Popham, W.J. & Husek, T.R. (1969). Implications of criterion-referenced measurement, *Journal of Educational Measurement* **6**, 1–9.
- [16] Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees, *Practical Assessment, Research & Evaluation* **7**(14).
- [17] Rudner, L.M. (2004). Expected classification accuracy, in Paper Presented at the Meeting of the National Council on Measurement in Education, San Diego.
- [18] Subkoviak, M.J. (1976). Estimating reliability from a single administration of a criterion-referenced test, *Journal of Educational Measurement* **13**, 265–276.
- [19] Swaminathan, H., Hambleton, R.K. & Algina, J. (1974). Reliability of criterion-referenced tests: a decision-theoretic formulation, *Journal of Educational Measurement* **11**, 263–267.
- [20] Swaminathan, H., Hambleton, R.K. & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests, *Journal of Educational Measurement*, **12**, 87–98.

RONALD K. HAMBLETON AND SHUHONG LI

Critical Region

RAYMOND S. NICKERSON

Volume 1, pp. 440–441

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Critical Region

The basis of null hypothesis testing is a theoretical probability curve representing the theoretical distribution of some statistic (e.g., difference between two means) assuming the independent random selection of two samples from the same population (*see Sampling Distributions*). To decide whether to reject the null hypothesis (the hypothesis that the samples were selected from the same population), one may, following Neyman and Pearson (*see Neyman–Pearson Inference*), establish a criterion value for the statistic and reject the hypothesis if and only if the statistic exceeds that value. All values that exceed that criterion value fall in a region of the decision space that is referred to as the *critical region* (*see Neyman–Pearson Inference*). Conventionally, again following Neyman and Pearson, the decision criterion, α , is selected so as to make the conditional probability of committing a Type I error (rejecting the null hypothesis, given that it is correct) quite small, say .05 or .01.

Suppose, for example, that the question of interest is whether two samples, one with Mean_1 and the other with Mean_2 , can reasonably be assumed to have been randomly drawn from the same population (the null hypothesis), as opposed to having been randomly drawn from different populations, the sample with Mean_2 being drawn from a population

with a larger mean than that of the population from which the sample with Mean_1 was drawn (the alternative hypothesis). A test of the null hypothesis in this case would make use of a theoretical distribution of differences between means of random samples of the appropriate size drawn from the same population. Assuming a strong desire to guard against rejection of the null hypothesis if it is true, the critical region would be defined as a region composed of a small portion of the right tail of the distribution of differences between means, and the null hypothesis would be rejected if and only if the observed differences between the means in hand exceeded the value representing the beginning of the critical region. Critical regions can be defined for tests of hypotheses involving statistics other than means, with similar rationales.

If the alternative to the null hypothesis is the hypothesis of an effect (e.g., a difference between two means) without specification of the direction of the effect (e.g., that Mean_2 is greater than Mean_1), a ‘two-tail’ test of the null hypothesis may be used in which the critical region is composed of two subregions, one on each tail of the distribution. In this case, the area under the curve for each critical subregion would be half the size of the area under the curve in the critical region for a corresponding ‘one-tail’ test.

RAYMOND S. NICKERSON

Cross-classified and Multiple Membership Models

JON RASBASH

Volume 1, pp. 441–450

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cross-classified and Multiple Membership Models

Introduction

Multilevel models also known as variance component models, random effects models and hierarchical linear models, see [5], [7], [14], and [15], have seen rapid growth and development over the last twenty years and are now becoming a standard part of the quantitative social scientist's toolkit (*see Generalized Linear Mixed Models; Linear Multilevel Models*).

Multilevel models provide a flexible regression modeling framework for handling data sampled from clustered population structures (*see Clustered Data*). Examples of clustered population structures are students within classes within schools, patients within hospitals, repeated measurements within individuals or children within families. Ignoring the multilevel structure can lead to incorrect inferences because the standard errors of regression coefficients are incorrect. Also, if the higher-level units are left out of the model we cannot explore questions about the effects of the higher-level units. Most social data have a strong hierarchical structure, which is why multilevel models are becoming so widely used in social science.

The basic multilevel model assumes that the classifications, which determine the multilevel structures, are nested. For example, see Figure 1, which shows a diagram of patients nested within hospitals nested within areas. Often classifications are not nested. Two types of nonnested multilevel models are considered in this chapter, cross-classified models and multiple membership models. These models are also described in [2], [5], [9], [10], and [12]. This entry gives examples of data sets, which have crossed and multiple membership classifications, some diagrammatic tools to help conceptualize these structures and statistical models to describe the structure. We then look at situations where nested, crossed, and multiple membership relationships between classifications can exist in a single population structure and show how the basic diagrams and statistical models are extended to handle this complexity.

Cross-classified Models

Two-way Cross-classifications

Suppose we have data on a large number of high school students and we also know what elementary school they attended and we regard student, high school, and elementary school all as important sources of variation for an educational outcome measure we wish to study. Typically, high schools will draw students from more than one elementary school and elementary schools will send students to more than one high school. The classifications of student, elementary school and high school are not described by a purely nested set of relationships, rather students are contained within a cross-classification of elementary school by high school. Many studies show this simple two-way crossed structure. For example,

- Health: patients contained within a cross-classification of hospital by area of residence
- Survey data: individuals cross-classified by interviewer and area of residence
- Repeated measures cross-classified by the individuals on whom the measurements are made and the set of raters who make the measurements; here different occasions within an individual are assessed by different raters and raters assess many individuals.

Diagrams Representing the Relationship Between Classifications

We find two types of diagrams useful for conveying the relationship between classifications. Firstly, *unit* diagrams where every unit (for example, patient, hospital, and area) appears as a node in the diagram. Lower level units are then connected to higher-level units. A full unit diagram, including all nodes, is prohibitively large. However, a schematic unit diagram conveying the essence of the structure is useful. Figure 2 shows a schematic unit diagram for patients contained within a cross-classification of hospital by area.

The crossing lines in Figure 2 arise because the data structure is cross-classified. When we have many classifications present even the schematic forms of these unit diagrams can be hard to read. In this case an alternative diagrammatic form, which has one node per classification, can be useful. In a *classification* diagram nodes connected by a single arrow represent

2 Cross-classified and Multiple Membership Models

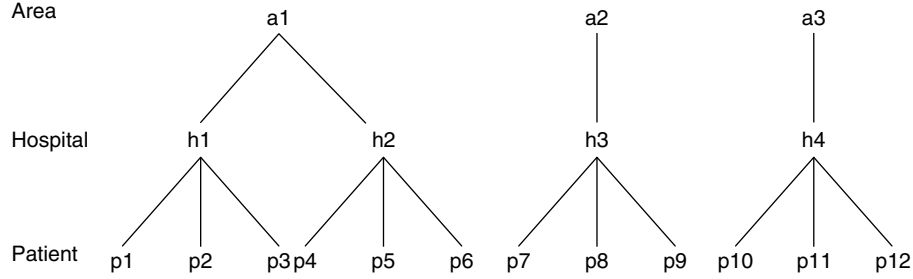


Figure 1 Unit diagram for a nested model

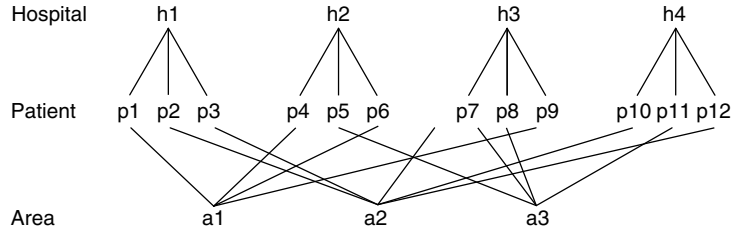


Figure 2 Unit diagram for a cross-classified model

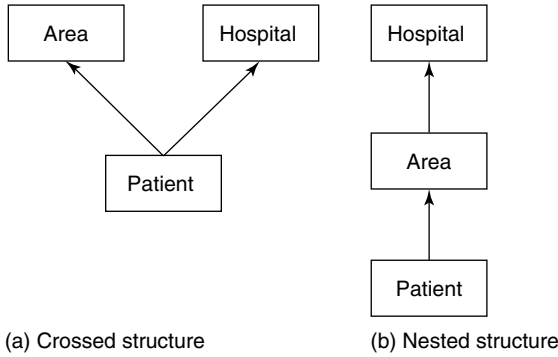


Figure 3 Classification diagrams for crossing and nesting

a nested relationship, nodes connected by a double arrow represent a multiple membership relationship (discussed later), and unconnected nodes represent a cross-classified relationship. Corresponding classification diagrams for the crossed structure in Figure 2 and the completely nested structure in Figure 1 are shown in Figure 3.

Writing the Model Down

In an earlier paper [9], we suggested a notation, which used one subscript per classification, so that

a variance components model for patients within a cross-classification of area by hospital would be written as

$$y_{i(j_1, j_2)} = \beta_0 + u_{j_1} + u_{j_2} + e_{i(j_1, j_2)}, \quad (1)$$

where j_1 indexes hospital, j_2 indexes area and $i(j_1, j_2)$ indexes the i th patient for the cell in the cross-classification defined by hospital j_1 and area j_2 . One problem with this notation is that as more classifications become involved with complex patterns of nesting, crossing, and multiple membership, the subscript formations to describe these patterns become very cumbersome.

An alternative notation, which only involves one subscript no matter how many classifications are present, is given in Browne et al. [2]. In the single subscript notation we write model (1) as:

$$y_i = \beta_0 + u_{\text{area}(i)}^{(2)} + u_{\text{hosp}(i)}^{(3)} + e_i, \quad (2)$$

where i indexes patient and $\text{area}(i)$ and $\text{hosp}(i)$ are functions that return the unit number of the area and hospital that patient i belongs to. For the data structure in Figure 2 the values of $\text{area}(i)$ and $\text{hosp}(i)$ are shown in Table 1. Therefore the model for patient

Table 1 Indexing table for areas by hospitals

i	Area(i)	Hosp(i)
1	1	1
2	2	1
3	2	1
4	1	2
5	3	2
6	1	2
7	2	3
8	3	3
9	1	3
10	2	4
11	3	4
12	2	4

1 would be

$$y_1 = \beta_0 + u_1^{(2)} + u_1^{(3)} + e_1, \quad (3)$$

and for patient 5 would be

$$y_5 = \beta_0 + u_3^{(2)} + u_2^{(3)} + e_5. \quad (4)$$

We use superscripts from 2 upwards to label the random effect corresponding to different classifications reserving the number 1 for the elementary classification. We identify the elementary classification with the letter e . The simplified subscript notation has the advantage that subscripting complexity does not increase as we add more classifications. However, the notation does not describe the patterns of nesting and crossing present. It is therefore useful to use this notation in conjunction with the classification diagrams shown in Figure 3, which display these patterns explicitly.

An Example Analysis for a Two-way Cross-classified Model

Here we look at a two-way cross-classification with students nested within a cross-classification of elementary school by high school. The data comes from Fife in Scotland. The response is the exam score of 3435 children at age 16. There are 19 high schools and 148 elementary schools. We partition variance in the response between student, elementary school, and high school. The classification diagram is shown in Figure 4.

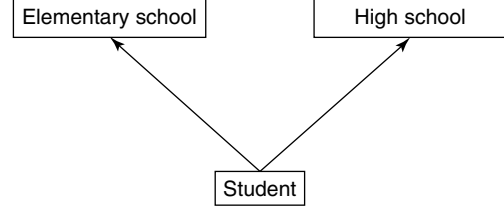


Figure 4 Classification diagram for the Fife educational example

Table 2 Results for the Fife educational data set

Parameter	Description	Estimate(se)
β_0	Mean achievement	5.50 (0.17)
$\sigma_{u^{(2)}}^2$	Elementary school variance	0.35 (0.16)
$\sigma_{u^{(3)}}^2$	High school variance	1.12 (0.20)
σ_e^2	Student variance	8.10 (0.20)

The model is written as:

$$\begin{aligned}
 y_i &= \beta_0 + u_{\text{elem}(i)}^{(2)} + u_{\text{high}(i)}^{(3)} + e_i \\
 u_{\text{elem}(i)}^{(2)} &\sim N(0, \sigma_{u^{(2)}}^2) \\
 u_{\text{high}(i)}^{(3)} &\sim N(0, \sigma_{u^{(3)}}^2) \\
 e_i &\sim N(0, \sigma_e^2).
 \end{aligned} \quad (5)$$

The results in Table 2, show that more of the variation in the achievement at 16 is attributable to high school than elementary school.

Models for More Complex Population Structures

We now consider two examples where the crossed classified structure is more complex than a simple two-way cross-classification.

Social Network Analysis. In social network studies (see **Social Networks**) and family studies we often have observations on how individuals behave or relate to each other. These measurements are often directional. That is we have recorded the amount of behavior from individual A to individual B and also from individual B to individual A. Snijders and Kenny [16] develop a cross-classified multilevel model for handling such data. They use the term *actor* for the individual from whom the behavior originates and the term *partner* for the individual to whom the

4 Cross-classified and Multiple Membership Models

behavior is directed. For a family with two parents and two children, we have 12 directed scores (ds):

$$\begin{aligned} c1 \rightarrow c2, c1 \rightarrow m, c1 \rightarrow f, c2 \rightarrow c1, c2 \rightarrow m, \\ c2 \rightarrow f, m \rightarrow c1, m \rightarrow c2, m \rightarrow f, f \rightarrow c1, \\ f \rightarrow c2, f \rightarrow m \end{aligned}$$

where $c1, c2, f, m$ denote child 1, child 2, father and mother. These directed scores can be classified by actor and by partner. They can also be classified into six dyad groupings:

$$\begin{aligned} (c1 \rightarrow c2, c2 \rightarrow c1), (c1 \rightarrow m, m \rightarrow c1), \\ (c1 \rightarrow f, f \rightarrow c1), (c2 \rightarrow m, m \rightarrow c2), \\ (c2 \rightarrow f, f \rightarrow c2), (m \rightarrow f, f \rightarrow m) \end{aligned}$$

Schematically, the structure is as shown in Figure 5.

Note that the notion of a classification is different from the set of units contained in a classification. For example, the actor and partner classifications are made up of the same set of units (family members). What distinguishes the actors and partners as different classifications is that they have a different set of

connections (or a different mapping) to the level 1 units (directed scores). See [2] for a mathematical definition of classifications as mappings between sets of units.

In this family network data the directed scores are contained within a cross-classification of actors, partners, and dyads and this crossed structure is nested within families. The classification diagram for this structure is shown in Figure 6.

The model can be written as:

$$\begin{aligned} y_i &= (X\beta)_i + u_{\text{actor}(i)}^{(2)} + u_{\text{partner}(i)}^{(3)} + u_{\text{dyad}(i)}^{(4)} \\ &\quad + u_{\text{family}(i)}^{(5)} + e_i \\ u_{\text{actor}(i)}^{(2)} &\sim N(0, \sigma_{u(2)}^2), \quad u_{\text{partner}(i)}^{(3)} \sim N(0, \sigma_{u(3)}^2) \\ u_{\text{dyad}(i)}^{(4)} &\sim N(0, \sigma_{u(4)}^2) \\ u_{\text{family}(i)}^{(5)} &\sim N(0, \sigma_{u(5)}^2), \quad e_i \sim N(0, \sigma_e^2). \end{aligned} \quad (6)$$

These models have not been widely used but they offer great potential for decomposing between and within family dynamics. They can address questions such as:

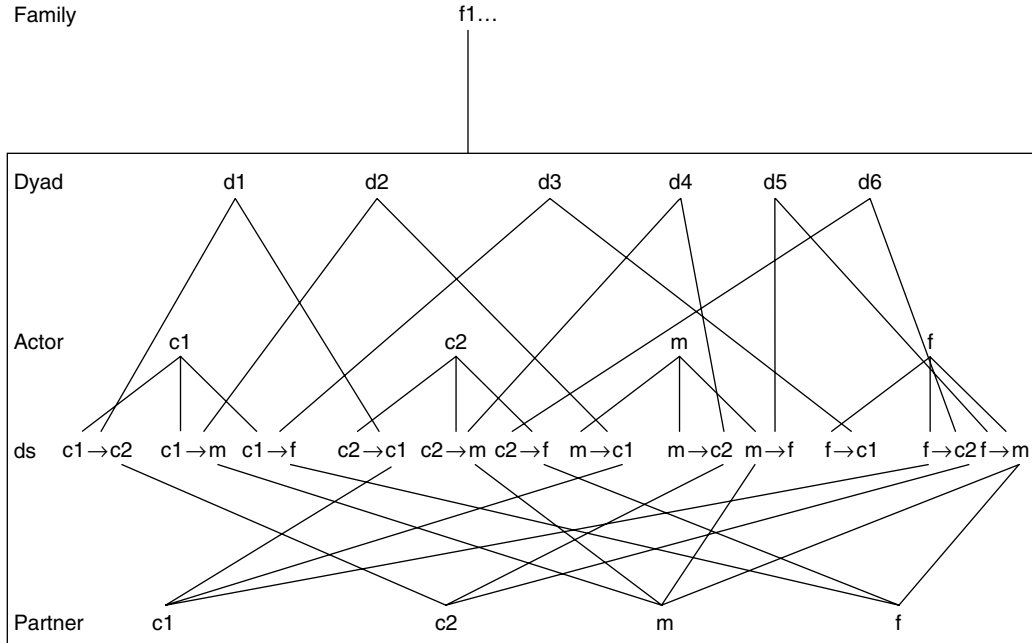


Figure 5 Unit diagram for the family network data

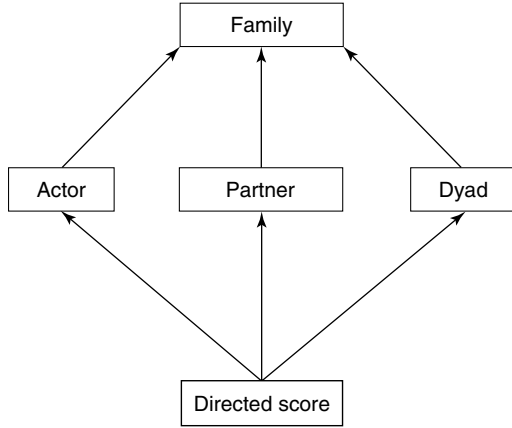


Figure 6 Classification diagram for family network data

how stable are an individual's actor effects across other family members?
 how stable are an individual's partner effects across other family members?
 what are the relative sizes of family, actor, partner, and dyad effects?

The reader is directed to [16] for a more detailed exposition on these models and some example analyses.

A Medical Example. We consider a data set concerning artificial insemination by donor shown in

Figure 7. A detailed description of this data set and the substantive research questions addressed by modeling it within a cross-classified framework are given in Ecochard and Clayton [4]. The data were reanalyzed by Clayton and Rasbash [3] as an example demonstrating the properties of a data augmentation algorithm for estimating cross-classified models.

The data consist of 1901 women who were inseminated by sperm donations from 279 donors. Each donor made multiple donations and there were 1328 donations in all. A single donation is used for multiple inseminations. Each woman received a series of inseminations, one per ovulatory cycle. The data contain 12 100 ovulatory cycles within the 1901 women. The response is a binary variable indicating whether conception occurs in a given cycle.

There are two crossed hierarchies, a hierarchy for donors and a hierarchy for women. The women hierarchy is cycles within women and the donor hierarchy is cycles within donations within donors. Within a series of cycles a women may receive sperm from multiple donors/donations. The model is schematically represented Figure 7.

Here cycles are positioned on the diagram within women, so the topology of the diagram reflects the hierarchy for women. When we connect the male hierarchy to the diagram we see crossing connections between donations and cycles, revealing the crossed structure of the data set. The classification diagram for this structure is shown in Figure 8.

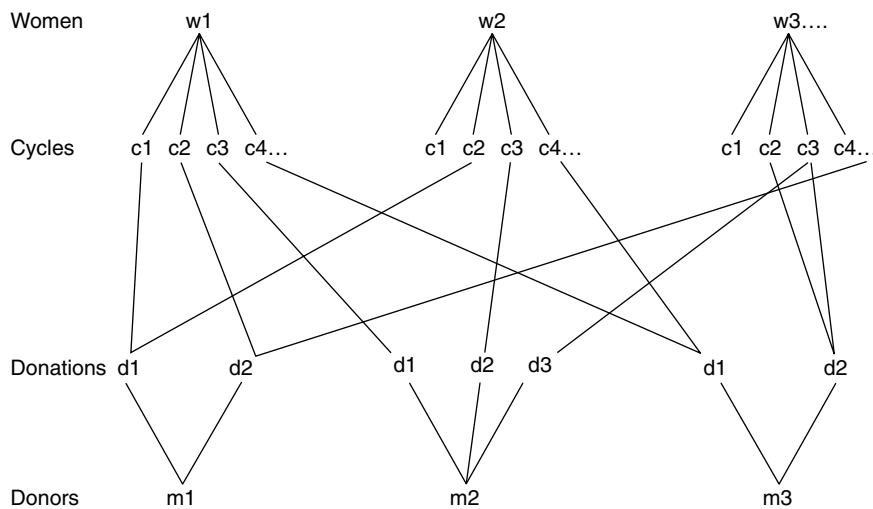


Figure 7 Unit diagram for the artificial insemination example

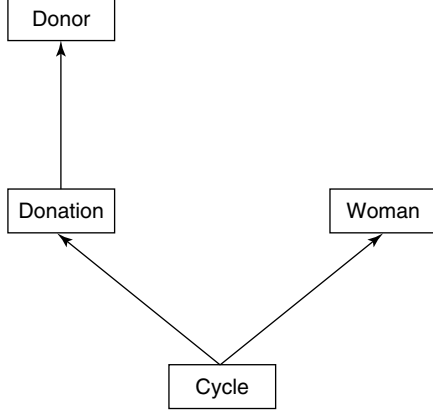


Figure 8 Classification diagram for the artificial insemination example

We can write the model as

$$\begin{aligned}
 y_i &\sim \text{Binomial}(1, \pi_i) \\
 \text{logit}(\pi_i) &= (X\beta)_i + u_{\text{woman}(i)}^{(2)} + u_{\text{donation}(i)}^{(3)} + u_{\text{donor}(i)}^{(4)} \\
 u_{\text{woman}(i)}^{(2)} &\sim N(0, \sigma_{u(2)}^2) \\
 u_{\text{donation}(i)}^{(3)} &\sim N(0, \sigma_{u(3)}^2) \\
 u_{\text{donor}(i)}^{(4)} &\sim N(0, \sigma_{u(4)}^2). \tag{7}
 \end{aligned}$$

The results are shown in Table 3. Azoospermia is a dichotomous variable indicating whether the fecundability of the women is not impaired. The probability of conception is increased with azoospermia and increased sperm motility, count, and quality but decreases in older women or if insemination is attempted too early or late in the monthly cycle. After

Table 3 Results for the artificial insemination data

Parameter	Description	Estimate(se)
β_0	Intercept	-3.92 (0.21)
β_1	Azoospermia	0.21 (0.09)
β_2	Semen quality	0.18 (0.03)
β_3	Women's age >35	-0.29 (0.12)
β_4	Sperm count	0.002 (0.001)
β_5	Sperm motility	0.0002 (0.0001)
β_6	Insemination too early	-0.69 (0.17)
β_7	Insemination too late	-0.27 (0.09)
$\sigma_{u(2)}^2$	Women variance	1.02 (0.15)
$\sigma_{u(3)}^2$	Donation variance	0.36 (0.074)
$\sigma_{u(4)}^2$	Donor variance	0.11 (0.06)

inclusion of covariates, there is considerably more variation in the probability of a successful insemination attributable to the women hierarchy than the donor/donation hierarchy.

Multiple Membership Models

In the models we have fitted so far, we have assumed that lower level units are members of a single unit from each higher-level classification. For example, students are members of a single high school and a single elementary school. Where lower-level units are influenced by more than one higher-level unit from the same classification, we have a multiple membership model. For example, if patients are treated by several nurses, then patients are “multiple members” of nurses. Each of the nurses treating a patient contributes to the patient’s treatment outcome. In this case the treatment outcome for patient i is modeled by a fixed predictor, a weighted sum of the random effects for the nurses that treat patient i and a patient level residual. This model can be written as

$$\begin{aligned}
 y_i &= (X\beta)_i + \sum_{j \in \text{nurse}(i)} w_{i,j}^{(2)} u_j^{(2)} + e_i \\
 u_j^{(2)} &\sim N(0, \sigma_u^2) \\
 e_i &\sim N(0, \sigma_e^2) \\
 \sum_{j \in \text{nurse}(i)} w_{i,j}^{(2)} &= 1 \tag{8}
 \end{aligned}$$

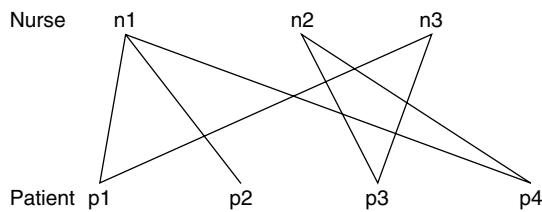
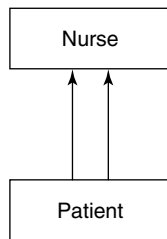
where $\text{nurse}(i)$ is the set of nurses treating patient i and $w_{i,j}^{(2)}$ is the weight given to nurse j for patient i . To clarify this, let’s look at the situation for the first four patients. The weighted membership matrix is shown in Table 4, where patient 1 is treated 0.5 of the time by nurse 1 and 0.5 of the time by nurse 3, patient 2 is seen only by nurse 1 and so on.

Writing out the model for the first four patients gives:

$$\begin{aligned}
 y_1 &= X_1\beta + 0.5u_1^{(2)} + 0.5u_3^{(2)} + e_1 \\
 y_2 &= X_2\beta + 1.0u_1^{(2)} + e_2 \\
 y_3 &= X_3\beta + 0.5u_2^{(2)} + 0.5u_3^{(2)} + e_3 \\
 y_4 &= X_4\beta + 0.5u_1^{(2)} + 0.5u_2^{(2)} + e_4. \tag{9}
 \end{aligned}$$

Table 4 An example weighted membership matrix for patients and nurses

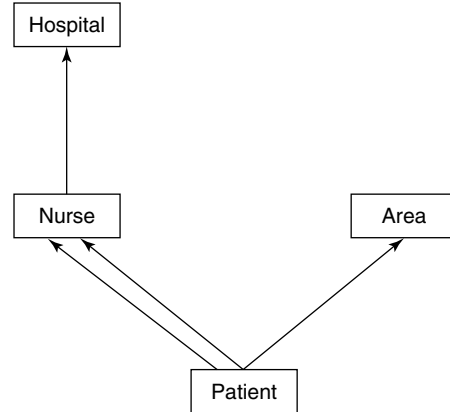
	Nurse 1 ($j = 1$)	Nurse 2 ($j = 2$)	Nurse 3 ($j = 3$)
Patient 1 ($i = 1$)	0.5	0	0.5
Patient 2 ($i = 2$)	1	0	0
Patient 3 ($i = 3$)	0	0.5	0.5
Patient 4 ($i = 4$)	0.5	0.5	0

**Figure 9** Unit diagram for patient nurse classification structure**Figure 10** Classification diagram for patient nurse classification structure

The unit diagram for this structure is shown in Figure 9 and the classification diagram, which denotes multiple membership with a double arrow, is shown in Figure 10.

Combining Nested, Crossed, and Multiple Membership Structures

If we extend the patient/nurse example so that nurses are nested within hospitals and patients' area of residence is crossed with hospitals, we now have a model containing two crossed hierarchies. The first hierarchy is patient within area and the second hierarchy is patients within nurses within hospital, where patients are multiple members of nurses. The

**Figure 11** Classification diagram for nested, crossed multiple membership structure

classification diagram for this structure is shown in Figure 11.

An Example Analysis Combining Nested, Crossed, and Multiple Membership Classifications

The example considered is from veterinary epidemiology. The data has been kindly supplied by Mariann Chriel. It is concerned with causes and sources of variability in outbreaks of salmonella in flocks of chickens in poultry farms in Denmark between 1995 and 1997. The data have a complex structure. There are two main hierarchies in the data. The first is concerned with production. Level 1 units are flocks of chickens and the response is binary, whether there was any instance of salmonella in the flock. Flocks live for a short time, about two months, before they are slaughtered for consumption. Flocks are kept in houses, so in a year a house may have a throughput of five or six flocks. Houses are grouped together in farms. There are 10 127 child flocks in 725 houses in 304 farms.

The second hierarchy is concerned with breeding. There are two hundred parent flocks in Denmark; eggs are taken from parent flocks to four hatcheries. After hatching, the chicks are transported to the farms in the production hierarchy, where they form the production (child) flocks. Any given child flock may draw chicks from up to six parent flocks. Child flocks are therefore multiple members of parent flocks. Chicks from a single parent flock go to many production farms and chicks on a single production

8 Cross-classified and Multiple Membership Models

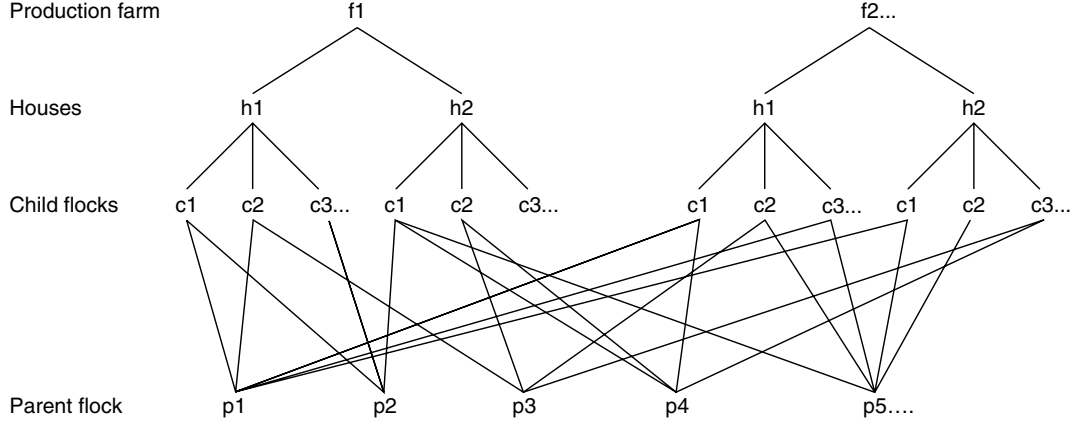


Figure 12 Unit diagram for Danish poultry example

farm come from more than one parent flock. This means the multiple membership breeding hierarchy is cross-classified with the production hierarchy. A unit diagram for the structure is shown in Figure 12 and a classification diagram in Figure 13.

We can write the model as

$$\begin{aligned}
 y_i &\sim \text{Bin}(\pi_i, 1) \\
 \text{logit}(\pi_i) &= (XB)_i + u_{\text{house}(i)}^{(2)} + u_{\text{farm}(i)}^{(3)} \\
 &\quad + \sum_{j \in \text{parent}(i)} w_{i,j}^{(4)} u_j^{(4)} \\
 u_{\text{house}(i)}^{(2)} &\sim N(0, \sigma_{u(2)}^2), \quad u_{\text{farm}(i)}^{(3)} \sim N(0, \sigma_{u(3)}^2)
 \end{aligned}$$

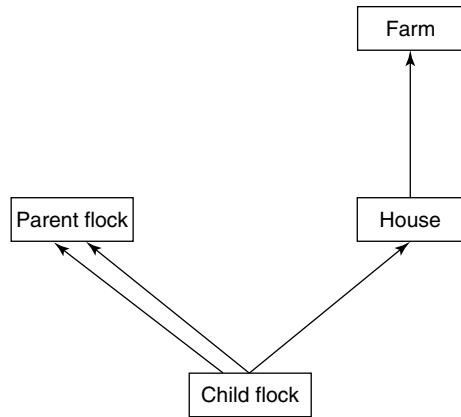


Figure 13 Classification diagram for Danish poultry example

Table 5 Results for Danish poultry data

Parameter	Description	Estimate(se)
β_0	Intercept	-1.86 (0.187)
β_1	1996	-1.04 (0.131)
β_2	1997	-0.89 (0.151)
β_3	Hatchery 2	-1.47 (0.22)
β_4	Hatchery 3	-0.17 (0.21)
β_5	Hatchery 4	-0.92 (0.29)
$\sigma_{u(2)}^2$	House variance	0.19 (0.09)
$\sigma_{u(3)}^2$	Farm variance	0.59 (0.11)
$\sigma_{u(4)}^2$	Parent flock variance	1.02 (0.22)

$$u_j^{(4)} \sim N(0, \sigma_{u(4)}^2). \quad (10)$$

Five covariates were considered in the model: year = 1996, year = 1997, hatchery = 2, hatchery = 3, hatchery = 4. The intercept corresponds to hatchery 1 in 1995. The epidemiological questions of interest revolve around how much of the variation of salmonella incidence is attributable to houses, farms, and parent flocks. The results in Table 5 show there is a large parent flock variance indicating that a parent flock process is having a substantial effect on the variability in the probability of child flock infection. This could be due to genetic variation across parent flocks in resistance to salmonella, or it may be due to differential hygiene standards in parent flocks. There is also a large between farm variance and a relatively small between house within farm variance.

Estimation Algorithms

All the models in this paper were fitted using MCMC estimation [1] (*see Markov Chain Monte Carlo and Bayesian Statistics*) in the *MLwiN* software package [11]. MCMC algorithms for cross-classified and multiple membership models are given by Browne et al. [2]. Two alternative algorithms available in *MLwiN* for cross-classified models are an Iterative Generalised Least Squares (IGLS) algorithm [12] and a data augmentation algorithm [3]. In a forthcoming book chapter Rasbash and Browne [10] give an overview of these algorithms for cross-classified models and compare results for the different estimation procedures. They also give details of an IGLS algorithm for multiple membership models.

Raudenbush [13] gives an empirical Bayes algorithm for two-way cross-classifications. Pan and Thompson [8] give a Gauss–Hermite quadrature algorithm for cross-classified structures and Lee and Nelder [6] give details of an Hierarchical Generalised Linear Model (HGLM), which can handle cross-classified structures.

Summary

Many data sets in the social, behavioral, and medical sciences exhibit crossed and multiple membership structures. This entry sets out some statistical models and diagrammatic tools to help conceptualize and model data with this structure.

References

- [1] Browne, W.J. (2002). *MCMC Estimation in MLwiN*, Institute of Education, University of London, London.
- [2] Browne, W.J., Goldstein, H. & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models, *Statistical Modelling* **1**, 103–124.
- [3] Clayton, D. & Rasbash, J. (1999). Estimation in large crossed random effects models by data augmentation, *Journal of the Royal Statistical Society, Series A* **162**, 425–436.
- [4] Ecochard, R. & Clayton, D. (1998). Multilevel modelling of conception in artificial insemination by donor, *Statistics in Medicine* **17**, 1137–1156.
- [5] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold, London.
- [6] Lee, Y. & Nelder, J. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random effects models and structured dispersion, *Biometrika* **88**, 987–1006.
- [7] Longford, N.T. (1995). *Multilevel Statistical Models*, Oxford University Press, New York.
- [8] Pan, J.X. & Thompson, R. (2000). Generalised linear mixed models: an improved estimating procedure, in *COMPSTAT: Proceedings in Computational Statistics*, J.G. Bethlehem & P.G.M. van der Heijden, eds, Physica-Verlag, Heidelberg, pp. 373–378.
- [9] Rasbash, J. & Browne, W.J. (2001). Modelling non-hierarchical multilevel models, in *Multilevel Modelling of Health Statistics*, A.H. Leyland & H. Goldstein, eds, Wiley, London.
- [10] Rasbash, J. & Browne, W.J. (2005). Non-hierarchical multilevel models, To appear in *Handbook of Quantitative Multilevel Analysis*, J. De Leeuw & I.G.G. Kreft, eds, Kluwer.
- [11] Rasbash, J., Browne, W., Healy, M., Cameron, B. & Charlton, C. (2000). *The MLwiN Software Package Version 1.10*, Institute of Education, London.
- [12] Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model, *Journal of Educational and Behavioural Statistics* **19**, 337–350.
- [13] Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research, *Journal of Educational Statistics* **18**, 321–350.
- [14] Raudenbush, S.W. & Bryk, A.S. (2001). *Hierarchical Linear Modelling*, 2nd Edition, Sage, Newbury Park.
- [15] Snijders, T. & Bosker, R. (1998). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage, London.
- [16] Snijders, T.A.B. & Kenny, D.A. (1999). The social relations model for family data: a multilevel approach, *Personal Relationships* **6**, 471–486.

JON RASBASH

Cross-lagged Panel Design

DAVID A. KENNY

Volume 1, pp. 450–451

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cross-lagged Panel Design

A cross-lagged panel design is simple to describe. Two variables, X and Y , are measured at two times, 1 and 2, resulting in four measures, X_1 , Y_1 , X_2 , and Y_2 . With these four measures, there are six possible relations among them – two synchronous or cross-sectional relations (*see* **Cross-sectional Design**) (between X_1 and Y_1 and between X_2 and Y_2), two stability relations (between X_1 and X_2 and between Y_1 and Y_2), and two cross-lagged relations (between X_1 and Y_2 and between Y_1 and X_2). As is typical in most considerations of the design, X and Y are treated as continuous variables.

As an example of cross-lagged panel design, Bentler and Speckart [1] examined the variables attitudes to alcohol and alcohol behavior (i.e., consumption) measured at two times. Their key research question was whether attitudes determined behavior or whether behavior determined attitudes.

Despite the simplicity in defining the design, there is considerable debate and difficulty in the statistical analysis of the data from such a design. By far, the most common analysis strategy is to employ **multiple linear regression**. Each measure at the second time or wave is predicted by the set of time-one measures. The measure X_2 is predicted by both X_1 and Y_1 , and Y_2 is predicted by X_1 and Y_1 . The result is four regression coefficients, two of which are stabilities and two of which are cross-lagged. Many assumptions are required before these cross-lagged relations can be validly interpreted as causal effects. Among them are (1) no measurement error in the time-one measures, (2) no unmeasured third variables that cause both X and Y , and (3) the correct specification of the causal lag. One strategy for handling the problem of measurement error is to employ a **latent variable** analysis (*see* **Structural Equation Modeling: Overview**) ([1]). However, it is much more difficult to know whether the assumptions of no unmeasured variables and correct specification of the causal lag have been met.

An alternative and rarely used approach was developed from Campbell's [2] cross-lagged panel correlation, or CLPC. For this approach, we begin with the assumption that the association between X and Y is due to some unmeasured latent variable.

That is, it is assumed that there are no direct causal effects between the two variables. As discussed by Kenny [3], given assumptions about certain parameters, such a model implies that the two cross-lagged relations should be equal. Also infrequently used are methods based on computing the relations using the average of each variable, $(X_1 + X_2)/2$, and their difference $X_2 - X_1$ ([4]).

Almost always there are additional variables in the model. More than two variables are being measured at both times. Other variables are demographic or control variables. Sometimes there are also intermediary variables (variables that measure what happens between time points). One additional complication arises in that typically some of the units that are measured at time one are not measured at time two, and less frequently some of those measured at time two are not measured at time one. Finally, it may not be clear that X and Y are 'measured at the same time'.

Several analysts (e.g., Singer & Willett [5]) have argued that two time points are insufficient to make strong causal claims about the relations between the variables. Thus, while a cross-lagged panel design might provide some information about causal ordering, it may be less than optimal.

References

- [1] Bentler, P.M. & Speckart, G. (1979). Models of attitude-behavior relations, *Psychological Review* **86**, 452–464.
- [2] Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching, in *Handbook of Research on Teaching*, N.L. Gage, ed., Rand-McNally, Chicago, pp. 171–246.
- [3] Kenny, D.A. (1975). Cross-lagged panel correlation: a test for spuriousness, *Psychological Bulletin* **82**, 887–903.
- [4] Kessler, R.C. & Greenberg, D.F. (1981). *Linear Path Analysis: Models of Quantitative Change*, Academic Press, New York.
- [5] Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, New York.

(*See also* **Multitrait–Multimethod Analyses; Structural Equation Modeling: Checking Substantive Plausibility**)

DAVID A. KENNY

Crossover Design

MICHAEL G. KENWARD

Volume 1, pp. 451–452

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Crossover Design

In experimental settings where the effects of treatments are reversible, and comparatively short-lived, the potential exists for increasing the precision of treatment effects through the use of within-subject comparisons: if repeated measurements (*see Repeated Measures Analysis of Variance*) on a subject are highly correlated, then differences among these will be much more precise than the differences among the same outcomes measured on different subjects. A *crossover design* exploits this by using repeated measurements from each subject under different treatments, and the gain in terms of precision compared with a **completely randomized design** increases with increasing within-subject **correlation**.

With two treatments (A and B, say) we might consider giving all the subjects A followed after a suitable gap by B. Such a design is flawed, however, because we cannot separate the effect of time or order of treatment from the treatment comparison itself. We can circumvent this problem by randomly allocating half the subjects to receive the treatments in the order AB and half in the order BA. When the treatment difference is calculated from such a design, any effects associated with time (termed *period effects*) cancel out. This simple design is known as the two-period two-treatment, or 2×2 , crossover. In an example described in [2, Section 2.11], this design was used in the investigation of the relationship between plasma estradiol levels in women and visuospatial ability. The ‘treatments’ in this case were defined by periods of relatively low- and high-estradiol levels in women undergoing *in-vitro* fertilization, with the women randomized to start the trial at either the high level or the low level.

Crossover designs have potential disadvantages along with the benefits. After the first period, subjects in different sequences have experienced different treatment regimes and are therefore not comparable in the sense of a completely randomized design. The justification for inferences about treatment effects cannot be based purely on **randomization** arguments. Additional assumptions are required, principally, that previous treatment allocation does not affect subsequent treatment comparisons. If this is *not* true, there is said to be a **carryover effect**. There are

many ways in which carryover effects can in principle occur. For example, in an experiment with drugs as treatments, it may be that these have not cleared a subject’s system completely at the end of a period, and so continue to affect the response in the following period. Often *washout* periods, intervals without treatment, are used between treatment periods to minimize the likelihood of this. Sometimes, for practical reasons, washout periods may not be possible, and other forms of carryover may be unaffected by washout periods. Another source of potential problem is the existence of treatment-by-period interaction. Should there be nonnegligible changes over time in response, learning effects being one common cause of this, it is possible in some settings that treatment comparisons will differ according to the period in which they are made. In the two-period two-treatment design, there is little that can be done about such problems, and the conclusions from such experiments rest on the assumption that carryover effects and treatment-by-period interaction are small compared to the direct treatment comparison.

There are many other crossover designs that may be used and each has particular practical and theoretical advantages and disadvantages. In particular, many allow the separation of direct treatment, carryover, and treatment-by-period effects. Although this may seem advantageous, such separation is always still firmly based on modeling assumptions that cannot be fully checked from the data under analysis. Also, correction for many extra terms can lead to great inefficiency. Hence, such statistical manipulations and corrections should not be seen as a blanket solution to the issues surrounding the assumptions that need to be made when analyzing crossover data.

In general, a crossover design may have more than two treatments, more than two periods, and more than two sequences. For a full description of the many crossover designs that have been proposed, see [2, Chapters 3 and 4]. One common feature of all crossover designs is the requirement that there be at least as many sequences as treatments, for otherwise it is not possible to separate treatment and period effects.

Practical restrictions permitting, it is generally desirable to balance as far as possible, the three main components of the design, that is, sequences, periods, and treatments, and this implies that the

number of periods will equal the number of treatments. Examples of such designs that are in perfect balance are so-called balanced **Latin square designs** in which each treatment occurs equally often in each period and each sequence [2, Section 4.2]. These are the most efficient designs possible, provided no adjustment need be made for other effects such as carryover. Additional forms of balance can be imposed to maintain reasonable efficiency when adjusting for simple forms of carryover effect, the so-called *Williams squares* are the commonest example of these.

The analysis of continuous data from crossover trials usually follows conventional factorial ANOVA type procedures (*see Factorial Designs*), incorporating fixed subject effects (*see Fixed and Random Effects*) [1; 2, Chapter 5]. This maintains simplicity but does confine the analysis to within-subject information only. In efficient designs, most or all information on treatment effects is within-subject, so it is rarely sensible to deviate from this approach. However, it is sometimes necessary to use designs that are not efficient, for example, when the number of treatments exceeds the number of periods that can be used, and for these it is worth considering the recovery of *between-subject* or inter-block information. This is accomplished by using an analysis with

random as opposed to fixed subject effects (*see Fixed and Random Effects*). In psychological research, it is not unusual to have crossover designs with very many periods in which treatments are repeated. In the analysis of data from such trials, there are potentially many period parameters. For the purposes of efficiency, it may be worth considering replacing the categorical period component by a smooth curve to represent changes over time; a polynomial or non-parametric smooth function might be used for this [2, Section 5.8]. For analyzing nonnormal, particularly, categorical data, appropriate methods for nonnormal clustered or repeated measurements data (*see Generalized Linear Mixed Models*) can be adapted to the crossover setting [2, Chapter 6].

References

- [1] Cotton, J.W. (1998). *Analyzing Within-subject Experiments*, Lawrence Erlbaum Associates, New Jersey.
- [2] Jones, B. & Kenward, M.G. (2003). *The Design and Analysis of Cross-over Trials*, 2nd Edition, Chapman & Hall/CRC, London.

MICHAEL G. KENWARD

Cross-sectional Design

PATRICIA L. BUSK

Volume 1, pp. 453–454

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cross-sectional Design

Cross-sectional research involves collecting data at the same time from groups of individuals who are of different ages or at different stages of development. Suppose a researcher is interested in studying self-concept of girls as they mature. It would take a long time to collect data on a group of first-grade girls following them through graduation from high school, measuring their self-concept every few years. Instead, the researcher could select samples of girls in the first, fourth, seventh, tenth, and twelfth grades and measure the self-concept of the girls in each of the samples. Differences between average self-concept scores at the different grade levels could be interpreted as reflecting developmental changes in female self-concept.

Longitudinal research (*see* **Longitudinal Data Analysis**) is a study of changes of a sample over time. There are several disadvantages to conducting a longitudinal research study. The study takes a long time to complete, and subjects are lost over the time period. As an alternative, researchers have turned to cross-sectional studies where large groups are studied at one time. Instead of studying one group over time, cohorts are investigated at the same time. The cohorts usually vary on the basis of age, stages of development, or different points in a temporal sequence, for example, college graduates, those with an M.A. degree, and those with a Ph.D. or Ed.D. degrees (*see* **Cohort Studies**). The advantage is that a study can be conducted in a relatively short period of time and with little loss of subjects. The major disadvantage is that any differences between the cohorts based on the variable or variables under study are confounded with cohort differences, which may be due to age differences or other extraneous factors unrelated to those being investigated. The confounding that can occur in the cross-sectional design may be attributed to personological or environmental variables or to the subjects' history and is called a *cohort effect*. The larger the difference in the variable for the cohort groups, then the greater potential for cohort effects, that is, if age is the variable that is the basis for the cohort, then the greater the age differences, the more likely subject history may contribute to cohort differences than the variable under investigation. In fact, if the age groups include an extremely large

range, then generational differences may be the confounding factor. If individuals are matched across cohorts on relevant variables, then the effects of subject history may be diminished. Another problem with cross-sectional designs is selecting samples that truly represent the individuals at the levels being investigated.

Many developmental studies employ a cross-sectional design. If a researcher is interested in developmental trends or changes in moral beliefs, a sample of students may be taken from different grade levels, for example, fourth, sixth, eighth, tenth, and twelfth. The students would be administered the moral beliefs instrument at about the same time in all of the grades and analyze for trends. Given that the data are collected within the same time period, individuals will not be lost over time owing to moving out of the area, death, and so on. **Factorial designs** can be employed within each of the cohorts. Using the same moral beliefs example, if the researcher were interested in gender and ethnic differences, then those variables would be assessed and analyzed. Main effects for age, gender, and development may be tested along with the interactions between these variables.

Nesselroade and Baltes [2, p. 265] have argued that 'cross-sectional designs confound interindividual growth and consequently are not adequate for the study of developmental processes'. Their premise is that repeated measurements on the same individual are essential for the assessment of individual growth and change (*see* **Growth Curve Modeling**). Purkey [3], however, used a number of cross-sectional studies to establish a persistent relationship between measures of self-concept and academic achievement. Causal inferences cannot be made from Purkey's cross-sectional studies, however, only from longitudinal research, which points to the limitations of cross-sectional studies.

As an example of a repeated cross-sectional design, consider the High School and Beyond (HS&B) Study [1] that began in 1980 and sampled 1647 high-school students from all regions of the United States. The HS&B survey included two cohorts: the 1980 senior class and the 1980 sophomore class. Both cohorts were surveyed every 2 years through 1986, and the 1980 sophomore class was also surveyed again in 1992. The purpose of the study was to follow students as they began to take on adult roles. A cross-sectional study

2 Cross-sectional Design

would include seniors, individuals with 2-, 4-, and 6-years graduation in the same geographic regions, same gender composition, and same socioeconomic status.

Additional information on repeated cross-sectional studies can be found under **cohort sequential designs** under accelerated longitudinal designs.

References

- [1] High School and Beyond information can be retrieved from HYPERLINK "[http://nces.ed.gov/sur-](http://nces.ed.gov/surveys/hsb/)

veys/hsb/” <http://nces.ed.gov/surveys/hsb/>

- [2] Nesselroade, J.R. & Baltes, P.B., (eds) (1979). *Longitudinal Research in the Study of Behavior and Development*, Academic Press, New York.
- [3] Purkey, W.W. (1970). *Self-concept and School Achievement*, Prentice-Hall, Engelwood Cliffs.

PATRICIA L. BUSK

Cross-validation

WERNER STUETZLE

Volume 1, pp. 454–457

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cross-validation

In a generic regression problem (see **Regression Models**), we want to model the dependence of a *response variable* Y on a collection $\mathbf{X} = (X_1, \dots, X_m)$ of *predictor variables*, based on a *training sample* $\mathcal{T} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$. The kinds of models as well as the goodness-of-fit criteria depend on the nature of the response variable Y . If Y is a continuous variable, we might fit a linear model by least squares, whereas for binary Y , we might use **logistic regression** or **discriminant analysis**. In the following, we will focus on the case of continuous response, but the ideas are very general.

Goals of Regression Analysis

There are (at least) two goals of regression analysis. The first one is to ‘understand how Y depends on $\mathbf{X}$.’ By this, we mean making statements like ‘ Y tends to increase as X_1 increases’ or ‘ X_5 seems to have little influence on Y .’ If we used a **linear model**

$$Y \sim b_0 + b_1 X_1 + \dots + b_m X_m, \quad (1)$$

we would base such statements on the estimated regression coefficients $\hat{b}_1, \dots, \hat{b}_m$. If the predictor variables are correlated – the usual situation for observational data – such interpretations are not at all innocuous; Chapters 12–13 of Mosteller and Tukey [2] contain an excellent discussion of the problems and pitfalls.

The second goal, and the one we will focus on, is prediction: generate a *prediction rule* $\phi(\mathbf{x}; \mathcal{T})$ that predicts the value of the response Y from the values $\mathbf{x} = (x_1, \dots, x_m)$ of the predictor variables. In the case of a linear model, an obvious choice is

$$\phi(\mathbf{x}; \mathcal{T}) = \hat{b}_0 + \hat{b}_1 x_1 + \dots + \hat{b}_m x_m, \quad (2)$$

where the regression coefficients are estimated from the training sample $\mathcal{T}$.

In the social and behavioral sciences, the dominant goal has traditionally been to understand how the response depends on the predictors. Even if understanding is the primary goal, it might still be prudent, however, to evaluate the predictive performance of a model. Low predictive power can indicate a lack of

understanding of the underlying mechanisms and may call any conclusions into question.

Measuring the Performance of a Prediction Rule

Once we have generated a prediction rule from our training sample, we want to assess its performance. We first have to choose a *loss function* $L(y, \hat{y})$ that specifies the damage done if the rule predicts $\hat{y}$, but the true response is y . A standard choice for continuous response is squared error loss: $L(y, \hat{y}) = (y - \hat{y})^2$. We then measure performance by the *risk* of the rule, the expected loss when applying the prediction rule to new *test observations* assumed to be randomly drawn from the same population as the training observations. For squared error loss, the case we will consider from now on, the risk is

$$R(\phi) = E((Y - \phi(\mathbf{X}; \mathcal{T}))^2), \quad (3)$$

where the expectation is taken over the population distribution of $(\mathbf{X}, Y)$. The question is how to estimate this risk.

Example

We now describe a simple example which we will use to illustrate risk estimation. There are $m = 50$ predictor variables that are independent and uniformly distributed on the unit interval $[0, 1]$. The response Y is a linear function of the predictors, plus Gaussian noise ϵ with mean 0 and variance σ^2 :

$$Y = b_1 X_1 + \dots + b_m X_m + \epsilon. \quad (4)$$

Each of the true regression coefficients $b_1, \dots, b_m$ is zero with probability 0.8, and an observation of a standard Gaussian with probability 0.2. Therefore, only about 10 of the true regression coefficients will be nonvanishing. The noise variance σ^2 is chosen to be the same as the ‘signal variance’:

$$\sigma^2 = \mathbf{V}(b_1 X_1 + \dots + b_m X_m). \quad (5)$$

The Resubstitution Estimate of Risk

The simplest and most obvious approach to risk estimation is to see how well the prediction rule does

2 Cross-validation

for the observations in the training sample. This leads to the *resubstitution estimate* of risk

$$\hat{R}_{\text{resub}}(\phi) = \frac{1}{n} \sum (y_i - \phi(\mathbf{x}_i; \mathcal{T}))^2, \quad (6)$$

which is simply the average squared residual for the training observations. The problem with the resubstitution estimate is that it tends to underestimate the risk, often by a substantial margin. Intuitively, the reason for this optimism is that, after all, the model was chosen to fit the training data well.

To illustrate this effect, we generated a training sample $\mathcal{T}$ of size $n = 100$ from the model described in the previous section, estimated regression coefficients by least squares, and constructed the prediction rule (2). The resubstitution estimate of risk is $\hat{R}_{\text{resub}}(\phi) = 0.64$. Because we know the population distribution of $(\mathbf{X}, Y)$, we can compute the true risk of the rule: We generate a very large ($N = 10\,000$) *test set* of new observations from the model and evaluate the average loss incurred when predicting those 10 000 responses from the corresponding predictor vectors. The true risk turns out to be $R(\phi) = 3.22$; the resubstitution estimate underestimates the risk by a factor of 5!

Of course, this result might be a statistical fluke – maybe we just got a bad training sample? To answer this question, we randomly generated 50 training samples of size $n = 100$, computed the true risk and the resubstitution estimate for each of them, and averaged over training samples. The average resubstitution estimate was 0.84, while the average true risk was 3.48; the result was not a fluke.

The Test Set Estimate of Risk

If we had a large data set at our disposal – a situation not uncommon in this age of automatic, computerized data collection – we could choose not to use all the data for making up our prediction rule. Instead, we could use half the data as the training set $\mathcal{T}$ and compute the average loss when making predictions for the test set. The average loss for the test set is an unbiased estimate for the risk; it is not systematically high or systematically low.

Often, however, we do not have an abundance of data, and using some of them just for estimating the risk of the prediction rule seems wasteful, given that we could have obtained a better rule by using all the data for training. This is where cross-validation comes in.

The Cross-validation Estimate of Risk

The basic idea of cross-validation, first suggested by Stone [3] is to use each observation in both roles, as a training observation and as a test observation. Cross-validation is best described in algorithmic form:

Randomly divide the training sample $\mathcal{T}$ into k subsets $\mathcal{T}_1, \dots, \mathcal{T}_k$ of roughly equal size (choice of k is discussed below). Let $\mathcal{T}^{-i}$ be the training set with the i -th subset removed.

For $i = 1 \dots k$ {

- Generate a prediction rule $\phi(\mathbf{x}, \mathcal{T}^{-i})$ from the training observations not in the i -th subset.
- Compute the total loss L^i when using this rule on the i -th subset:

$$L^i = \sum_{j \in \mathcal{T}^i} (y_j - \phi(\mathbf{x}_j, \mathcal{T}^{-i}))^2 \quad (7)$$

}

Compute the cross-validation estimate of risk

$$\hat{R}_{\text{cv}}(\phi) = \frac{1}{n} \sum_{i=1}^k L^i. \quad (8)$$

In our example, the cross-validation estimate of risk is $\hat{R}_{\text{cv}} = 2.87$, compared to the true risk $R = 3.22$, and the resubstitution estimate $\hat{R}_{\text{resub}} = 0.64$. So the cross-validation estimate is much closer to the true risk than the resubstitution estimate. It still underestimates the risk, but this is a statistical fluke. If we average over 50 training samples, the average cross-validation estimate is 3.98. The cross-validation estimate of risk is not optimistic, because the observations that are predicted are not used in generating the prediction rule. In fact, cross-validation tends to be somewhat pessimistic, partly because it estimates the performance of a prediction rule generated from a training sample of size roughly $n(1 - 1/k)$.

A question which we have not yet addressed is the choice of k . In some situations, such as linear least squares, leave-one-out cross validation, corresponding to $k = n$, has been popular, because it can be done in a computationally efficient way. In general, though, the work load increases with k because we have to generate k prediction rules instead of one. Theoretical analysis of cross-validation has proven to be surprisingly difficult, but a general consensus, based mostly

on empirical evidence, suggests that $k = 5$ or 10 is reasonable (see [1, Chapter 7.10]).

Using Cross-validation for Model Selection

In a situation like the one in our example, where we have many predictor variables and a small training sample, we can often decrease the prediction error by reducing model complexity. A well-known approach to reducing model complexity in the context of linear least squares (see **Least Squares Estimation**) is *stepwise regression*: Find the predictor variable that, by itself, best explains Y ; find the predictor variable that best explains Y when used together with the variable found in step 1; find the variable that best explains Y when used together with the variables found in steps 1 and 2; and so on. The critical question is when to stop adding more variables.

The dotted curve in Figure 1 shows the resubstitution estimate of risk – the average squared residual – plotted against the number of predictor variables in the model. It is not helpful in choosing a good model size.

The dashed curve shows the cross-validation estimate of risk. It is minimized for a rule using six predictor variables, suggesting that we should end the process of adding variables after six steps.

The solid curve shows the true risk. It exhibits the same pattern as the cross-validation estimate and is

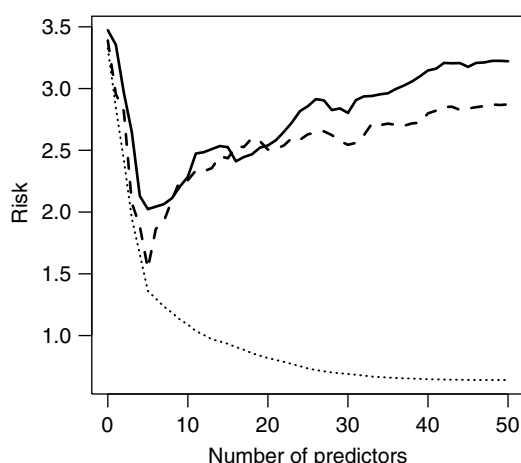


Figure 1 True risk (solid), resubstitution estimate (dotted) and cross-validation estimate (dashed) for a single training sample as a function of the number of predictor variables

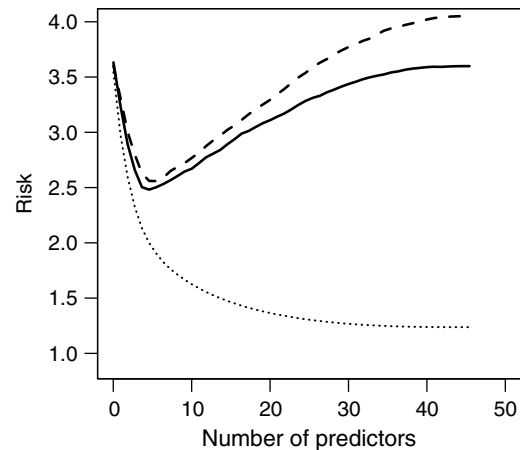


Figure 2 Average of true risk (solid), resubstitution estimate (dotted) and cross-validation estimate (dashed) over 50 training samples as a function of the number of predictor variables

also minimized for a model with six variables. This is not surprising, given that only about 10 of the true regression coefficients are nonvanishing, and some of the remaining ones are small. Adding more predictor variables basically just models noise in the training sample; complex models typically do not generalize well.

Figure 2 shows the corresponding plot, but the curves are obtained by averaging over 50 training samples. Note that the cross-validation estimate of risk tracks the true risk well, especially for the lower ranges of model complexity, which are the practically important ones.

Alternatives to Cross-validation

Many alternatives to cross-validation have been suggested. There is **Akaike's Information Criterion (AIC)**, the **Bayesian Information Criterion (BIC)**, **Minimum Description Length (MDL)**, and so on; see [1, Chapter 7] for a survey and references. These criteria all consist of two components: a measure of predictive performance for the training data, and a penalty for 'model complexity.' In principle, this makes sense – more complex models are more prone to modeling the noise in the training data, which makes the resubstitution estimate of risk more optimistic. However, it is often not obvious how to

measure model complexity, and the degree of optimism depends not only on the set of models under consideration but also on the thoroughness of the search process, the amount of data dredging. There are also risk estimates based on the **bootstrap**, which are similar in spirit to cross-validation, in that they substitute calculations on a computer for mathematical derivation of mostly asymptotic results. Overall, though, cross-validation is appealing because it is intuitive, easy to understand, and trivial to implement.

References

- [1] Hastie, T., Tibshirani, R. & Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
- [2] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression*, Addison Wesley, Cambridge.
- [3] Stone, M. (1974). Cross-validated choice and assessment of statistical predictions, *Journal of the Royal Statistical Society* **36**, 111–147.

WERNER STUETZLE

Cultural Transmission

SCOTT L. HERSHBERGER

Volume 1, pp. 457–458

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Cultural Transmission

Cultural transmission facilitates the transfer, via teaching and imitation, of the knowledge, values, and other factors of a civilization that influence behavior. Cultural transmission creates novel evolutionary processes through the existence of socially transmitted traditions that are not directly attributable to genetic factors and immediate environmental contingencies; instead, social learning *interacts* with environmental contingencies to produce behavior (*see* **Gene-Environment Interaction**). As a nongenetic form of evolution by natural selection, if the behavior that arises from cultural forces is not adaptive, it will not be transmitted across generations. Thus, the relationship between culture and behavior created by cultural transmission is analogous to the relationship between **genotype** and phenotype caused by genetic transmission.

Cultural transmission is not simply another term to describe how the 'environment' influences behavior. Behavior caused by cultural transmission must itself be modifiable in later generations by cultural influences contemporary to that generation. For example, the local climate and the kinds of food items available are part of the population's environment but are not affected by evolutionary changes in the population. Individuals form the environment relevant to cultural transmission.

Cultural transmission also relies specifically on social learning and not on other forms of learning. Learning mechanisms that influence a specific individual's behavior, and not the population, are not relevant to cultural transmission. For instance, given a criterion of reinforcement, such as a sense of pain or a taste for rewards, behavior will change. Behaviors acquired through reinforcement schedules are lost with the individual's death; culturally acquired behaviors are transmitted from generation to generation, and like genes, they are evolving properties of the population.

Different types of cultural transmission can be defined. Vertical transmission refers to transmission from parents to offspring, and can be uniparental or biparental. Vertical transmission does not require genetic relatedness – adoptive parents can transfer knowledge to their adopted children. Horizontal transmission refers to transmission between any two individuals, related or not, who are of the

same generation, such as siblings and age peers. Oblique transmission occurs from a member of a given generation to a later generation, but does not include parent–offspring transmission (which is specific to vertical transmission). Examples of oblique sources of cultural information are members of society in general, family members other than parents, and teachers.

Mathematical models of cultural transmission can be placed in two categories: (a) evolutionary models (e.g., [1, 2]), and (b) behavior genetic models (e.g., [5]). Evolutionary models emphasize similarities among individuals within a population, while behavior genetic models emphasize differences among individuals within a population.

Cultural transmission can create correlations between genotypes and environments. **Genotype–environment correlation** refers to a correlation between the genetic and environmental influences affecting a trait. As in any other correlation, where certain values of one variable tend to occur in concert with certain values of another variable, a significant genotype–environment correlation represents the nonrandom distribution of the values of one variable (genotype) across the values of another variable (environment).

Passive genotype–environment correlation [8] arises from vertical cultural transmission. That is, because parents and their offspring are on the average 50% genetically related, both genes and cultural influences are transmitted to offspring from a common source (parents), inducing a correlation between genotypes and the environment. Genotype–environment correlation can also arise through horizontal cultural transmission. In sibling effects (*see* **Sibling Interaction Effects**) models [3, 4], sibling pairs are specified as sources of horizontal transmission. When siblings compete or cooperate, genotype–environment correlation occurs if the genes underlying a trait in one sibling also influence the environment of the cosibling. Reactive genotype–environment correlation [7] can result from oblique cultural transmission. In this type of correlation, society reacts to the *level* of an individual's genetically influenced behavior by providing cultural information that it might not provide to other individuals who do not show the same level of behavior. For example, children who show unusual talent for the violin will more likely be instructed by masters of the violin

2 Cultural Transmission

than those children who do not demonstrate such talent. Active genotype–environment correlation [6] can also arise from oblique cultural transmission: Individuals select certain cultural artifacts on the basis of genetically influenced proclivities. The child with unusual talent for the violin will choose to master that musical instrument over the piano.

References

- [1] Boyd, R. & Richerson, P.J. (1985). *Culture and the Evolutionary Process*, University of Chicago Press, Chicago.
- [2] Cavalli-Sforza, L.L. & Feldman, M.W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton University Press, Princeton.
- [3] Eaves, L.J. (1976a). A model for sibling effects in man, *Heredity* **36**, 205–214.
- [4] Eaves, L.J. (1976b). The effect of cultural transmission on continuous variation, *Heredity* **37**, 41–57.
- [5] Eaves, L.J., Eysenck, H.J. & Martin, N.G. (1989). *Genes, Culture, and Personality*, Academic Press, London.
- [6] Hershberger, S.L. (2003). Latent variable models of genotype-environment correlation, *Structural Equation Modeling* **10**(3), 423–434.
- [7] Plomin, R., DeFries, J.C. & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior, *Psychological Bulletin* **84**, 309–322.
- [8] Scarr, S. & McCartney, K. (1983). How people make their own environments: a theory of genotype → environment effects, *Child Development* **54**, 424–435.

SCOTT L. HERSHBERGER

Data Mining

DAVID J. HAND

Volume 1, pp. 461–465

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Data Mining

Data mining is the technology of discovering structures and patterns in large data sets. From this definition, it will be immediately apparent that the discipline has substantial overlap with other data analytic disciplines, especially statistics, machine learning, and **pattern recognition**. However, while there is substantial overlap, each of these disciplines has its own emphasis. Data mining, in particular, is distinguished from these other disciplines by the (large) size of the data sets, often by the poor quality of the data, and by the breadth of the kind of structures sought. Each of these aspects is explored in more detail below.

Data mining exercises and tools can be conveniently divided into two types. The first is *model building* and the second is *pattern detection*. In model building, one seeks to summarize (large parts of) the data in a convenient form. Model building is a familiar exercise, especially to statisticians, and many of the issues in data mining modeling are the same as those in statistics. Differences do arise, however, due to the sizes of the data sets involved, and some of these are discussed below. In pattern detection on the other hand, one seeks anomalies in a data set. Although there are subdisciplines of statistics devoted to this aim (**outlier detection**-, *scan statistics*, and, especially in the behavioral sciences, **configural frequency analysis**-, are examples), in general, far less attention has been devoted to it in the past, at least in statistics. There are probably sound historical reasons for this: many pattern detection problems depend on having a large data set available, and this is a relatively recent development. One might characterize these two rather complementary aspects of data mining as being, on the one hand, like coal mining or oil recovery, where one is extracting and refining large masses of material (this is the modeling aspect), and, on the other, as being like diamond or gold mining, where one is seeking the occasional gem or nugget from within a mass of dross (this is the pattern detection aspect).

Much of the original motivation for data mining has come from the commercial sector: from the promise that the large databases now collected contain, hidden within them, potential discoveries, which could give an organization a market edge. This would include, for example, a superior partition

of customers into groups for sales purposes, better models for predicting customer behavior, and identification of potential fraudulent transactions. However, the tools are now widely used in scientific problems where sometimes truly vast datasets are collected. Scientific areas, which are using data mining methods, include bioinformatics (genomics, proteomics, microarray data (*see* **Microarrays**), astronomy (with giant star catalogues), and medicine (with things like image analysis, scanning technologies, and direct electronic feeds from intensive care patients). The application of such tools in the behavioral sciences is relatively new, but is growing.

Model Building

There are various possible aims in model building. In particular, one can distinguish between summary and prediction. In the former, one merely seeks to summarize data in a manner which is convenient and accurate, while, in the latter, one has in mind the subsequent use of the model for, for example, predicting possible future outcomes or the values of some variables from others. Furthermore, there are different types of models according to whether it is to be based on an underlying theory or is purely empirical. In all cases, however, the objective is to produce a description, which summarizes the data in a convenient and relatively simple form. It will be apparent from this that models typically decompose the entire data set, or at least large chunks of it, into parts: we may decompose a **time series** into trend and seasonal components using signal processing, a distribution of data points into groups using cluster analysis (*see* **Cluster Analysis: Overview; Hierarchical Clustering**), a related set of variables into cliques in a **graphical model**, and so on.

So far, all of the above applies equally to data mining, statistics, and other data analytic technologies. However, it is clear that fitting a model to a data set describing customer behavior consisting of a billion transaction records, or a data set describing microarray or image data with tens of thousands of variables, will pose very different problems from building one for 10 measurements on each of a hundred subjects. One of the most obvious differences is that significance and hypothesis tests become irrelevant with very large data sets. In conventional statistical modeling (in the frequentist approach, at least),

one evaluates the quality of a model by looking at the probability of obtaining the observed data (or more extreme data) from the putative model. Unfortunately, when very large data sets are involved, this probability will almost always be vanishingly small: even a slight structural difference will translate into many data points, and hence be associated with a very small probability and so high significance. For this reason, measures of model quality other than the conventional statistical ones are widely used. Choice between models is then typically based on the relative size of these measures.

Apart from analytic issues such as the above, there are also more mundane difficulties associated with fitting models to large data sets. With a billion data points, even a simple **scatterplot** can easily reduce to a solid black rectangle: a contour plot is a more useful summary. For such reasons, and also because many of the data sets encountered in data mining involve large numbers of variables, dynamic interactive graphical tools are quite important in certain data mining applications.

Pattern Detection

Whereas models are generally large-scale decompositions of the data, splitting the data into parts, patterns are typically small-scale aspects: in pattern detection, we are interested only in particular small localities of the data, and the rest of the data are irrelevant. Just as there are several aims in model building, so there are several aims in pattern detection.

In *pattern matching*, we are told the structure of the pattern we are seeking, and the aim is to find occurrences of it in the data. For example, we may look for occurrences of particular behavioral sequences when studying group dynamics, or purchasing patterns when studying shopping activities. A classic example of the latter has been given the name *bread dominoes*, though it describes a much more general phenomenon. The name derives from shoppers who normally purchase a particular kind of bread. If their preferred kind is not present, they tend to purchase the most similar kind, and, if that is not present, the next most similar kind, and so on, in a sequence similar to a row of dominoes falling one after the other. Pattern matching has been the focus of considerable research in several areas, especially

genomics, espionage, and text processing. The latter is especially important for the web, where search engines are based on these ideas.

In *supervised pattern detection*, we are told the values of some variables, y , and the aim is to find the values of other variables, x , which are likely to describe data points with the specified y values. For example, we might want to find early childhood characteristics, x , which are more likely to have a high value of a variable y , measuring predisposition to depression.

In *pattern discovery*, the aim is both to characterize and to locate unusual features in the data. We have already mentioned outliers as an example of this: we need to define what we mean by an outlier, as well as test each point to identify those, which are outliers. More generally, we will want to identify those regions of the data space, which are associated with an anomalously high local density of data points. For example, in an EEG trace, we may notice repeated episodes, separated by substantial time intervals, in which *'a brief interval of low voltage fast or desynchronized activity [is] followed by a rhythmic (8–12 Hz) synchronized high amplitude discharge ... The frequency then begins to slow and spikes to clump in clusters, sometimes separated by slow waves ... Finally, the record is dominated by low amplitude delta activity'*. Any signal of a similar length can be encoded so that it corresponds to some point in the data space, and similar signals will correspond to similar points. In particular, this means that whenever a signal similar to that above occurs, it will produce a data point in a specific region of the data space: we will have an anomalously high local density of data points corresponding to such patterns in the trace. In fact, the above quotation is from Toone [8]; it describes the pattern of electrical activity recorded in an EEG trace during a grand mal epileptic seizure.

Pattern discovery is generally a more demanding problem than pattern matching or supervised pattern detection. All three of these have to contend with a potentially massive search space, but, in pattern discovery, one may also be considering a very large number of possible patterns. For this reason, most of the research in the area has focused on developing effective search algorithms. Typically, these use some measure of 'interestingness' of the potential patterns and seek local structures, which maximize this. An important example arises in *association analysis*. This describes the search, in a multivariate categorical data

space, for anomalously high cell frequencies. Often the results are couched in the form of *rules*: ‘If A and B occur, then C is likely to occur’. A classic special case of this is *market basket analysis*, the analysis of supermarket purchasing patterns. Thus, one might discover that people who buy sun-dried tomatoes also have a higher than usual chance of buying balsamic vinegar. Note that, latent behind this procedure, is the idea that, when one discovers such a pattern, one can use it to manipulate the future purchases of customers. This, of course, does not necessarily follow: merely because purchases of sun-dried tomatoes and balsamic vinegar are correlated does not mean that increasing the purchases of one will increase the purchases of the other. On the other hand, sometimes such patterns can be taken advantage of. Hand and Blunt [3] describe how the discovery of patterns revealing surprising local structures in petrol purchases led to the use of free offers to induce people to spend more.

Exhaustive search is completely infeasible, so various forms of constrained search have been developed. A fundamental example is the *a priori algorithm*. This is based on the observation that if the pattern AB occurs too few times to be of interest, then there is no point in counting occurrences of patterns which include AB. They must necessarily occur even less often. In fact, this conceals subtleties: it may be that the frequency which is sufficient to be of interest should vary according to the length of the proposed pattern.

Search for data configurations is one aspect of pattern discovery. The other is inference: is the pattern real or could it have occurred by chance? Since there will be many potential patterns thrown up, the analyst faces a particularly vicious version of the *multiplicity* problem: if each pattern is tested at the 5% level, then a great many ‘false’ patterns (i.e. not reflecting real underlying structure in the distribution) will be flagged; if one controls the overall level of flagging any ‘false’ pattern as real at the 5% level, then the test for each pattern will be very weak. Various strategies have been suggested for tackling this, including the use of false discovery rate now being promoted in the medical statistics literature, and the use of likelihood as a measure of evidence favoring each pattern, rather than formal testing. Empirical Bayesian ideas (see **Bayesian Statistics**) are also used to borrow strength from the large number of similar potential patterns.

The technology of scan statistics has much to offer in this area, although most of its applications to date have been to relatively simple (e.g. mainly one-dimensional) situations.

In general, because the search space and the space of potential patterns are so vast, there will be a tendency to throw up many possible patterns, most of which will be already known or simply uninteresting. One particular study [1] found 20 000 rules and concluded ‘the rules that came out at the top were things that were obvious’.

Data Quality

The issue of data quality is important for all data analytic technologies, but perhaps it is especially so for data mining. One reason is that data mining is typically secondary data analysis. That is, the data will normally have been collected for some purpose other than data mining, and it may not be ideal for mining. For example, details of credit card purchases are collected so that people can be properly billed, and not so that, later, an analyst can pore through these records seeking patterns (or, more immediately, so that an automatic system can examine them for signs of fraud).

All data are potentially subject to distortion, errors, and missing values, and this is probably especially true for large data sets. Various kinds of errors can occur, and a complete taxonomy is probably impossible, though it has been attempted [7]. Important types include the following:

- *Missing data*. Entire records may be missing; for example, if people have a different propensity to enter into a study, so that the study sample is not representative of the overall population. Or individual fields may be missing, so distorting models; for example, in studies of depression, people may be less likely to attend interview sessions when they are having a severe episode (see **Missing Data; Dropouts in Longitudinal Studies; Methods of Analysis**).
- *Measurement error*. This includes ceiling and floor effects, where the ranges of possible scores are artificially truncated.
- *Deliberate distortion*. This, of course, can be a particular problem in the behavioral sciences, perhaps especially when studying sensitive topics such as sexual practices or alcohol

consumption. Sometimes, in such situations, sophisticated data capture methods, such as **randomized response** [9], can be used to tackle it.

Clearly, distorted or corrupted data can lead to difficulties when fitting models. In such cases, the optimal approach is to include a model for the data distortion process. For example, Heckman [6] models both the sample selection process and the target regression relationship. This is difficult enough, but, for pattern discovery, the situation is even more difficult. The objective of pattern discovery is the detection of anomalous structures in the data, and data errors and distortion are likely to introduce anomalies. Indeed, experience suggests that, in addition to patterns being ‘false’, and, in addition to them being uninteresting, obvious, or well known, most of the remainder are due to data distortion of some kind. Hand et al. [4] give several examples.

Traditional approaches to handling data distortion in large data sets derive from work in survey analysis, and include tools such as automatic edit and imputation. These methods essentially try to eliminate the anomalies before fitting a model, and it is not obvious that they are appropriate or relevant when the aim is pattern detection in data mining. In such cases, they are likely to smooth away the very features for which one is searching. Intensive human involvement seems inescapable.

The Process of Data Mining

Data mining, like statistical data analysis, is a cyclical process. For model fitting, one successively fits a model and examines the quality of the fit using various diagnostics, then refines the model in the light of the fit, and so on. For pattern detection, one typically locates possible patterns, and then searches for others in the light of those that have been found. In both cases, it is not a question of ‘mining the data’ and being finished. In a sense, one can never finish mining a data set: there is no limit to the possible questions that could be asked.

Massive data sets are now collected almost routinely. In part, this is a consequence of automatic electronic data capture technologies, and, in part, it is a consequence of massive electronic storage facilities. Moreover, the number of such massive data sets

is increasing dramatically. It is very clear that there exists the potential for great discoveries in these data sets, but it is equally clear that making those discoveries poses great technical problems. Data mining, as a discipline, may suffer from a backlash, as it becomes apparent that the potential will not be achieved as easily as the media hype accompanying its advent may have led us to believe. However, there is no doubt that such a technology will be needed more and more in our increasingly data-dependent world. Data mining will not go away.

General books on data mining, which describe the tools in computational or mathematical detail, include [2], [5], and [10].

References

- [1] Brin, S., Motwani, R., Ullma, J.D. & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, ACM Press, Tucson, pp. 255–264.
- [2] Giudici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*, Wiley, Chichester.
- [3] Hand, D.J. & Blunt, G. (2001). Prospecting for gems in credit card data, *IMA Journal of Management Mathematics* **12**, 173–200.
- [4] Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M. (2000). Data mining for fun and profit, *Statistical Science* **15**, 111–131.
- [5] Hand, D.J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*, MIT Press, Cambridge.
- [6] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models, *Annals of Economic and Social Measurement* **5**, 475–492.
- [7] Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K. & Lee, D. (2003). A taxonomy of dirty data, *Data Mining and Knowledge Discovery* **7**, 81–99.
- [8] Toone, B. (1984). The electroencephalogram, in *The Scientific Principles of Psychopathology*, P. McGuffin, M.F. Shanks & R.J. Hodgson, eds, Grune and Stratton, London, pp. 36–55.
- [9] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.
- [10] Witten, I.H. & Franke, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco.

DAVID J. HAND

de Finetti, Bruno

SANDY LOVIE

Volume 1, pp. 465–466

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

de Finetti, Bruno

Born: June 13, 1906, in Innsbruck, Austria.

Died: July 20, 1985, in Rome, Italy.

De Finetti's early education was as a mathematician, graduating from the University of Milan in 1927, and then working as government statistician at the National Institute of Statistics in Rome until 1946. His final move was to the University of Rome first in the Department of Economics, then attached to the School of Probability. A considerable amount of his time was spent undertaking actuarial work, not too surprising given his employment. This added both a sense of time and finality to his views on the nature of probability and risk, and a strong conviction that probability was only meaningful in a human context. According to Lindley [1], de Finetti was fond of saying that 'Probability does not exist', by which he meant that probability cannot be thought of independently of the (human) observer. Jan von Plato's recent history of modern probability theory [3] even compares de Finetti to the eighteenth-century Empiricist philosopher David Hume in that probability is likened to sense data and hence *entirely* dependent on the judgment of the observer. This radical vision was formulated by de Finetti around the late 1920s and early 1930s, appearing in a series of papers listed in von Plato [3, p. 298–300]. The central concept through which this subjective approach was worked out is *exchangeability*, that is, the notion that there exist probabilistically 'equivalent sequences of events' (to quote Savage, [2]), which allowed a somewhat weaker form of statistical independence to be used in a subjective context, although most historians note that the notion of exchangeability was not

discovered by de Finetti[3]. De Finetti also pioneered the use of what some people subsequently referred to as the *scoring rule method* of extracting subjective probabilities. Thus, the generation of personal probabilities and probability distributions was relocated by de Finetti to a gambling setting. This orientation was combined by de Finetti with his notion of exchangeability to argue that people who held coherent views on subjective probability would eventually converge on the same subjective probability values if they were faced with, say, separate sequences from a binomial source with a common probability of success. Although de Finetti came from a rather different mathematical tradition and culture than the Anglo-American one, as Lindley points out [1], his influence on key figures in the Bayesian statistics movement (see **Bayesian Statistics**), including Lindsey himself and the American statistician **Leonard Savage**, was decisive. For instance, on page 4 of the introduction to the latter's best-known book [2], one finds the extent to which de Finetti's notion of subjective probability had become the foundation for Savages's own work. More generally, the post-WWII rise of Bayesian statistics gave an enormous boost to de Finetti's international reputation.

References

- [1] Lindley, D.V. (1997). Bruno de finetti, in *Leading Personalities in Statistical Science*, N.L. Johnson & S. Kotz, eds, Wiley, New York.
- [2] Savage, L.J. (1954, 1972). *The Foundations of Statistics*, Wiley, New York.
- [3] von Plato, J. (1994). *Creating Modern Probability*, Cambridge University Press, Cambridge.

SANDY LOVIE

de Moivre, Abraham

DAVID C. HOWELL

Volume 1, pp. 466–466

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

de Moivre, Abraham

Born: May 26, 1667, in Vitry-le-François, France.
Died: November 27, 1754, in London, England.

Abraham de Moivre was born in France in 1667 into a protestant family at a time when being a protestant in France was difficult. He attended a variety of schools, but in 1685, when Louis XIV revoked the Edict of Nantes, he was imprisoned for a time and then emigrated to England, where he remained for the rest of his life.

In England, de Moivre made his living as a tutor of mathematics. Soon after his arrival, he met Halley and Newton, and was heavily influenced by both of them. He soon mastered Newton's *Principia*, which was no mean feat, and Halley arranged for his first paper in mathematics to be read before the Royal Society in 1695. In 1697, he was made a fellow of the Society.

De Moivre's most important work was in the theory of probability and analytic geometry. He published his *The Doctrine of Chance: A method of calculating the probabilities of events in play* in 1718, and it was revised and republished several times during his life. He published *Miscellanea Analytica* in

1730, and in that work, he developed Stirling's formula to derive the normal distribution (*see Catalogue of Probability Density Functions*) as an approximation to the binomial. (The original discovery was de Moivre's, though he later credited Stirling, who provided a later simplification.) The normal distribution is one of the critical foundations of modern statistical theory, and its use as an approximation to the binomial is still an important contribution. In developing this approximation, de Moivre had to make use of the standard deviation, though he did not provide it with a name.

Despite de Moivre's eminence as a mathematician, he was never able to obtain a university chair, primarily owing to that fact that he was a foreigner. Even his friends Halley and Newton were not sufficiently influential in supporting his application. As a result, he remained throughout his life a private tutor of mathematics, and died in poverty in 1754.

Further Reading

Moivre, Abraham de. (2004). Available at http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/De_Moivre.html

DAVID C. HOWELL

Decision Making Strategies

ROB RANYARD

Volume 1, pp. 466–471

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Decision Making Strategies

Representations of Decision Problems

Decision making is a form of problem solving in which the need to choose among alternative courses of action is the focus of the problem and not merely part of the journey towards its solution. For example, if a student had offers of places on two or more university courses, a decision problem has to be solved. The goal-state of such a problem is usually ill-defined and often a matter of personal preference. At least four types of preferential decision have been defined and studied: multiattribute, risky, sequential and dynamic decisions. Table 1 illustrates a typical multiattribute decision dilemma for a student choosing a course of study: should she choose

Business Studies, which is more relevant to her future career, or Psychology, which she expects to be more interesting? The central issue in multiattribute decision making is how to resolve such conflicts. Table 1 represents this dilemma as a decision matrix of *alternatives* (different courses) varying on several *attributes* (relevance to future career, interest, and so on).

A second important type of decision involves risk or uncertainty. Figure 1 illustrates some of the risky aspects of our student's decision problem represented as a decision tree. If she aspired to graduate level employment after her course, then the safe option would be Business Studies if it led to a job at that level for certain. In comparison, the risky option would be Psychology if that resulted in a only a one-third chance of her target job as a graduate psychologist, but a two-thirds chance of a nongraduate level post which would be a particularly bad outcome for her. Should she take the risk, or opt for

Table 1 Conflicts in course decisions represented as a multiattribute choice

Course	Course attribute			Utility
	Career relevance	Interest of content	Quality of teaching	
Social Anthropology	3	6	8	102
Psychology	5	7	5	105
Business Studies	8	4	4	108
Attribute importance	8	5	6	

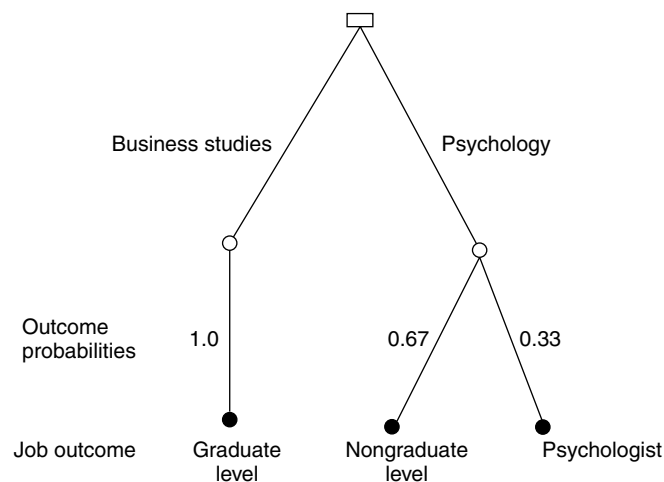


Figure 1 Risk or uncertainty represented as a decision tree

the safer course in terms of future employment? That is the typical risky decision dilemma. Decision making where outcome probabilities are known, such as in a game of roulette, are conventionally described as decision under risk. However, decision making where outcome probabilities are unknown, and the stated probabilities represent subjective estimates, as in our example, are known as decision under uncertainty [6, 20].

The other types of decision making (sequential and dynamic) involve situations where people's decisions are dependent on previous decisions they have made. This entry will only consider strategies for multiattribute and risky decision making, for reasons of space. For the same reason, the entry is specifically concerned with evaluative choice rather than evaluative judgment. There is a subtle difference between the two. Evaluative *judgments*, as opposed to decisions, are made when a person is asked to evaluate decision alternatives separately rather than choosing one of them. In consumer contexts, for example, people might be asked to say what they might be willing to pay for an item. A substantial body of research has demonstrated that people apply different strategies in evaluative judgment as opposed to evaluative decision tasks [13, 14, 22]. For example, there is evidence that an anchoring-and-adjustment heuristic is often applied to judgments of the prices of consumer items, but not to choices between them [3].

Two Theoretical Frameworks: Structural Models and Process Descriptions

Two major traditions that have shaped contemporary psychological theories of decision making are utility theory, which aims to predict decision behavior, and the information processing approach, which models underlying cognitive processes. It is assumed in the former that the goal of the rational decision maker is to maximize utility or expected utility. Specific variants of utility theory have been proposed as structural models to describe and predict decision behavior (see entry on utility theory). Such models are not strictly decision strategies, since they do not necessarily correspond to the mental processes underlying decisions. Nevertheless, they can be interpreted as strategies, as discussed below.

Cognitive, or information processing approaches to the study of decision making began to emerge

in the early sixties. Herbert Simon [12] argued that human information processing capacities limit the rationality of decision making in important ways. He believed that utility maximization is beyond cognitive limitations in all but the simplest environments. His *bounded rationality* model assumed that people construct simplified mental representations of decision problems, use simple decision strategies (or heuristics), and often have the goal of making a satisfactory, rather than an optimal, decision. Since Simon's seminal work, much research has adopted his bounded rationality perspective to develop process models describing mental representations, decision strategies and goals. Payne, Bettman, and Johnson [10, p. 9] defined a decision strategy as 'a sequence of mental and effector (actions on the environment) operations used to transform an initial state of knowledge into a final state of knowledge where the decision maker views the particular decision problem as solved'. Process models can be categorized as either single or multistage. Single-stage models describe decision strategies in terms of single sequences of elementary information processing operators (eips) that lead to a decision, whereas multistage models describe sequences of eips that form interacting components of more complex strategies [2].

A Taxonomy of Decision Strategies to Resolve Multiattribute Conflicts

Compensatory Strategies: Additive Utility

Decision strategies can be classified as either compensatory or noncompensatory. The former involve trade-offs between the advantages and disadvantages of the various choice alternatives available, whereas noncompensatory strategies do not. Utility models can be recast within an information processing framework as compensatory decision strategies appropriate for certain contexts. Consider the student's dilemma over which course to choose, illustrated in Table 1. The example assumes a short list of three courses: Social Anthropology, Psychology, and Business Studies. It assumes also that the three most important attributes to the student are the interest of the subject, the quality of teaching, and the relevance of the course to her future career. She has been gathering information about these courses and her assessments are shown in the decision matrix on nine-point scales, where 1 = very poor and 9 = excellent.

In addition, the decision will depend on the relative importance of these attributes to the student: Which is more important, career-relevance or interest? Are they both equally important? The answer will be different for different people.

Utility theorists have proposed a rational strategy to resolve the student's dilemma which takes attribute importance into account. Multiattribute utility theory (MAUT) assumes that each attribute has an *importance weight*, represented in the bottom row of the decision matrix in the table on a nine-point scale (9 = most important). According to MAUT, the utility of each alternative is calculated by adding the utilities of each aspect multiplied by its importance weight [15]. This weighting mechanism ensures that a less important attribute makes a smaller contribution to the overall utility of the alternative than a more important one. For example, utility (psychology) = $(8 \times 5) + (5 \times 7) + (6 \times 5) = 105$ units. Here, the interest value has made a smaller contribution than career relevance, even though the course was rated high in interest value. However, for this student, career relevance is more important, so it contributes more to overall utility. This is a compensatory strategy, since positively evaluated aspects, such as the career-relevance of the Business Studies course, compensate for its less attractive attributes.

Noncompensatory Strategies: Satisficing

Even with our simplified example in Table 1, nine items of information plus three attribute importance weights must be considered. Most real-life decisions involve many more attributes and alternatives and we have limited time and cognitive capacity to cope with all the information we receive. A range of decision strategies we might use in different contexts has been identified. Beach and Mitchel [1] argued that the cognitive effort required to execute a strategy is one of the main determinants of whether it is selected. Building on this, Payne, Bettman, and Johnson [10] developed the more precise *effort-accuracy* framework, which assumes that people select strategies adaptively, weighing the accuracy of a strategy against the cognitive effort it would involve in a given decision context.

One of the earliest noncompensatory strategies to be proposed was Simon's *satisficing principle*:

choose the first alternative, that is at least satisfactory on all important attributes. This saves time and effort because only part of the available information is processed and all that is required is a simple comparison of each aspect with an acceptability criterion. In our example, (Table 1), one could work through each alternative, row by row, perhaps starting at the top. Suppose a rating of 5 was considered satisfactory. Working from left to right across the matrix, the first alternative, Social Anthropology, could be rejected on career-relevance and the other two aspects would not need to be examined. The second alternative would be selected because it passes satisficing tests on all three attributes and consequently, the third alternative would not be considered at all. In this example, satisficing leads to a decision after processing less than half the available information and the conflict across attributes is avoided rather than resolved. Although such a choice process may be good enough in some contexts, satisficing often fails to produce the best decision: the gain in reduced effort is at the expense of a loss of 'accuracy', as Payne et al. argue.

Direction of Processing: Attribute-based or Alternative-based

Amos Tversky [18] proposed two alternative decision strategies to explain how evaluative decisions can violate one of the basic principles of rationality, *transitivity of preference*. Intransitive preference occurs when someone prefers A to B, B to C, but C to A. (A preference for a Psychology over a Business Studies course, Business Studies over Anthropology, but Anthropology over Psychology, would clearly need to be sorted out). Tversky observed similar intransitive cycles of preference with risky decisions, and explained them in terms of attribute-based processing. This is in contrast to alternative-based processing strategies, such as additive utility and satisficing, in which each alternative is processed one at a time. In an attribute-based strategy, alternatives are initially compared on some or all of their important attributes. In Tversky's, *additive difference* strategy, alternatives are compared systematically, two at a time, on each attribute. The differences between them are added together in a compensatory manner, to arrive at an evaluation as to which is the best. For example, in Table 1, the differences between Business Studies and Psychology on career relevance and interest would cancel

out, and the small difference in quality of teaching would tip the balance in favor of Psychology. Tversky showed that intransitive preference could occur if the evaluations of differences were nonlinear ([18], see also [11]).

A noncompensatory attribute-based strategy could also account for intransitive preferences. In a *lexicographic* strategy, the decision maker orders the attributes by importance and chooses the best alternative on the most important attribute (for our hypothetical student, Business Studies is chosen because career-relevance is the most important attribute). If there is a tie on the first attribute, the second attribute is processed in the same way. The process stops as soon as a clear favorite on an attribute is identified. This requires less cognitive effort because usually information on several attributes is not processed (it is noncompensatory since trade-offs are not involved). Tversky argued that if preferences on any attribute form a semi-order, involving intransitive indifference, as opposed to a full rank-order, then intransitivity could result. This is because small differences on a more important attribute may be ignored, with the choice being based on less important attributes, whereas large differences would produce decisions based on more important attributes. He termed this the *lexicographic semi-order* strategy.

Other important attribute-based strategies include dominance testing [9] and elimination by aspects [19]. With respect to the former, one alternative is said to dominate another if it is at least as good as the other on all attributes and strictly better on at least one of them. In such a case there is no conflict to resolve and the rational decision strategy is obvious: choose the dominant alternative. From a cognitive and a rational perspective it would seem sensible to test for dominance initially, before engaging in deeper processing involving substantially more cognitive effort. There is a lot of empirical evidence that people do this. Turning to the elimination-by-aspects strategy, this is similar to satisficing, except that evaluation is attribute-based. Starting with the most important, alternatives are evaluated on each attribute in turn. Initially, those not passing a satisficing test on the most important attribute are eliminated. This process is repeated with the remaining alternatives on the next most important attribute, and so on. Ideally, the process stops when only one alternative remains. Unfortunately, as is the case for most

noncompensatory strategies, the process can fail to resolve a decision problem, either because all alternatives are eliminated, or more than one remain.

Two Stage and Multistage Strategies

It has been found that with complex choices, involving perhaps dozens of alternatives, several decision strategies are often used in combination. For example, all alternatives might initially be screened using the satisficing principle to produce a short-list. Additional screening could apply dominance testing to remove all dominated alternatives, thereby shortening the short-list still further. Following this, a variety of strategies could be employed to evaluate the short-listed alternatives more thoroughly, such as the *additive equal weight* heuristic. In this strategy, the utilities of attribute values are combined additively in a similar manner to MAUT but without importance weights, which are assumed to be equal. Job selection procedures based on equal opportunity principles are often explicitly structured in this way, to make the selection process transparent. Applicants can be given a clear explanation as to why they were selected, short-listed or rejected, and selection panels can justify their decisions to those to whom they are accountable. Beach and his colleagues have developed image theory [1], which assumes a two-stage decision process similar to that described above, that is, a screening stage involving noncompensatory strategies, followed by a more thorough evaluation involving the selection of appropriate compensatory strategies. Other multistage process models describe rather more complex combinations of problem structuring operations and strategies [9, 16, 17].

Strategies for Decisions Involving Risk and Uncertainty

All the strategies discussed so far can be applied to decisions involving risk and uncertainty, if the outcomes and probabilities of the decision tree illustrated in Figure 1 are construed as attributes. However, various specific strategies have been proposed that recognize probability as being fundamentally different to other attributes. The most influential structural models to predict decisions under risk are variants of the subjectively expected utility (SEU) model,

which assume that outcome probabilities are transformed into weights of the subjective values of outcomes. For example, sign and rank dependent models such as prospect theory [4] and cumulative prospect theory [21]. In addition to compensatory strategies derived from SEU, various noncompensatory strategies have been suggested for decision under risk. These include the minimax strategy: choose the alternative with the lowest maximum possible loss. In Figure 1 this would lead to the selection of the alternative with the certain outcome, however attractive the possible gain of the other option might be.

Current Issues

It is important to distinguish between strategies for preferential as opposed to judgmental decisions. In the latter, a choice has to be made concerning whether A or B is closer to some criterion state of the world, present or future (e.g., which has the greater population, Paris or London?). Decision strategies for judgmental choice are discussed in the entry on fast and frugal heuristics. Preferential choice, as defined earlier, is fundamentally different since there is no real world 'best' state of the world against which the accuracy of a decision can be measured. Consequently, decision strategies for preferential choice, though superficially similar, are often profoundly different. In particular, preferential choice often involves some slow and costly, rather than fast and frugal thinking. Certainly, research within the naturalistic decision making framework [7] has identified several fast and frugal decision heuristic used to make preferential decisions (*see* **Heuristics: Fast and Frugal**). For example, the lexicographic strategy described earlier is essentially the same as the fast and frugal 'take the best' heuristic. Similarly, the recognition-primed decisions identified by Klein and his colleagues [5] have the same characteristics, and it has been found that people often switch to strategies that use less information under time pressure [8]. However, evidence related to multistage process theory points to a rather more complex picture. Especially when important decisions are involved, people spend considerable effort seeking and evaluating as much information as possible in order to clearly differentiate the best alternative from the others [16, 17]. One of the main challenges for future research is to model such multistage

decision strategies and validate them at the individual level.

References

- [1] Beach, L.R. & Mitchell, T.R. (1998). A contingency model for the selection of decision strategies, in *Image Theory: Theoretical and Empirical Foundations*, L.R. Beach ed., *LEA's Organization and Management Series; Image Theory: Theoretical and Empirical Foundations*. Lawrence Erlbaum, Mahwah, pp. 145–158.
- [2] Huber, O. (1989). Information-processing operators in decision making, in *Process and Structure in Human Decision Making*, H. Montgomery & O. Svenson, eds, Wiley, Chichester.
- [3] Kahneman, D. & Tversky, A. (1974). Judgment under uncertainty: heuristics and biases, *Science* **185**(4157), pp. 1124–1131.
- [4] Kahneman, D. & Tversky, A. (1979). Prospect theory: analysis of decision under risk, *Econometrica* **47**, 263–291.
- [5] Klein, G. (1989). Recognition-primed decisions, *Advances in Man-Machine Systems Research* **5**, 47–92.
- [6] Knight, F.H. (1921). *Risk, Uncertainty and Profit*, Macmillan, London.
- [7] Lipshitz, R., Klein, G., Orasanu, J. & Salas, E. (2001). Taking stock of naturalistic decision making, *Journal of Behavioral Decision Making* **14**, pp. 331–352.
- [8] Maule, A.J. & Edland, A.C. (1997). The effects of time pressure on human judgement and decision making, in *Decision Making: Cognitive Models and Explanations*, R. Ranyard, W.R. Crozier & O. Svenson, eds, Routledge, London.
- [9] Montgomery, H. (1983). Decision rules and the search for a dominance structure: Towards a process model of decision making, in *Analyzing and Aiding Decision Processes*, P.D. Humphries, O. Svenson & A. Vari, eds, North-Holland, Amsterdam.
- [10] Payne, J.W., Bettman, J.R. & Johnson, E.J. (1993). *The Adaptive Decision Maker*, Cambridge University Press, New York.
- [11] Ranyard, R. (1982). Binary choice patterns and reasons given for simple risky choice, *Acta Psychologica* **52**, 125–135.
- [12] Simon, H.A. (1957). *Models of Man*, John Wiley, New York.
- [13] Slovic, P. & Lichtenstein, S. (1971). Reversals of preference between bids and choices in gambling decisions, *Journal of Experimental Psychology* **89**(1), 46–55.
- [14] Slovic, P. & Lichtenstein, S. (1973). Response-induced reversals of preference in gambling: an extended replication in Las Vegas, *Journal of Experimental Psychology* **101**(1), 16–20.
- [15] Srivastava, J., Connolly, T. & Beach, L. (1995). Do ranks suffice? a comparison of alternative weighting approaches in value elicitation, *Organizational Behavior and Human Decision Processes* **63**, 112–116.

6 Decision Making Strategies

- [16] Svenson, O. (1992). Differentiation and consolidation theory of human decision making: a frame of reference for the study of pre- and post-decision processes, *Acta Psychologica* **80**, 143–168.
- [17] Svenson, O. (1996). Decision making and the search for fundamental psychological realities: what can be learned from a process perspective?, *Organizational Behavior and Human Decision Processes* **65**, 252–267.
- [18] Tversky, A. (1969). Intransitivity of preferences, *Psychological Review* **76**, 31–48.
- [19] Tversky, A. (1972). Elimination by aspects: a theory of choice, *Psychological Review* **79**(4), 281–299.
- [20] Tversky, A. & Fox, C.R. (1995). Weighing risk and uncertainty, *Psychological Review* **102**, 269–283.
- [21] Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: cumulative representations of uncertainty, *Journal of Risk and Uncertainty* **5**, 297–323.
- [22] Tversky, A., Sattah, S. & Slovic, P. (1988). Contingent weighting in judgment and choice, *Psychological Review* **95**(3), 371–384.

ROB RANYARD

Deductive Reasoning and Statistical Inference

JAMES CUSSENS

Volume 1, pp. 472–475

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Deductive Reasoning and Statistical Inference

Let us begin by differentiating between *probabilistic* and *statistical* inference. The former is the process of inferring probabilistic statements from other probabilistic statements and is entirely deductive in nature. For example, from $P(h) = 0.5$, $P(e|h) = 0.1$ and $P(e) = 0.2$ one can use Bayes theorem (see **Bayesian Belief Networks**) to infer, with absolute certainty, that $P(h|e) = P(h)P(e|h)/P(e) = 0.5 \times 0.1/0.2 = 0.25$.

Probabilistic inference is thus a subcategory of deductive inference. Statistical inference, in contrast, is the process of inferring general hypotheses about the world from particular observations (see **Classical Statistical Inference: Practice versus Presentation**). Such hypotheses cannot be held with certainty: from the observation of a hundred white swans we might infer that all swans are white, but we cannot do so with certainty. It is tempting to assert that from a hundred white swans we can at least infer either that *probably* all swans are white, or infer the statistical hypothesis that the next observed swan *probably* will be white. However, even these probabilistic options do not hold, as pointed out by the great empiricist philosopher David Hume (1711–76):

... all our experimental conclusions proceed upon the supposition, that the future will be conformable to the past. To endeavour, therefore, the proof of this last supposition by probable arguments, or arguments regarding existence, must be evidently going in a circle, and taking that for granted, which is the very point in question. [5, p. 35–36]

This is Hume's infamous *Problem of Induction* (although Hume did not use the word 'induction' in this context). Any hypothesis worth its salt will make predictions about future observations, but its success on past observations does not *on its own* guarantee successful future prediction. What is more, any claim concerning mere probabilities of various future events also make an unjustified claim that there is some connection between past and future.

The rest of this entry will consider two opposing ways out of this conundrum: the Bayesian approach and the Classical approach. In both cases, the connections with deductive inference will be emphasized.

The Bayesian approach evades the problem of induction by replacing statistical inference by probabilistic inference, which is deductive. To see how this is possible, let us consider a routine case of statistical inference, which will serve as a running example throughout: estimating an unknown probability (p) from a sequence (e) containing r 'successes' and $n - r$ 'failures'. So p might be the unknown probability that a particular coin lands heads, and e might be a sequence of n tosses of that coin, r of which land heads. No matter how large n is, there remains the possibility that the sample mean (r/n) is arbitrarily far from p . It follows that inferring that $p = r/n$ is nondeductive. More interestingly, if $0 < \epsilon < 1$, we cannot even deduce that $|p - r/n| < \epsilon$. If the coin tosses are somehow known to be *independent*, then the Weak **Law of Large Numbers** does at least gives us that $P(|p - r/n| < \epsilon) \rightarrow 1$ as $n \rightarrow \infty$. However, even in this case there is the underlying assumption that each coin toss is a so-called 'identical trial' governed by a fixed unknown probability p . Clearly, such trials are not 'identical' since different results occur. What is really being assumed here is that each coin toss is generated by a fixed set of 'generating conditions' [8] or, as Hacking would have it, a single 'chance set-up' [3]. A satisfactory account of just what is fixed in a given chance set-up and what may vary is still lacking.

The Bayesian approach avoids these difficulties by taking as given a *prior distribution* $f(p)$ over possible values of p (see **Bayesian Statistics**). It also assumes, in common with the Classical approach, that the *likelihood* $f(e|p)$ is given. Since $f(p, e) = f(p)f(e|p)$ this amounts to assuming that the joint distribution $f(p, e)$ is known. All that is then required is to compute the *posterior distribution* $f(p|e)$ (see **Bayesian Statistics**) and this is an entirely deductive process. How successful is this 'deductivization' of statistical inference? What has in fact happened is that any nondeductive inferences have been shoved to the beginning and to the end of the process of statistical inference.

In the beginning is the prior, whose derivation we will now consider. It is not possible to adopt a particular prior distribution by simply observing it to be the case, since, like all probability distributions, it is not directly observable. It is often argued, however, that, at least in some cases, it is possible to infer the true prior. Suppose we had absolutely no knowledge about p , then (one version of) the *Principle of*

2 Deductive Reasoning and Statistical Inference

Indifference permits us to infer that $f(p) = 1$, so that the prior is uniform and $P(a \leq p \leq b) = b - a$ for any a, b such that $0 \leq a \leq b \leq 1$. However, if we are indifferent about p , then surely we should be indifferent about p^2 also. Applying the Principle of Indifference to p^2 yields the prior $f'(p^2) = 1$ so that $P'(a \leq p^2 \leq b) = b - a$. But it then follows that $f'(p) = 1/(2\sqrt{p})$ and $P'(a \leq p \leq b) = \sqrt{b} - \sqrt{a}$. We have ‘logically inferred’ two inconsistent priors. Where the prior is over a single variable, it is possible to avoid this sort of inconsistency by adopting not a uniform distribution but Jeffreys’ noninformative distribution [2, p. 53]. However, when extended for multi-parameter distributions Jeffreys’ distributions are more problematic.

None of the vast literature attempting to rescue the Principle of Indifference from problems such as these is successful because the fundamental problem with the Principle is that it fallaciously claims to generate knowledge (in the form of a prior probability distribution) from ignorance. As Keynes memorably noted:

No other formula in the alchemy of logic has exerted more astonishing powers. For it has established the existence of God from total ignorance, and it has measured with numerical precision the probability that the sun will rise tomorrow. [6, p. 89], quoted in [4, p. 45]

Prior distributions can, in fact, only be deduced from other distributions. If the coin being tossed were selected at random from a set of coins all with *known* probabilities for heads, then a prior over possible values of p can easily be deduced. In such cases, a Bayesian approach is entirely uncontroversial. Classical statisticians will happily use this prior and any observed coin tosses to compute a posterior, because this prior is ‘objective’ (although, in truth, it is only as objective as the probabilities from which it was deduced). In any case, such cases are rare and rest on the optimistic assumption that we somehow know certain probabilities ahead of time. In most cases, the formulation of the prior is a mathematical incarnation of exogenously given *assumptions* about the likely whereabouts of p . If we adopt it as the ‘true’ prior, this amounts to positing these assumptions to be true: a nondeductive step. If we put it forward as merely an expression of our prior beliefs, this introduces an element of subjectivity – it is this to which the Classical view most strongly objects.

Consider now the end result of Bayesian inference: the posterior distribution. A pure Bayesian approach views the posterior as the end result of Bayesian inference:

Finally, never forget that the goal of Bayesian computation is not the posterior mode, not the posterior mean, but a representation of the entire *distribution*, or summaries of that distribution such as 95% intervals for estimands of interest [2, p. 301] (*italics in the original*)

However, the posterior is often used in a semi-Bayesian manner to compute a point estimate of p – the mean of the posterior is a favourite choice. But, as with all point estimates, acting as if some estimated value were the true value is highly nondeductive.

To sum up the principal objections to the Bayesian approach: these are that it simply does not tell us (a) what the right prior is and (b) what to do with the posterior. But as Howson and Urbach point out, much the same criticism can be made of deductive inference:

Deductive logic is the theory (though it might be more accurate to say ‘theories’) of deductively valid inferences from premisses whose truth-values are exogenously given. Inductive logic – which is how we regard subjective Bayesian theory – is the theory of inference from some exogenously given data and prior distribution of belief to a posterior distribution. . . . neither logic allows freedom to individual discretion: both are quite *impersonal* and objective. [4, p. 289–90]

The Classical approach to statistical inference is closely tied to the doctrine of *falsificationism*, which asserts that only statements that can be refuted by data are scientific. Note that the refutation of a scientific hypothesis (‘all swans are white’) by a single counterexample (‘there is a black swan’) is entirely deductive. Statistical statements are not falsifiable. For example, a heavy preponderance of observed tails is logically consistent with the hypothesis of a fair coin ($p = 0.5$). However, such an event is at least *improbable* if $p = 0.5$. The basic form of Classical statistical inference is to infer that a statistical hypothesis (the ‘null’ hypothesis) is refuted if we observe data that is sufficiently improbable (5% is a favourite level) on the assumption that the null hypothesis is true. In this case, the null hypothesis is regarded as ‘practically refuted’.

It is important to realize that, in the Classical view, improbable data does not make the null hypothesis

improbable. There is no probability attached to the truth of the hypothesis, since this is only possible in the Bayesian view. (We will assume throughout that we are not considering those rare cases where there exist the so-called ‘objective’ priors.) What then does it mean, on the Classical view, for a null hypothesis to be rejected by improbable data? Popper justified such a rejection on the grounds that it amounted to a ‘methodological decision to regard highly improbable events as ruled out – as prohibited’ [7, p. 191]. Such a decision amounts to adopting the nondeductive inference rule $P(e) < \epsilon \vdash \neg e$, for some sufficiently small ϵ . In English, this inference rule says: ‘From $P(e) < \epsilon$ infer that e is not the case’. With this nondeductive rule it follows that if h_0 is the null hypothesis and $h_0 \vdash P(e) < \epsilon$, then the observation of e falsifies h_0 in the normal way. A problem with this approach (due to Fisher [1]) concerns the choice of e . For example, suppose the following sequence of 10 coin tosses were observed $e = H, H, T, H, H, H, H, H, H, T$. If h_0 states that the coin is fair, then $h_0 \vdash P(e) = (1/2)^{10} < 0.001$. It would be absurd to reject h_0 on this basis, since any sequence of 10 coin tosses is equally improbable. This shows that care must be taken with the word ‘improbable’: with a large enough space of possible outcomes and a distribution that is not too skewed, then something improbable is bound to happen. If we could sample a point from a continuous distribution (such as the Normal distribution), then an event of probability zero would be guaranteed to occur!

Since e has been observed, we have also observed the events $e' = 'r = 8'$ and $e'' = 'r \geq 8'$. Events such as e'' are of the sort normally used in statistical testing; they assert that a *test statistic* (r) has been found to lie in a *critical region* (≥ 8). $h_0 \vdash P(e'') = 5.47\%$ (to 3 significant figures), so if we choose to test h_0 with e'' as opposed to e , h_0 is not rejected at significance level 5%, although it would, had we chosen 6%. e'' is a more sensible choice than e but there is no justified way of choosing the ‘right’ combination of test statistic and critical region.

In the modern (Neyman–Pearson) version of the Classical approach (see **Neyman–Pearson Inference**) a null hypothesis is compared to competing, alternative hypotheses. For example, suppose there were only one competing hypothesis h_1 , then h_0 would be rejected if $P(e|h_0)/P(e|h_1) \leq k$, where k is determined by the desired significance level. This turns out to solve the problem of choosing

e , but at the expense of a further move away from falsificationism. In the standard falsificationist approach, a black swan refutes ‘all swans are white’ *irrespective* of any other competing hypotheses.

Having dealt with these somewhat technical matters let us return to the deeper question of how to interpret the rejection of a hypothesis at significance level, say, 5%. Clearly, this is not straight logical refutation, nor (since it is non-Bayesian) does it even say anything about the probability that the hypothesis is false. In the literature, the nearest we get to an explanation is that one can *act as if the hypothesis were refuted*. This is generally justified on the grounds that such a decision will only rarely be mistaken: if we repeated the experiment many times, producing varying data due to sampling variation, and applied the same significance test then the hypothesis would not be erroneously rejected too often. But, in fact, it is only possible to infer that *probably* erroneous rejection would not occur often: it is possible (albeit unlikely) to have erroneous rejection every time. Also note that such a justification appeals to what *would* (or more properly *probably would*) happen if imaginary experiments were conducted. This is in sharp contrast to standard falsificationism, which, along with the Bayesian view, makes use only of the data we actually have; not any imaginary data. Another, more practical, problem is that rejected hypotheses cannot be resurrected if strongly supportive data is collected later on, or if other additional information is found. This problem is not present in the Bayesian case since new information can be combined with an ‘old’ posterior to produce a ‘new’ posterior, using Bayes theorem as normal. Finally, notice that confidence is not invested in any particular hypothesis rejection, but on the *process* that leads to rejection. Separating out confidence in the process of inference from confidence in the results of inference marks out Classical statistical inference from both Bayesian inference and deductive inference.

To finish this section on Classical statistical inference note that the basic inferential features of hypothesis testing also apply to Classical estimation of parameters. The standard Classical approach is to use an estimator – a function mapping the data to an estimate of the unknown parameter. For example, to estimate a probability p the proportion of successes r/n is used. If our data were e above, then the estimate for p would be $8/10 = 0.8$. Since we cannot ask questions about the likely accuracy of any particular

4 Deductive Reasoning and Statistical Inference

estimate, the Classical focus is on the *distribution* of estimates produced by a fixed estimator determined by the distribution of the data. An analysis of this distribution leads to **confidence intervals**. For example, we might have a 95% confidence interval of the form $(\hat{p} - \epsilon, \hat{p} + \epsilon)$, where $\hat{p}$ is the estimate. It is important to realize that such a confidence interval does not mean that the true value p lies within the interval $(\hat{p} - \epsilon, \hat{p} + \epsilon)$ with probability 95%, since this would amount to treating p as a random variable. See the entry on confidence intervals for further details.

References

- [1] Fisher, R.A. (1958). *Statistical Methods for Research Workers*, 13th Edition, Oliver and Boyd, Edinburgh. First published in 1925.
- [2] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, London.
- [3] Hacking, I. (1965). *Logic of Statistical Inference*, Cambridge University Press, Cambridge.
- [4] Howson, C. & Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle, Illinois.
- [5] Hume, D. (1777). *An Enquiry Concerning Human Understanding*, Selby-Bigge, London.
- [6] Keynes, J.M. (1921). *A Treatise on Probability*, Macmillan, London.
- [7] Popper, K.R. (1959). *The Logic of Scientific Discovery*, Hutchinson, London. Translation of *Logik der Forschung*, 1934.
- [8] Popper, K.R. (1983). *Realism and the Aim of Science*, Hutchinson, London. Written in 1956.

JAMES CUSSENS

DeFries–Fulker Analysis

RICHARD RENDE AND CHERYL SLOMKOWSKI

Volume 1, pp. 475–477

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

DeFries–Fulker Analysis

DeFries–Fulker, or DF, analysis [5] refers traditionally to a class of regression models that can be used to provide estimates of the fundamental behavioral genetic constructs **heritability** (h^2) and shared or common environment (c^2) (e.g., [1–3], [7], and [9]) (see **ACE Model**). Initially, two regression models were proposed. The basic model provided effect sizes of h^2 and c^2 on the basis of the individual differences in continuously measured traits. The critical addition of the DF approach, however, was to provide an augmented regression model that provided alternative measures of heritability and shared environment, specific to the extremes in the distribution. The conceptual advance was to provide measurable indicators of extreme group membership (i.e., elevated or diminished scores on a trait) that could be used in modeling heritability and shared environment. This approach offered an alternative to the more traditional biometrical liability model that utilizes statistical transformations of categorical indicators of familial association for diagnoses (e.g., twin concordance rates) (see **Twin Designs**) to yield quantitative genetic parameters on a hypothetical (rather than measured) continuum of liability to disease. The parameters from the two types of regression models used in DF analysis could then be compared to determine if the effect sizes differed for individual differences versus extreme scores, providing one empirical method for assessing if the etiology of the extremes was due in part to unique influences (either genetic or environmental) that are not influential for individual differences.

The basic DF model predicts one sibling's phenotype from the cosibling's as follows:

$$SP_1 = b_0 + b_1 SP_2 + b_2 R + b_3 R \times SP_2 + e \quad (1)$$

The dependent variable is sibling 1's phenotype (SP_1), which is a score on a continuously distributed trait. The independent variables are: (a) the second sibling's phenotype (SP_2); (b) the sibling pair's coefficient of genetic relatedness (R), which reflects the usual coding based on average genetic resemblance for additive traits (e.g., 1 for monozygotic twin, 0.5 for dizygotic twins and full siblings, 0.25 for half-siblings, and 0 for biologically unrelated siblings); and (c) the interaction between the second

sibling's score and the coefficient of genetic relatedness ($SP_2 \times R$). The intercept (b_0) and an error term (e) are also included.

The unstandardized β -weight on the SP_2 variable (b_1) estimates shared environment variance (c^2). The unstandardized β -weight on the interaction term (b_3) estimates heritability (h^2). For these DF analyses, all pairs are double-entered with adjustment for the standard errors of the parameter estimates (which are increased by a multiplicative constant because a standard error is based on the square root of the sample size). It should be noted that this regression model provides a method for conducting biometrical analyses that does not require specialized software and is quite flexible. For example, this basic model can be extended to examine other relevant variables (such as environmental factors) that may serve as moderators of both genetic and environmental influences (e.g., [10]).

Analyses of the extremes require two modifications to this basic model. First, probands are identified, typically as those individuals exceeding a preselected cutoff score in the distribution. Once identified, the regression model is modified to predict cosibling scores as a function of proband score and genetic relatedness. As described by DeFries and Fulker [2, 3], dropping the interaction term from the basic model, when applied to the restricted proband sample and their cosiblings, provides estimates of group heritability (h_g^2) via the unstandardized β -weight on the genetic relatedness (R) variable (b_2) and group-shared environmental influence (c_g^2) via the unstandardized β -weight on the proband phenotype (P) variable (b_1), provided that raw scores are transformed prior to analyses:

$$S = b_0 + b_1 P + b_2 R + e \quad (2)$$

Conceptually, the model analyzes **regression toward the mean** in the cosibling's score as a function of genetic relatedness [7]; heritability is implied by the extent to which such regression is conditional on genetic relatedness (with regression toward the mean increasing with decreasing genetic similarity), and shared environmental effects are suggested by the extent to which there is only partial regression toward the mean (i.e., higher than average cosibling scores), not conditional on genetic relatedness. If both siblings in a pair exceed the cutoff, one is randomly assigned as proband and the other as cosibling, although the double-entry method may also be used, along

with variations in the computation of group-shared environment [4]. This method may be used with both selected and unselected samples and is assumed to be robust to violations of normality in the distribution.

After determining the effect sizes of heritability and shared environment from both the basic and the augmented model, the individual difference and group parameters may be compared by contrasting confidence intervals for each. For example, it has been demonstrated that shared environmental influences are notable for elevated levels of depressive symptoms in adolescence but not for the full range of individual differences (e.g., [4] and [8]).

A model-fitting implementation of the DF method has been introduced, which preserves the function of the regression approach but allows for particular advantages [6]. Fundamental advances in this implementation include the analysis of pairs rather than individuals (eliminating the need for double entry of twin pairs and requisite correction of standard error terms) and the facility to include opposite-sex pairs in a sex-limited analysis. As described in detail in Purcell and Sham [6], the fundamental analytic strategy remains the same, as each observation (i.e., pair) contains a zygosity coefficient, continuous trait score for each member of the pair, and proband status for each member of the pair (i.e., a dummy variable indicating if they have exceeded a particular cutoff). Other details of the analytic procedure can be found in Purcell and Sham [6], including a script for conducting this type of analysis in the statistical program MX.

References

- [1] Cherny, S.S., DeFries, J.C. & Fulker, D.W. (1992). Multiple regression analysis of twin data: a model-fitting approach, *Behavior Genetics* **22**, 489–497.
- [2] DeFries, J.C. & Fulker, D.W. (1985). Multiple regression analysis of twin data, *Behavior Genetics* **15**, 467–473.
- [3] DeFries, J.C. & Fulker, D.W. (1988). Multiple regression analysis of twin data: etiology of deviant scores versus individual differences, *Acta Geneticae Medicae et Gemellologiae (Roma)* **37**, 205–216.
- [4] Eley, T.C. (1997). Depressive symptoms in children and adolescents: etiological links between normality and abnormality: a research note, *Journal of Child Psychology and Psychiatry* **38**, 861–865.
- [5] Plomin, R. & Rende, R. (1991). Human behavioral genetics, *Annual Review of Psychology* **42**, 161–190.
- [6] Purcell, S. & Sham, P.D. (2003). A model-fitting implementation of the DeFries-Fulker model for selected twin data, *Behavior Genetics* **33**, 271–278.
- [7] Rende, R. (1999). Adaptive and maladaptive pathways in development: a quantitative genetic perspective, in *On the Way to Individuality: Current Methodological Issues in Behavioral Genetics*, LaBuda, M., Grigorenko, E., Ravich-Serbo, I. & Scarr, S., eds, Nova Science Publishers, New York.
- [8] Rende, R., Plomin, R., Reiss, D. & Hetherington, E.M. (1993). Genetic and environmental influences on depressive symptoms in adolescence: etiology of individual differences and extreme scores, *Journal of Child Psychology and Psychiatry* **34**, 1387–1398.
- [9] Rodgers, J.L. & McGue, M. (1994). A simple algebraic demonstration of the validity of DeFries-Fulker analysis in unselected samples with multiple kinship levels, *Behavior Genetics* **24**, 259–262.
- [10] Slomkowski, C., Rende, R., Novak, S., Lloyd-Richardson, E. & Niaura, R. (in press). Sibling effects of smoking in adolescence: evidence for social influence in a genetically-informative design, *Addiction*.

RICHARD RENDE AND CHERYL SLOMKOWSKI

Demand Characteristics

AGNES DE MUNTER

Volume 1, pp. 477–478

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Demand Characteristics

It was as early as 1933 when Saul Rosenzweig [4] expressed his presumption that unverifiable factors may distort the research results of a psychological experiment. This formed the prelude to Martin T. Orne's first description of the demand characteristics of the experimental situation as 'the totality of cues which convey an experimental hypothesis to the subject' [1, p. 779]. He gave this definition at the symposium of the American Psychological Association in 1959 [3].

Demand characteristics come into being when research data are collected through the exchange of information between human participants, that is, between the researcher or the experimenter on the one hand, and the respondent or the experimental subject on the other hand. The respondents are aware of the fact that they are being observed and know that they are expected to show a particular research behavior. In situations like these, the researcher has no total control over the indications regarding the objectives or hypotheses introduced by either himself or the context.

This complex of unintended or unverified indications that make the respondent think he is able to guess the hypotheses or objectives of the researcher is referred to as the demand characteristics of the research situation.

This obscure tangle of indications is formed by, amongst other things, the data that the respondent had learnt about the research and the researcher prior to the research; the experiences of the respondent as respondent; the expectations and convictions of the respondent regarding the research itself; the setting where the research is conducted; and the manner of measuring.

Demand characteristics elicit answers from the respondents that the researcher can neither verify nor anticipate. Not every respondent reacts in the same way to the research hypotheses and questions he presupposes. Respondents take on different roles [5]. A good respondent will try to support the presumed research hypotheses or questions. He or she will want to meet the researcher part of the way. The behavior of many volunteering respondents is that of a 'good' respondent. The negativist respondent will answer to the research questions in such a manner that the research is obstructed and hindered. In the

interpretation of the effects observed, demand characteristics are responsible for errors of inference or artifacts. The changes in the dependent variables are not so much caused by the independent variables as by a particular disturbing factor, namely, the reaction of the respondents to the demand characteristics of the research. Rosnow [5] calls these factors systematic errors. They are mostly an implicit threat to the 'construct validity' and the **external validity** if the research is conducted from postpositivistic paradigms. However, if the research is conducted from interpretative paradigms, they constitute a threat to the 'credibility' and 'transferability'.

It is the task of the researcher to find out whether there are any demand characteristics present in the research that have significantly influenced respondent behavior.

Orne [1, 2] has developed three 'quasi-control' strategies to detect demand characteristics in the research.

A first possible strategy is the 'postexperimental interview'. The respondents' emotions, perceptions, and thoughts with regard to the research and the researcher are assessed by means of questionnaires. This is done in order to find out whether the respondent had some understanding of the research hypotheses. Another technique is 'preinquiry' or 'nonexperiment'. The experimental procedures are carefully explained to prospective respondents. Subsequently, they are asked to react as if they had actively participated in the planned experiment. If the same data are obtained with these respondents as with the actual participants, there may have been a possible influence of demand characteristics.

In a third strategy, that of 'the use of simulators', the researcher works with respondents acting as if they have been part of the experiment, whereas this is evidently not the case. Now, it is for the researcher to tell the genuine respondents from those who are simulating. If he is unable to do so, certain data in the case may be the result of demand characteristics.

The 'quasi-control' strategies may help the researcher to detect the presence of demand characteristics. The actual bias, however, cannot be proved. Some scientists hope to neutralize this form of bias by applying techniques to reduce the systematic errors caused by the demand characteristics to unsystematic errors, if possible.

2 Demand Characteristics

Robson [3] proposes to restrict the interaction between the researcher and the respondent as much as possible. In his view, this can be done by making use of taped instruction or of the automated presentation of material.

References

- [1] Orne, M. (1962). On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications, *American Psychologist* **17**(11), 776–783.
- [2] Orne, M. (1969). Demand characteristics and the concept of quasi control, in *Artifact in Behavioral Research*, R. Rosenthal & R.L. Rosnow, eds, Academic Press, New York, pp. 143–179.
- [3] Robson, C. (1993). *Real World Research. A Resource for Social Scientists and Practitioner-researchers*, Blackwell, Oxford.

- [4] Rosenzweig, S. (1933). The experimental situation as a psychological problem, *Psychological Review* **40**, 337–354.
- [5] Rosnow, R.L. (2002). The nature and role of demand characteristics in scientific inquiry, *Prevention & Treatment* **5**, Retrieved July 16, 2003, from <http://www.Journals.apa.org/prevention/volume5/pre0050037c.html>

Further Reading

- Orne, M. (1959). The demand characteristics of an experimental design and their implications, in *Paper Presented at the Symposium of the American Psychological Association*, Cincinnati.

AGNES DE MUNTER

Deming, Edwards William

SANDY LOVIE

Volume 1, pp. 478–479

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Deming, Edwards William

Born: October 14, 1900, in Iowa.

Died: December 20, 1993, in Washington.

William Deming's career was as colorful and varied as it was long. Thus, he began his working life as a mathematical physicist, but then switched to the role of consultant and propagandist to the newly emerging field of American industrial and business statistics, a field that has produced much of modern statistical theory and practice. This change had been prefigured in his earliest publication titled *Statistical Adjustment of Data* [1], first published in 1938 but available in a paperback edition as late as 1985. The book represented 16 years of work both as teacher and consultant and researcher, and is mainly concerned with integrating all the various flavors of curve fitting, where such curves provide the replacement for or adjustment of the original data. Thus, adjustment here means the choice of a representative number or function for a sample. The main criterion for this selection is that the number supplies the information for *action*. There is, in other words, a strong operational and instrumental force behind Deming's views on statistics. The motivation for his work was, therefore, highly practical and as a consequence led him to work in both academic and business and Government settings: He was, for example, employed by the American Department of Agriculture and then was with the Bureau of the Census, for nearly 20 years in total. During this period, he also introduced American statisticians to such British and European luminaries as **Ronald Fisher** and **Jerzy Neyman**, both of whom he had studied with in the United Kingdom in 1936. He had also collected and edited a volume of the latter's unpublished lectures and conference contributions.

After what appears to be an idyllic but hard early life on a farm in Wyoming, Deming had graduated from the University of Wyoming in 1921 as an electrical engineer. His subsequent degrees included a Masters from the University of Colorado in 1924 (he had taught at the Colorado School of Mines for 2 years prior to this degree) and a Ph.D. in mathematical physics from Yale in 1928. In 1927, Deming began his long association with Government

until 1946 when he was appointed to a Chair in the Graduate School of Business Administration at the University of New York (he also held the position of 'Distinguished Lecturer' in Management at Columbia University from 1986). Here, his career as international consultant and management *guru* flourished (but not curiously in postwar America), with extensive tours in South America, Europe, and Asia, particularly in Japan where he had been adviser on sampling techniques to the Supreme Command of the Allied Powers in Tokyo as early as 1947. His long-term espousal of quality control methods, particularly those invented by Shewhart, together with his extensive practical experience of sampling in various industrial settings (see [2]) and his enormous capacity for hard work meant that he was a natural to help Japan become an industrial world leader from the 1960s onward. It also appears that the Japanese businessmen who visited Deming in America were equally impressed by him as a person, for example, that he lived in an unpretentious house in Washington, DC, and that he had clearly not attempted to make any serious money from his efforts [3]. According to Paton, therefore, many leading Japanese businessmen and politicians viewed Deming as a 'virtual god' [4]. Indeed, the somewhat religious aspect of his work was revealed during his many Four Day Seminars on quality control and management, where the famous Fourteen Points to successful management had to be accepted with minimal dissent by the participants (see [4], for an eyewitness account of a seminar given in 1992, when Deming had only a year to live). Only after American industry had been seriously challenged by Japan in the late 1970s and early 1980s had there arisen a call to adopt Deming's winning formula, a neglect that he seemed never to have really forgotten or forgiven.

References

- [1] Deming, W.E. (1938). *Statistical Adjustment of Data*, Wiley, New York.
- [2] Deming, W.E. (1950). *Some Theory of Sampling*, Wiley, New York.
- [3] Halberstam, D. (1966). *The Reckoning*, William Morrow, New York.
- [4] Paton, S.M. (1993). Four days with Dr Deming, *Quality Digest*. February

SANDY LOVIE

Design Effects

NEIL KLAR AND ALLAN DONNER

Volume 1, pp. 479–483

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Design Effects

Introduction

Survey researchers can only be assured of obtaining a representative sample of subjects when a probability sampling plan is used [9]. Simple random sampling (SRS) in which all eligible study subjects have the same probability of being selected, is the most basic of such plans (*see Randomization; Survey Sampling Procedures*). However, this design is usually recommended only when one has little additional information beyond a list of eligible subjects. More sophisticated sampling plans tend to be used in most actual surveys.

Consider, for example, the multistage, stratified, cluster sample used by the National Crime Victimization Survey (NCVS) [9]. This survey estimates the amount of crime in the United States, supplementing crime statistics compiled by the Federal Bureau of Investigation (FBI), known to be limited to crimes reported to law enforcement agencies. According to the NCVS sampling plan, random samples of census enumeration districts are selected from within strata defined on the basis of geographic location, demographic data, and rates of reported crime. This sampling plan increases the precision with which crime statistics are estimated, while ensuring national and local representation. Clusters of households are then sampled from within enumeration districts and finally all household members at least 12 years of age are asked about their experience as victims of crime over the past six months. The selection of subjects from within household clusters is an effective and cost-efficient method of identifying eligible subjects across a wide age range.

Sample size estimation for sampling plans typical of that used in the NCVS can be quite complicated. A method of simplifying calculations is to assume data will be obtained using SRS and then accounting for the use of **stratification** and clustering using design effects. We define this concept in section ‘what is a design effect?’, and demonstrate its role in sample size estimation for a cluster randomized trial in section ‘worked example’. This example also demonstrates that, while design effects were originally defined by Kish [6, 7, 13] in the context of survey sampling, they are in fact much more

broadly applicable. Some related issues are discussed in section ‘discussion’.

What is a Design Effect?

A goal of the NCVS is to estimate p , the proportion of respondents who were victims of crime in the six months prior to being surveyed. The design effect associated with $\hat{p}$ (the estimator of p) in the NCVS is given by

$$deff = \frac{Var_{NCVS}(\hat{p})}{Var_{SRS}(\hat{p})}, \quad (1)$$

where $Var_{NCVS}(\hat{p})$ denotes the variance of $\hat{p}$ obtained using the NCVS sampling plan while $Var_{SRS}(\hat{p})$ denotes the variance of this statistic obtained using SRS. More generally, a design effect is defined as the ratio of the variance of an estimator for a specified sampling plan to the variance of the same estimator for SRS, assuming a fixed overall sample size. In this sense, the design effect measures the statistical efficiency of a selected sampling plan relative to that of SRS.

The selection of subjects from within strata and the use of cluster sampling are often motivated by practical concerns, as is seen in the NCVS. Stratification typically results in gains in statistical efficiency ($deff < 1$) to the extent that selected strata are predictive of the study outcome. In this case, greater precision is secured by obtaining an estimate of variability that is calculated within strata. In the NCVS, FBI crime statistics were used to define strata, a strategy intended to increase the precision of $\hat{p}$. Conversely, cluster sampling typically reduces statistical efficiency ($deff > 1$), since subjects from the same cluster tend to respond more similarly than subjects from different clusters, thus inflating estimates of variance. The competing effects of stratification and cluster sampling combine in the NCVS to provide a design effect for $\hat{p}$ of about two [9], indicating that SRS is twice as efficient as the selected sampling plan. Consequently NCVS investigators would need to enroll twice as many subjects as they would under SRS in order to obtain an equally efficient estimate of p .

Worked Example

Greater insight into the interpretation and calculation of design effects may be obtained by considering a

2 Design Effects

worked example from a school-based smoking prevention trial. As part of the Four Group Comparison Study [10] children were randomly assigned by school to either one of three smoking prevention programs or to a control group. Randomization by school was adopted for this trial since allocation at the individual level would have been impractical and could also have increased the possibility that children in an intervention group could influence the behavior of control group children at the same school.

Unfortunately, the selected programs in this trial proved ineffective in reducing tobacco use among adolescents. However, suppose investigators decided to design a new trial focusing specifically on p , the proportion of children using smokeless tobacco. The corresponding design effect is given by

$$deff = \frac{Var_{SR}(\hat{p})}{Var_{IR}(\hat{p})}, \quad (2)$$

where $\hat{p}$ denotes the sample estimate of p , $Var_{SR}(\hat{p})$ denotes the variance of $\hat{p}$ assuming random assignment by school and $Var_{IR}(\hat{p})$ denotes the variance of $\hat{p}$ assuming individual random assignment. One can show [3] that in this case, $deff \approx 1 + (\bar{m} - 1)\rho$, where $\bar{m}$ denotes the average number of students per school and ρ is the **intraclass correlation** coefficient measuring the similarity in response between any two students in the same school. With the additional assumption that ρ is nonnegative, this parameter may also be interpreted as the proportion of overall variation in response that can be accounted for by between-school variation.

Data are provided in Table 1 for the rates of smokeless tobacco use among the students from the 12 control group schools randomized in the Four Group Comparison Study [3, 10], where the average number of students per school is given by $\bar{m} = 1479/12 = 123.25$. Therefore, the sample estimate of ρ may be calculated [3] as

$$\begin{aligned} \hat{\rho} &= 1 - \frac{\sum_{j=1}^k m_j \hat{p}_j (1 - \hat{p}_j)}{k(\bar{m} - 1) \hat{p} (1 - \hat{p})} \\ &= 1 - \frac{[152 \times 0.0132(1 - 0.0132)] + \dots + [125 \times 0.1280(1 - 0.1280)]}{12(123.25 - 1)0.0615(1 - 0.0615)} \\ &= 0.013324, \end{aligned} \quad (3)$$

where $\hat{p}_j = y_j/m_j$ and $\hat{p} = Y/M = 91/1479 = 0.0615$ denote the prevalence of smokeless tobacco

Table 1 Number and proportion of children who report using smokeless tobacco after 2 years of follow-up in each of 12 schools [3, 10]

j	y_j	m_j	$\hat{p}_j$
1	2	152	0.0132
2	3	174	0.0172
3	1	55	0.0182
4	3	74	0.0405
5	5	103	0.0485
6	12	207	0.0580
7	7	104	0.0673
8	7	102	0.0686
9	6	83	0.0723
10	6	75	0.0800
11	23	225	0.1022
$k = 12$	16	125	0.1280
Total	$Y = 91$	$M = 1479$	$\hat{p} = Y/M = 0.0615$

use in the j th school and among all 12 schools, respectively.

The estimated design effect is then seen to be given approximately by

$$1 + (123.25 - 1)0.013 = 2.6, \quad (4)$$

indicating that random assignment by school would require more than twice the number of students as compared to an individually randomized trial having the same power. Note that the design effect is a function of both the degree of intraclass correlation and the average number of students per school, so that even values of $\hat{\rho}$ close to zero can dramatically inflate the required sample size when $\bar{m}$ is large.

Now suppose investigators believe that their new intervention can lower the rate of smokeless tobacco use from six percent to three percent. Then using standard sample size formula for an individually randomized trial [3], approximately 746 students would be required in each of the two groups at a 5% two-sided significance level with 80% power. However, this result needs to be multiplied by $deff = 2.6$, implying that at least 16 schools need to be randomized per intervention group assuming approximately 123 students per school ($746 \times 2.6/123.25 = 15.7$). In practice, investigators should consider a range of plausible values for the design effect as it is frequently estimated with low precision in practice.

Discussion

Trials randomizing schools are an example of a more general class of intervention studies known as cluster randomization trials. The units of randomization in such studies are diverse, ranging from small clusters such as households or families, to entire neighborhoods and communities but also including worksites, hospitals, and medical practices [3]. Cluster randomization has become particularly widespread in the evaluation of nontherapeutic interventions, including lifestyle modification, educational programmes, and innovations in the provision of health care.

We have limited attention here to trials in which clusters are randomly assigned to intervention groups without the benefit of matching or stratification with respect to baseline factors perceived to have prognostic importance. Design effects for these relatively simple designs are analogous to those obtained for surveys involving a one-stage cluster sampling scheme [9]. One distinction is that investigators designing surveys often incorporate finite population correction factors since then clusters are typically selected without replacement [9]. However, such correction factors have very little impact when the selected clusters represent only a small fraction of the target population.

Comparisons of design effects computed across studies have led to some interesting empirical findings. For instance, estimates of design effects often vary less across similar variables or studies than does the estimated variance [1]. This greater 'portability' argues for the use of design effects in place of the variance of a statistic when estimating sample size in practice. Data from cluster randomization trials and from surveys using cluster sampling also reveal that the degree to which responses of cluster members are correlated, and consequently the size of the resulting design effect, will tend to vary as a function of cluster size [3, 5]. Not surprisingly, responses among subjects in smaller clusters (e.g., households) tend to be more highly correlated than responses among subjects in larger clusters (e.g., communities). This is intuitively sensible, since people from the same household tend to be more alike than randomly selected subjects who live in the same city. However, although the magnitude of the intraclass correlation coefficient tends to decline with cluster size, it does so at a relatively slow rate. Thus, design effects of greater magnitude are usually obtained in

studies enrolling large clusters such as communities than in studies randomizing households, even though the value of the intraclass correlation tends to be much greater for the latter.

Our discussion has focused on the application of design effects for sample size estimation when the outcome of interest is binary. However, sample size estimation procedures that incorporate design effects have also been reported for studies having a range of other outcomes (e.g., continuous [3], count [3], time to event [15]) and more generally to allow for covariate adjustment (*see Analysis of Covariance*) [11]. Extensions of these formulae have also been derived for surveys where analyses need to consider the effects of both cluster sampling and weighting to account for unequal selection probabilities [4]. Design effects can furthermore be incorporated to compare costs of different sampling plans, as discussed by Kish [6] and Connelly [2].

The application of design effects is not limited to sample size estimation and to comparing the efficiency of competing sampling plans. They can also be used to adjust standard test statistics for features of the sampling plan [3, 9, 12] although these procedures become more approximate when covariate adjustment is required [9, 11, 14]. In this case it is preferable to use more sophisticated regression models requiring software capable of simultaneously accounting for weighting, stratification, and clustering (see, e.g. [8]).

Confidentiality concerns may impose limits on the amount of information that can be released about the design of a survey. For example, the public-use data from the NCVS did not include information about who belonged to the same cluster [9]. Data analysts will then have to approximate the effects of clustering using average design effects [14] or, in their absence, exploit the availability of external data to assign an *a priori* value for the design effect. Given the sensitivity of statistical inferences to the assigned values of the design effect, we would discourage this latter practice unless very reliable and representative external data are available.

References

- [1] Clark, R.G. & Steel, D.G. (2002). The effect of using household as a sampling unit, *International Statistical Review* 70, 289–314.

4 Design Effects

- [2] Connelly, L.B. (2003). Balancing the number and size of sites: an economic approach to the optimal design of cluster samples, *Controlled Clinical Trials* **24**, 544–559.
- [3] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*, Arnold, London.
- [4] Gabler, S., Haeder, S. & Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering, *Survey Methodology* **25**, 105–106.
- [5] Hansen, M.H. & Hurwitz, W.N. (1942). Relative efficiencies of various sampling units in population inquiries, *Journal of the American Statistical Association* **37**, 89–94.
- [6] Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.
- [7] Kish, L. (1982). Design Effect, *Encyclopedia of Statistical Sciences*, Vol. 2, John Wiley & Sons, New York, pp. 347–348.
- [8] LaVange, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. & Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples, *Statistical Methods in Medical Research* **5**, 311–329.
- [9] Lohr, S.L. (1999). *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove.
- [10] Murray, D.M., Perry, C.L., Griffin, G., Harty, K.C., Jacobs Jr, D.R., Schmid, L., Daly, K. & Pallonen, U. (1992). Results from a statewide approach to adolescent tobacco use prevention, *Preventive Medicine* **21**, 449–472.
- [11] Neuhaus, J.M. & Segal, M.R. (1993). Design effects for binary regression models fit to dependent data, *Statistics in Medicine* **12**, 1259–1268.
- [12] Rao, J.N.K. & Scott, A.J. (1992). A simple method for the analysis of clustered binary data, *Biometrics* **48**, 577–585.
- [13] Rust, K. & Frankel, M. (2003). Issues in inference from survey data, in *Leslie Kish, Selected Papers*, S. Heeringa & G. Kalton, eds, John Wiley & Sons, New York, pp. 125–129.
- [14] Williams, D.A. (1982). Extra-binomial variation in logistic linear models, *Applied Statistics* **31**, 144–148.
- [15] Xie, T. & Waksman, J. (2003). Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint, *Statistics in Medicine* **22**, 2835–2846.

(See also **Survey Sampling Procedures**)

NEIL KLAR AND ALLAN DONNER

Development of Statistical Theory in the 20th Century

PETER M. LEE

Volume 1, pp. 483–485

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Development of Statistical Theory in the 20th Century

At the beginning of the twentieth century, the dominant figures in statistics were **Francis Galton (1822–1911)**, **Karl Pearson (1857–1936)**, and **George Udny Yule (1871–1951)**. Galton, whose mathematics was rudimentary but whose insight was truly remarkable, had developed the theory of regression and, subsequently, of correlation. Karl Pearson was an enthusiastic follower of Galton who wished to apply statistical methods to substantial quantities of real data. In doing so, he developed various families of statistical distributions, which he tried to fit to his data. This led naturally to the search for a method of estimating parameters and hence to his ‘method of moments’ and, subsequently, to the need for a measure of ‘goodness of fit’ and hence to the development of the chi-squared test (*see Goodness of Fit for Categorical Variables*). His methods were unsuitable for small samples, and it was ‘Student’ (a pen name for **William Sealy Gosset, 1876–1937**) who began to develop small-sample methods with his statistic, subsequently modified to give what is now known as Student’s t .

The towering genius of twentieth-century statistics was **Ronald Aylmer Fisher (1890–1962)** who further developed the theory of small samples, giving rigorous proofs of the distributions of χ^2 , t , and Fisher’s z (which is equivalent to F). He further established the principle of maximum likelihood (*see Maximum Likelihood Estimation*) as a method of estimation, which proved far more successful than the method of moments. Unfortunately, both Karl Pearson and Fisher were of a quarrelsome nature and they fell out over various issues, including the distribution of the correlation coefficient, the correct number of degrees of freedom in a 2×2 contingency table and methods of estimation.

Fisher worked from 1919 to 1933 at the Rothamsted agricultural research station, which led to his development of most of the present-day techniques of the design and analysis of experiments. Subsequently, he was Galton Professor of Eugenics at University College, London, and Arthur Professor of Genetics

at Cambridge, and he always regarded himself primarily as a scientist. His books [1, 2, 4] (especially the first two) were immensely influential, but can be difficult to follow because proofs and mathematical details are omitted.

Fisher’s use of significance tests is well illustrated by his discussion in Chapter III of [4] of the observation that the star β Capricorni has five close neighbors among 1500 bright stars for which he has data. Assuming that the stars are randomly distributed about the celestial sphere, Fisher shows that this event has a probability of 1 in 33 000. He concludes that *either* an exceptionally rare event has occurred *or* the theory of random distribution is not true. In Fisher’s view, scientific theories should be tested in this way, and, while they may, at a certain P value (1 in 33 000 in the example) be rejected, they can never be finally accepted.

Fisher’s theory of inference was largely based on his ‘fiducial argument’ (expounded, e.g., in [4]), which depends on the existence of ‘pivotal quantities’ whose distribution does not depend on unknown parameters. For example, if $x_1, \dots, x_n$ are independently normally distributed with mean μ and variance σ^2 and their sample mean and standard deviation are m and s respectively, then $t = (m - \mu)/(s/\sqrt{n})$ is a pivotal quantity with Student’s t distribution on $n - 1$ degrees of freedom, whatever μ and σ may be. Fisher then deduced that μ had a fiducial distribution, which was that of $m - (s/\sqrt{n})t$ (where, in finding the distribution, m and s are thought of as constants). Often, but not invariably, the fiducial distribution is the same as that given by the method of inverse probability, which is essentially the Bayesian argument discussed below using uniform prior distributions. A particularly controversial case is the Behrens–Fisher problem, in which the means μ_1 and μ_2 of two samples from independent normal distributions are to be compared. Then, if $\delta = \mu_1 - \mu_2$ and $d = m_1 - m_2$ the distribution of $\delta - d$ is that of $(s_1/\sqrt{n_1})t_1 - (s_2/\sqrt{n_2})t_2$, where t_1 and t_2 are independent variates with t distributions. In this case (and others), the probability given by the fiducial argument is not that with which rejection takes place in repeated samples if the theory were true. Although this fact ‘caused him no surprise’ [3, p. 96], it is generally regarded as a defect of the fiducial argument.

In the early 1930s, **Jerzy Neyman (1894–1981)** and Karl Pearson’s son **Egon Sharpe Pearson (1895–1980)** developed their theory of hypothesis

testing (see **Neyman–Pearson Inference**). In the simplest case, we are interested in knowing whether an unknown parameter θ takes the value θ_0 or the value θ_1 . The first possibility is referred to as the null hypothesis H_0 and the second as the alternative hypothesis H_1 . They argue that if we are to collect data whose distribution depends on the value of θ , then we should decide on a ‘rejection region’ R , which is such that we reject the null hypothesis if and only if the observations fall in the rejection region. Naturally, we want to minimize the probability $\alpha = P(R|\theta_0)$ of rejecting the null hypothesis when it is true (an ‘error of the first kind’), whereas we want to maximize the probability $1 - \beta = P(R|\theta_1)$ of rejecting it when it is false (thus avoiding an ‘error of the second kind’). Since increasing R decreases β but increases α , a compromise is necessary. Neyman and Pearson accordingly recommended restricting attention to tests for which α was less than some preassigned value called the *size* (for example 5%) and then choosing among such regions one with a maximum value of the ‘power’ $1 - \beta$. Fisher, however, firmly rebutted the view that the purpose of a test of significance was to decide between two or more hypotheses.

Neyman and Pearson went on to develop a theory of ‘**confidence intervals**’. This can be exemplified by the case of a sample of independently normally distributed random variables (as above), when they argued that if the absolute value of a t statistic on $n - 1$ degrees of freedom is less than $t_{n-1,0.95}$ with 95% probability, then the random interval $(m - st_{n-1,0.95}, m + st_{n-1,0.95})$ will contain the true, unknown value μ of the mean with probability 0.95. Incautious users of the method are inclined to speak as if the value μ were random and lay within that interval with 95% probability, but from the Neyman–Pearson standpoint, this is unacceptable, although under certain circumstances, it may be acceptable to Bayesians as discussed below.

Later in the twentieth century, the Bayesian viewpoint gained adherents (see **Bayesian Statistics**). While there were precursors, the most influential early proponents were **Bruno de Finetti** (1906–1985) and **Leonard Jimmie Savage** (1917–1971). Conceptually, Bayesian methodology is simple. It relies on Bayes theorem, that $P(H_i|E) \propto P(H_i)P(E|H_i)$, where the H_i constitute a set of possible hypotheses and E a body of evidence. In the

case where the ‘prior probabilities’ $P(H_i)$ are all equal, this is essentially equivalent to the method of ‘inverse probability’, which was popular at the start of the twentieth century. Thus, if we assume that $x_1, \dots, x_n$ are independently normally distributed with unknown mean μ and known variance σ^2 , then assuming that all values of μ are equally likely *a priori*, it is easy to deduce that *a posteriori* the distribution of μ is normal with mean m (the sample mean) and variance σ^2 . In the case where σ^2 is unknown, another conventional choice of prior beliefs for σ^2 , this time uniform in its logarithm, leads to the t Test, which Student and others had found by classical methods. Nevertheless, particularly in the continuous case, there are considerable difficulties in taking the standard conventional choices of prior, and these difficulties are much worse in several dimensions. Controversy over Bayesian methods has centered mainly on the choice of prior probabilities, and while there have been attempts to find an objective choice for prior probabilities (see, e.g., [5, Section 3.10]), it is now common to accept that the prior probabilities are chosen subjectively. This has led some statisticians and scientists to reject Bayesian methods out of hand. However, with the growth of **Markov Chain Monte Carlo Methods**, which have made Bayesian methods simple to use and made some previously intractable problems amenable to solution, they are now gaining in popularity.

For some purposes, formal statistical theory has declined in importance with the growth of ‘**Exploratory Data Analysis**’ as advocated by **John W Tukey (1915–2000)** and of modern graphical methods as developed by workers such as William S. Cleveland, (1943), but for many problems, the debate about the foundations of statistical inference remains lively and relevant.

References

- [1] Fisher, R.A. (1925). *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.
- [2] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [3] Fisher, R.A. (1935). The fiducial argument in statistics, *Annals of Eugenics* **6**, 391–398.
- [4] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Oliver & Boyd, Edinburgh.
- [5] Jeffreys, H. (1939). *Theory of Probability*, Oxford University Press, Oxford.

Further Reading

- Berry, D.A. (1996). *Statistics: A Bayesian Perspective*, Duxbury, Belmont.
- Box, J.F. (1978). *R.A. Fisher: The Life of a Scientist*, Wiley, New York.
- Cleveland, W.S. (1994). *The Elements of Graphing Data*, Revised Ed., AT&T Bell Laboratories, Murray Hill.
- Pearson, E.S. (1966). The Neyman-Pearson story: 1926–34. Historical sidelights on an episode in Anglo-Polish collaboration, in F.N. David ed., *Festschrift for J Neyman*, Wiley, New York; Reprinted on pages 455–477 of E.S. Pearson & M.G. Kendall eds, (1970). *Studies in the History of Statistics and Probability*, Griffin Publishing, London.
- Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.
- Porter, T.M. (2004). *Karl Pearson: The Scientific Life in a Statistical Age*, Princeton University Press, Princeton.
- Reid, C. (1982). *Neyman – From Life*, Springer-Verlag, New York.
- Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W H Freeman, New York.
- Savage, L.J. (1962). *The Foundations of Statistical Inference; a Discussion Opened by Professor L J Savage*, Methuen, London.
- Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

PETER M. LEE

Differential Item Functioning

H. JANE ROGERS

Volume 1, pp. 485–490

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Differential Item Functioning

Differential item functioning (DIF) occurs when individuals at the same level on the trait(s) or construct(s) being assessed but from different subpopulations have unequal probabilities of attaining a given score on the item. The critical element in the definition of DIF is that individuals from the different subpopulations are matched on the traits of interest before their responses are compared. Performance differences that remain after matching on the relevant variables must be because of group differences on an additional trait being tapped by the item. Note that while unidimensionality of the test is commonly assumed in practice, DIF can occur in multidimensional tests whenever a test item taps an unintended trait on which groups differ [16].

By its definition, DIF is distinguished from impact, which is defined as a simple mean difference in performance on a given item for the subpopulations of interest. Impact occurs whenever there is a mean difference between subpopulations on the trait. Impact is easily observed and understood by nonpsychometricians, and as a result has become the legal basis for exclusion of items in some high-stakes tests; psychometricians repudiate such practices because of the failure of nonpsychometricians to take into account valid group differences. A further distinction is made between DIF and bias. While DIF was once known as ‘item bias’, modern psychometric thought distinguishes the concepts; the term ‘bias’ carries an evaluative component not contained in the detection of differential performance on the item. Holland and Wainer [10] define item bias as ‘an informed judgment about an item that takes into account the purpose of the test, the relevant experiences of certain subgroups of examinees taking it, and statistical information about the item.’ Hence, evidence of DIF is necessary but not sufficient for the conclusion of bias.

In DIF studies, two groups are usually compared at a time, one of which is denoted the reference group; the other, the focal group. Typically, the majority group (e.g., Whites, males) is taken as the reference group and the focal group is the minority group of interest.

Two types of DIF are commonly differentiated: uniform and nonuniform [17]. Uniform DIF is present when the difference between (or ratio of) the probabilities of a given response in the reference and focal groups is constant across all levels of the trait; that is, the item is uniformly more difficult for one group than the other across the trait continuum. Nonuniform DIF is present when the direction of differences in performance varies or even changes direction in different regions of the trait continuum; that is, the item is differentially discriminating. Hanson [8] makes further distinctions between uniform, unidirectional, and parallel DIF.

The study of DIF began in the 1960s [1] and has been vigorously pursued in the intervening years. The large majority of research on DIF methods has focused on dichotomously scored items; only since the early 1990s has much attention been given to methods for detecting DIF in polytomously scored items. Early ‘item bias’ methods were based on item-by-group interactions and failed to adequately disentangle DIF from impact. These early methods have been reviewed and their shortcomings identified by a number of authors [26]; none of these methods is currently in wide use and hence are not described here.

As noted above, the central element in the definition of DIF is that the comparison of reference and focal groups be conditional on trait level. The manner in which that conditioning is operationalized provides a convenient basis for classifying DIF methods in current use. Millsap and Everson [18] distinguish methods based on unobserved conditional invariance (UCI) from those based on observed conditional invariance (OCI). UCI methods condition on model-based estimates of the trait, while OCI methods condition on an observed proxy for the trait, typically total test score. The primary UCI methods are those based on item response theory (*see Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data*).

Item response theory (IRT) provides a natural framework for defining and investigating DIF. Item response models specify the probability of a given response to an item as a function of the unobservable trait underlying performance and characteristics of the item [7]. If item response functions (IRFs) for a specific item differ for subpopulations of interest, this indicates that examinees at the same trait value do not

2 Differential Item Functioning

have the same probability of the response, and hence that DIF is present. Thus, IRT methods for detecting DIF involve comparison of IRFs. Approaches to quantifying this comparison are based on comparison of item parameters, comparison of item characteristic curves, or comparison of model fit. Millsap and Everson [18] provide a comprehensive review of IRT DIF methods that encompasses the majority of research done on these methods to date.

Lord [15] proposed a statistic, often referred to as D^2 , to test for equality of the vectors of item parameters for the two groups. The test statistic requires the vector of item parameter estimates and the variance-covariance matrices of the estimates in each group and has an approximate chi-square distribution with degrees of freedom equal to the number of item parameters compared. Lord [15] notes that the chi-square distribution is asymptotic and strictly holds only when the true theta values are known. A practical problem with the D^2 statistic is that the variance-covariance matrix of the item parameter estimates is often not well-estimated [28].

Differences between item characteristic functions have been quantified by computing the area between the curves. Bounded and unbounded area statistics have been developed. Bounded statistics are necessary when the c-parameters for the two groups differ, as in this case, the area between item characteristic curves is infinite. Kim and Cohen [12] developed a formula for bounded area, but no standard error or test statistic has been derived. Raju [21] provided a formula for computing the area between IRFs when the c-parameters are equal. Raju [22] further provided a standard error formula and derived an approximately normal test statistic.

Likelihood ratio statistics can be used to compare the fit of a model based on equality constraints on parameters across all items with that of a model in which the item parameters for the studied item are estimated separately within groups. Thissen et al. [28] described a likelihood ratio test statistic proportional to the difference in the log likelihoods under the constrained and unconstrained models. The likelihood ratio and D^2 statistics are asymptotically equivalent; Thissen et al. [28] argue that the likelihood ratio test may be more useful in practice because it does not require computation of the variance-covariance matrix of the parameter estimates. However, the likelihood ratio procedure requires fitting a

separate model for each item in the test; by comparison, the D^2 statistic requires only one calibration per group.

Limitations of IRT methods include the necessity for large sample sizes in order to obtain accurate parameter estimates, the requirement of model-data fit prior to any investigation of DIF, and the additional requirement that the parameter estimates for the two groups be on a common scale before comparison. Each of these issues has been subject to a great deal of research in its own right. IRT methods for detecting DIF remain current and are considered the theoretical ideal that other methods approximate; however, for the reasons given above, they are not as widely used as more easily computed methods based on OCI.

A variety of OCI-based methods have been proposed. All use total score in place of the latent variable. Problems with the use of total score as a conditioning variable have been noted by many authors [19]. A fundamental issue is that the total score may be contaminated by the presence of DIF in some of the items. ‘Purification’ procedures are routinely used to ameliorate this problem; these procedures involve an initial DIF analysis to identify DIF items, removal of these items from the total score, and recomputation of the DIF statistics using the purified score as the conditioning variable [9, 17].

Logistic regression procedures [23, 27] most closely approximate IRT methods by using an observed score in place of the **latent variable** and in essence fitting a two-parameter model in each group. Swaminathan and Rogers [27] reparameterized to produce an overall model, incorporating parameters for score, group, and a score-by-group interaction. An overall test for the presence of DIF is obtained by simultaneously testing the hypotheses that the group and interaction parameters differ from zero, using a chi-square statistic with two degrees of freedom. Separate one degree of freedom tests (or z tests) for nonuniform and uniform DIF are possible by testing hypotheses about the interaction and group parameters, respectively. Zumbo [29] suggested an R-square **effect size measure** for use with the logistic regression procedure. The logistic regression procedure is a generalization of the procedure based on the **loglinear model** proposed by Mellenbergh [17]; the latter treats total score as a categorical variable. The advantages of the logistic regression procedure over other OCI methods are

its generality and its flexibility in allowing multiple conditioning variables.

The standardization procedure of Dorans and Kulick [6] also approximates IRT procedures by comparing empirical item characteristic curves, using total score on the test as the proxy for the latent trait. Unlike the logistic regression procedure, the standardization index treats observed score as a categorical variable; differences between the probability of a correct response (in the case of dichotomously scored items) are computed at each score level, weighted by the proportion of focal group members at that score level, and summed to provide an index of uniform DIF known as the *standardized P-DIF statistic*. Dorans and Holland [5] provide standard errors for the standardization index but no test of significance. Because it requires very large sample sizes to obtain stable estimates of the item characteristic curves, the standardization index is not widely used; it is used largely by Educational Testing Service (ETS) as a descriptive measure for DIF in conjunction with the **Mantel–Haenszel (MH)** procedure introduced by Holland and Thayer [9].

The MH procedure is the most widely known of the OCI methods. The procedure for dichotomously scored items is based on contingency tables of item response (right/wrong) by group membership (reference/focal) at each observed score level. The null hypothesis tested under the MH procedure is that the ratio of the **odds** for success in the reference versus focal groups is equal to one at all score levels. The alternative hypothesis is that the **odds ratio** is a constant, denoted by α . This alternative hypothesis represents uniform DIF; the procedure is not designed to detect nonuniform DIF. The test statistic has an approximate chi-square distribution with one degree of freedom. Holland and Thayer [9] note that this test is the uniformly most powerful (*see Power*) unbiased test of the null hypothesis against the specified alternative. The common odds ratio, α , provides a measure of effect size for DIF. Holland and Thayer [9] transformed this parameter to the ETS delta scale so that it can be interpreted as the constant difference in difficulty of the item between reference and focal groups across score levels. Holland and Thayer [9] note that the MH test statistic is very similar to the statistic given by Mellenbergh [17] for testing the hypothesis of uniform DIF using the log-linear model. Swaminathan and Rogers [27] showed equivalence between the MH and logistic regression

procedures if the logistic model is reexpressed in logit form, the total score treated as categorical, and the interaction term omitted. In this case, the group parameter is equal to $\log \alpha$. The primary advantages of the MH procedure are its ease of calculation and interpretable effect size measure.

Also of current interest is the SIBTEST procedure proposed by Shealy and Stout [25]. Shealy and Stout [25] note that this procedure can be considered an extension of the standardization procedure of Dorans and Kulick [6]. Its primary advantages over the standardization index are that it does not require large samples and that it provides a test of significance. It is conceptually model-based but non-parametric. The procedure begins by identifying a ‘valid subtest’ on which conditioning is based and a ‘studied subtest’, which may be a single item or group of items. The SIBTEST test statistic is based on the weighted sum of differences in the average scores of reference and focal group members with the same valid subtest true score. Shealy and Stout [25] show that if there are group differences in the distribution of the trait, matching on observed scores does not properly match on the trait, so they base matching instead on the predicted valid subtest true score, given observed valid subtest score. Shealy and Stout [25] derive a test statistic that is approximately normally distributed. Like the standardization and MH procedures, the SIBTEST procedure is designed to detect only uniform DIF. Li and Stout [14] modified SIBTEST to produce a test sensitive to nonuniform DIF. Jiang and Stout [11] offered a modification of the SIBTEST procedure to improve Type I error control and reduce estimation bias. The advantages of the SIBTEST procedure are its strong theoretical basis, relative ease of calculation, and effect size measure.

Extensions of all of the DIF procedures described above have been developed for use with polytomous models. Cohen, Kim, and Baker [4] developed an extension of the IRT area method and an accompanying test statistic and provided a generalization of Lord’s D^2 . IRT likelihood ratio tests are also readily extended to polytomous item response models [2, 13]. Rogers and Swaminathan [24] described logistic regression models for unordered and ordered polytomous responses and provided chi-square test statistics with degrees of freedom equal to the number of item parameters being compared. Zwick, Donoghue,

and Grima [30] gave an extension of the standardization procedure for ordered polytomous responses based on comparison of expected responses to the item; Zwick and Thayer [31] derived the standard error for this statistic. Zwick et al. [30] also presented generalized MH and Mantel statistics for unordered and ordered polytomous responses, respectively; the test statistic in the unordered case is distributed as a chi-square with degrees of freedom equal to one less than the number of response categories, while the test statistic in the ordered case is a chi-square statistic with one degree of freedom. These authors also provide an effect size measure for the ordered case. Chang, Mazzeo, and Roussos [3] presented an extension of the SIBTEST procedure for polytomous responses based on conditional comparison of expected scores on the studied item or subtest. Potenza and Dorans [20] provided a framework for classifying and evaluating DIF procedures for polytomous responses.

Investigations of DIF remain an important aspect of all measurement applications; however, current research efforts focus more on interpretations of DIF than on development of new indices.

References

- [1] Angoff, W.H. (1993). Perspectives on differential item functioning methodology, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 3–23.
- [2] Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods, *Applied Measurement in Education* **15**, 113–141.
- [3] Chang, H.H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure, *Journal of Educational Measurement* **33**, 333–353.
- [4] Cohen, A.S., Kim, S.H. & Baker, F.B. (1993). Detection of differential item functioning in the graded response model, *Applied Psychological Measurement* **17**, 335–350.
- [5] Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 35–66.
- [6] Dorans, N.J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test, *Journal of Educational Measurement* **23**, 355–368.
- [7] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, Boston.
- [8] Hanson, B.A. (1998). Uniform DIF and DIF defined by differences in item response functions, *Journal of Educational and Behavioral Statistics* **23**, 244–253.
- [9] Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure, in *Test Validity*, H. Wainer & H.I. Braun, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 129–145.
- [10] Holland, P.W. & Wainer, H., eds (1993). *Differential Item Functioning*, Lawrence Erlbaum Associates, Hillsdale.
- [11] Jiang, H. & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST, *Journal of Educational Measurement* **23**, 291–322.
- [12] Kim, S.H. & Cohen, A.S. (1991). A comparison of two area measures for detecting differential item functioning, *Applied Psychological Measurement* **15**, 269–278.
- [13] Kim, S.-H. & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test, *Applied Psychological Measurement* **22**, 345–355.
- [14] Li, H. & Stout, W. (1996). A new procedure for detection of crossing DIF, *Psychometrika* **61**, 647–677.
- [15] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Erlbaum, Hillsdale.
- [16] Mazor, K., Kanjee, A. & Clauser, B.E. (1995). Using logistic regression and the Mantel-Haenszel procedure with multiple ability estimates to detect differential item functioning, *Journal of Educational Measurement* **32**, 131–144.
- [17] Mellenbergh, G.J. (1982). Contingency table models for assessing item bias, *Journal of Educational Statistics* **7**, 105–118.
- [18] Millsap, R.E. & Everson, H.T. (1993). Methodology review: statistical approaches for assessing measurement bias, *Applied Psychological Measurement* **17**, 297–334.
- [19] Millsap, R.E. & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias, *Applied Psychological Measurement* **16**, 389–402.
- [20] Potenza, M.T. & Dorans, N.J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation, *Applied Psychological Measurement* **19**, 23–37.
- [21] Raju, N.S. (1988). The area between two item characteristic curves, *Psychometrika* **53**, 495–502.
- [22] Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement* **14**, 197–207.
- [23] Rogers, H.J. & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning, *Applied Psychological Measurement* **17**, 105–116.
- [24] Rogers, H.J. & Swaminathan, H. (1994). Logistic regression procedures for detecting DIF in nondichotomous

- item responses, in *Paper Presented at the Annual AERA/NCME Meeting*, New Orleans.
- [25] Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF, *Psychometrika* **58**, 159–194.
- [26] Shepard, L., Camilli, G. & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research, *Journal of Educational Statistics* **9**, 93–128.
- [27] Swaminathan, H. & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement* **27**, 361–370.
- [28] Thissen, D., Steinberg, L. & Wainer, H. (1992). Use of item response theory in the study of group differences in trace lines, in *Test Validity*, H. Wainer & H.I. Braun, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 147–170.
- [29] Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*, Directorate of Human Resources Research and Evaluation, Department of National Defense, Ottawa.
- [30] Zwick, R., Donoghue, J.R. & Grima, A. (1993). Assessment of differential item functioning for performance tasks, *Journal of Educational Measurement* **30**, 233–251.
- [31] Zwick, R. & Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items, *Journal of Educational and Behavioral Statistics* **21**, 187–201.

H. JANE ROGERS

Direct and Indirect Effects

DOMINIQUE MULLER AND CHARLES M. JUDD

Volume 1, pp. 490–492

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Direct and Indirect Effects

Suppose that a researcher wants to study the impact of an independent variable X on a dependent variable Y . To keep things simple, imagine that X has two conditions, a treatment condition and a control condition, and, ideally (in order to make causal inferences a bit easier), that participants have been randomly assigned to one or the other of these conditions. In this context, the total effect will be defined as the effect of X on Y , and can be, for instance, estimated in the context of a linear regression model (see **Multiple Linear Regression**) as the following:

$$Y = \beta_{01} + \tau X + \varepsilon_1 \quad (1)$$

It should be noted that in some situations, for instance when Y is a latent construct, the X to Y relationship would be estimated using **structural equation modeling**. In (1), the slope τ thus represents the total effect of X on Y . This relationship can be schematized as in Figure 1. The estimate of τ in a given sample is often labeled c .

It is possible that the researcher will be interested in finding the mechanism responsible for this $X - Y$ relationship [1–3]. The researcher may hypothesize that a third variable is partially responsible for the observed effect. The question will then be: Does a part of the total effect go through a third variable, often called a mediator (or an intervening variable)? Is there an indirect effect of X on Y going through Me (as Mediator)? (see **Mediation**) This three-variable relationship can be schematized as in Figure 2. In

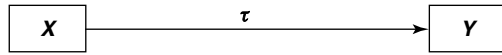


Figure 1 Total effect

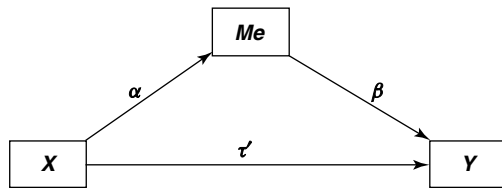


Figure 2 Direct and indirect effects

this context, the indirect effect will be the product of the path going from X to Me and the path going from Me to Y . Frequently, the former will be called α in the population or a in the sample, and, the latter, β in the population and b in the sample. This indirect effect will then be $\alpha * \beta$ in the population and $a * b$ in the sample. In regression terms, the two following models will be run in order to estimate both α and β :

$$Me = \beta_{02} + \alpha X + \varepsilon_2 \quad (2)$$

$$Y = \beta_{03} + \tau' X + \beta Me + \varepsilon_3 \quad (3)$$

As can be seen in (3), and in line with Figure 2, the β path is the effect of Me on Y controlling for X . In this same equation, we estimate the $X:Y$ partial relationship once Me has been controlled. This is equivalent to the remaining (or residual) effect of X on Y once the part of this relationship that goes through Me has been removed. This coefficient is frequently labeled τ' (or c' in the sample). This is the direct effect of X on Y . It should be noted that the term ‘direct’ must be understood in relative terms, given that there may be other mediators that potentially explain this direct effect. Hence, in the case of two mediators Me_1 and Me_2 , the direct effect would be the residual effect of X on Y not explained by either Me_1 or Me_2 .

As a summary, so far we have seen that:

$$\text{Total effect} = \tau$$

$$\text{Indirect effect} = \alpha\beta, \quad (4)$$

and

$$\text{Direct effect} = \tau'. \quad (5)$$

Furthermore, it can be shown, not surprisingly, that the total effect (τ) is equal to the direct effect (τ') plus the indirect effect ($\alpha\beta$) (e.g., [5]). In other words:

$$\tau = \tau' + \alpha\beta. \quad (6)$$

It follows that:

$$\tau - \tau' = \alpha\beta. \quad (7)$$

From this last equation, it can be seen that the indirect effect can be estimated either by $\alpha\beta$ or $\tau - \tau'$, the change in effect of X on Y when controlling and not controlling for Me . This is of importance because, if mediation is at stake, it should be the

2 Direct and Indirect Effects

case that $|\tau| > |\tau'|$. Hence, the magnitude of the $X - Y$ relationship should decrease once the mediator is controlled. The total effect must be of a higher magnitude than the direct effect. It could happen, however, that $|\tau| < |\tau'|$. If it is the case, the third variable is not a mediator but a suppressor (e.g., [4]). In this case, this third variable does not reflect a possible mechanism for the relationship between X and Y , but, on the contrary, hides a portion of this relationship. When the third variable is a suppressor, the relationship between X and Y is stronger once this third variable is controlled.

Extension to Models with More Than One Mediator

As suggested above, in some cases, multiple mediator models could be tested. Then, there will be multiple possible indirect effects. For instance, in a model like the one presented in Figure 3, there will be an indirect effect through Me_1 (i.e., $\alpha_1\beta_1$) and another one through Me_2 (i.e., $\alpha_2\beta_2$). In such a situation, it is still possible to calculate an overall indirect effect taking into account both the indirect effect through Me_1 and through Me_2 . This one will be the sum of $\alpha_1\beta_1$ and $\alpha_2\beta_2$ (or a_1b_1 and a_2b_2 in terms of their sample estimates). It also follows that:

$$\tau - \tau' = \alpha_1\beta_1 + \alpha_2\beta_2. \quad (8)$$

Illustration

Two researchers conducted a study with only female participants [6]. They first showed, in line with the

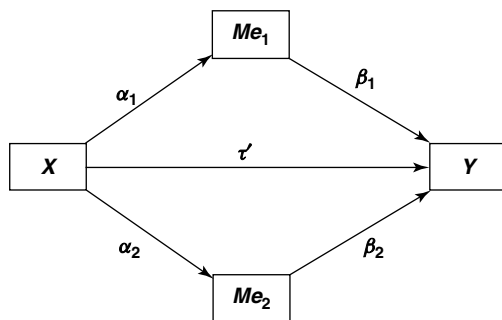


Figure 3 Two indirect effects

so-called stereotype threat literature (e.g., [7]), that these female participants performed less well at a math test in a condition making salient their gender (see [6] for details on the procedure) than in a control condition. The test of this effect of condition on math performance is, thus, the *total effect*. The linear regression conducted on these data revealed a standardized estimate of $c = -0.42$ (in this illustration, we used the standardized estimates simply because these are what these authors reported in their article. The same algebra would apply with the unstandardized estimates). Next, these authors wanted to demonstrate that this difference in math performance was due to a decrease in working memory in the condition where female gender was made salient. In other words, they wanted to show that working memory mediated the condition effect. In order to do so, they conducted two additional regression analyses: The first one, regressing a working memory measure on condition, and the second, regressing math performance on both condition and working memory (i.e., the mediator). The first one gives us the path from condition to working memory ($a = -0.52$); the second one gives us the path from the mediator to the math performance controlling for condition ($b = 0.58$) and the path from condition to math controlling for working memory ($c' = -0.12$). So, in this illustration, the *indirect effect* is given by $a * b = -0.30$ and the *direct effect* is $c' = -0.12$. This example also illustrates that, as stated before, $c - c' = ab$, so that $-0.42 - (-0.12) = -0.52 * 0.58$. The total effect, -0.42 is, thus, due to two components, the direct effect (-0.12) and the indirect effect (-0.30).

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] James, L.R. & Brett, J.M. (1984). Mediators, moderators, and tests for mediation, *Journal of Applied Psychology* **69**, 307–321.
- [3] Judd, C.M. & Kenny, D.A. (1981). Process analysis: estimating mediation in treatment evaluation, *Evaluation Review* **5**, 602–619.
- [4] MacKinnon, D.P., Krull, J.L. & Lockwood, C.M. (2000). Equivalence of the mediation, confounding and suppression effect, *Prevention Science* **1**, 173–181.

- [5] MacKinnon, D.P., Warsi, G. & Dwyer, J.H. (1995). A simulation study of mediated effect measures, *Multivariate Behavioral Research* **30**, 41–62.
- [6] Schmader, T. & Johns, M. (2002). Converging evidence that stereotype threat reduces working memory capacity, *Journal of Personality and Social Psychology* **85**, 440–452.
- [7] Steele, C.M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans, *Journal of Personality and Social Psychology* **69**, 797–811.

DOMINIQUE MULLER AND CHARLES M. JUDD

Direct Maximum Likelihood Estimation

CRAIG K. ENDERS

Volume 1, pp. 492–494

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Direct Maximum Likelihood Estimation

Many widely used statistical procedures (e.g., structural equation modeling, multilevel modeling) rely on **maximum likelihood estimation** (ML) to obtain estimates of model parameters. The basic goal of ML is to identify the population parameter values most likely to have produced a particular sample of data. The fit of the data to a set of parameter values is gauged by the *log likelihood* (see **Maximum Likelihood Estimation**), a value that quantifies the relative probability of a particular sample, given that the data originated from a normally-distributed population with a mean vector and covariance matrix, μ and Σ , respectively. Estimation usually requires the use of iterative algorithms that ‘try out’ different values for the unknown parameters, ideally converging on the parameter values that maximize the log likelihood.

ML estimation is ideally suited for analyses of incomplete datasets, and requires less stringent assumptions about the missingness mechanism than traditional methods such as listwise deletion. So-called *direct maximum likelihood* (direct ML; also known as *full information maximum likelihood*) is widely available in commercial software packages, and the number of models to which this estimator can be applied is growing rapidly. To better appreciate the benefits of direct ML estimation, it is necessary to understand Rubin’s [4] taxonomy of missing data mechanisms. A more detailed treatment of missing data can be found elsewhere in this volume (see **Missing Data; Dropouts in Longitudinal Data**), but a brief overview is provided here.

Data are said to be *missing completely at random* (MCAR) when missingness on a variable Y is unrelated to other variables as well as the values of Y itself (i.e., the observed data are essentially a random sample of the hypothetically complete data). *Missing at random* (MAR) is less stringent in the sense that missingness on Y is related to another observed variable, X , but is still unrelated to the values of Y . For example, suppose a counseling psychologist is investigating the relationship between perceived social support and depression, and finds that individuals with high levels of support are more likely to have missing values on the depression inventory (e.g., they

prematurely terminate therapy because of their perceived support from others). Finally, data are *missing not at random* (MNAR) when missing values on Y are related to the values of Y itself. Returning to the previous example, suppose that individuals with low levels of depression are more likely to have missing values on the depression inventory, even after controlling for social support (e.g., they prematurely terminate therapy because they don’t feel depressed).

Although behavioral researchers are unlikely to have explicit knowledge of the missing data mechanism (only MCAR can be empirically tested), the theoretical implications of Rubin’s [4] taxonomy are important. Rubin showed that unbiased and efficient parameter estimates could be obtained from likelihood-based estimation (e.g., direct ML, multiple imputations) under MAR. If one views the missing data mechanism as a largely untestable assumption, the implication is that ML estimation is more ‘robust’ in the sense that unbiased and efficient estimates can be obtained under MCAR and MAR, whereas traditional approaches generally produce unbiased estimates under MCAR only.

We will explore the basic principles of direct ML estimation using the dataset in Table 1. The data consist of 10 scores from a depression inventory and measure of perceived social support, and an MAR mechanism was simulated by deleting depression scores for the three cases having the highest levels of social support. Additionally, the social support variable was randomly deleted for a single case.

As described previously, the ML log likelihood quantifies the relative probability of the data, given a normally distributed population with an unknown

Table 1 Hypothetical depression and perceived social support data

Complete		Missing	
Depression	Support	Depression	Support
17	4	17	4
22	6	22	6
28	10	28	10
13	11	13	?
7	12	7	12
17	13	17	13
13	15	13	15
17	18	?	18
10	19	?	19
6	22	?	22

2 Direct Maximum Likelihood Estimation

mean vector and covariance matrix, μ and Σ , respectively. ML estimation proceeds by ‘trying out’ values for μ and Σ in an attempt to identify the estimates that maximize the log likelihood. In the case of direct ML, each case’s contribution to the log likelihood is computed using the observed data for that case. Assuming multivariate normality, case i ’s contribution to the sample log likelihood is given by

$$\log L_i = K_i - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i), \quad (1)$$

where $\mathbf{y}_i$ is the vector of raw data, $\boldsymbol{\mu}_i$ is the vector of population means, Σ_i is the population covariance matrix, and K_i is a constant that depends on the number of observed values for case i . The sample log likelihood is subsequently obtained by summing over the N cases, as shown in (2).

$$\log L = K - \frac{1}{2} \sum_{i=1}^N \log |\Sigma_i| - \frac{1}{2} \sum_{i=1}^N (\mathbf{y}_i - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i). \quad (2)$$

The case subscript i implies that the size and content of the arrays differ according to the missing data pattern for each case. To illustrate, let us return to the data in Table 1. Notice that there are three missing data patterns: six cases have complete data, a single case is missing the social support score (Y_1), and three cases have missing values on the depression variable (Y_2). The contribution to the log likelihood for each of the six complete cases is computed as follows:

$$\begin{aligned} \log L_i &= K_i - \frac{1}{2} \log \begin{vmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{vmatrix} - \frac{1}{2} \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)' \\ &\quad \times \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right) \end{aligned} \quad (3)$$

For cases with missing data, the rows and columns of the arrays that correspond to the missing values are removed. For example, the arrays for the three cases with missing depression scores (Y_2) would include only y_1 , μ_1 , and σ_{11}^2 . Substituting these quantities into (1), the contribution to the log likelihood for each of these three cases is computed as

$$\log L_i = K_i - \frac{1}{2} \log |\sigma_{11}| - \frac{1}{2} (y_i - \mu_1)'$$

Table 2 Maximum likelihood parameter estimates

Variable	Estimate		
	Complete	Direct ML	Listwise
Mean			
Support	13.00	13.32	10.00
Depression	15.00	15.12	17.33
Variance			
Support	29.00	30.99	15.00
Depression	40.80	44.00	43.56
Covariance			
Support/Depression	-19.90	-18.22	-10.00

$$[\sigma_{11}]^{-1} (y_i - \mu_1) \quad (4)$$

In a similar vein, the log likelihood for the case with a missing social support score (Y_1) is computed using y_2 , μ_2 , and σ_2^2 .

To further illustrate, direct ML estimates of μ and Σ are given in Table 2. For comparative purposes, μ and Σ were also estimated following a listwise deletion of cases ($n = 6$). Notice that the direct ML estimates are quite similar to those of the complete data, but the listwise deletion estimates are fairly distorted. These results are consistent with theoretical expectations, given that the data are MAR (for depression scores, missingness is solely due to the level of social support). Moreover, there is a straightforward conceptual explanation for these results. Because the two variables are negatively correlated ($r = -0.58$), the listwise deletion of cases effectively truncates the marginal distributions of both variables (e.g., low depression scores are systematically removed, as are high support scores). In contrast, direct ML utilizes all observed data during estimation. Although it may not be immediately obvious from the previous equations, cases with incomplete data are, in fact, contributing to the estimation of all parameters. Although depression scores are missing for three cases, the inclusion of their support scores in the log likelihood informs the choice of depression parameters via the linear relationship between social support and depression.

As mentioned previously, ML estimation requires the use of iterative computational algorithms. One such approach, the *EM algorithm* (see **Maximum Likelihood Estimation**), was originally proposed as a method for obtaining ML estimates of μ and Σ

with incomplete data [1], but has since been adapted to complete-data estimation problems as well (e.g., **hierarchical linear models**; [3]). EM is an iterative procedure that repeatedly cycles between two steps. The process begins with initial estimates of μ and Σ , which can be obtained via any number of methods (e.g., listwise deletion). The purpose of the *E*, or *expectation*, step is to obtain the sufficient statistics required to compute μ and Σ (i.e., the variable sums and sums of products) in the subsequent *M* step. The complete cases simply contribute their observed data to these sufficient statistics, but missing *Y*s are replaced by predicted scores from a regression equation (e.g., a missing Y_1 value is replaced with the predicted score from a regression of Y_1 on Y_2 , Y_3 , and Y_4 , and a residual variance term is added to restore uncertainty to the imputed value). The *M*, or *maximization*, step uses standard formulae to compute the updated covariance matrix and mean vector using the sufficient statistics from the previous *E* step. This updated covariance matrix is passed to the next *E* step, and is used to generate new expectations for the missing values. The two EM steps are repeated until the difference between covariance matrices in adjacent *M* steps becomes trivially small, or falls below some *convergence criterion*.

The previous examples have illustrated the estimation of μ and Σ using direct ML. It is important to

note that a wide variety of linear models (e.g., regression, structural equation models, multilevel models) can be estimated using direct ML, and direct ML estimation has recently been adapted to nonlinear models as well (e.g., logistic regression implemented in the Mplus software package). Finally, when provided with the option, it is important that direct ML standard errors be estimated using the observed, rather than expected, **information matrix** (a matrix that is a function of the second derivatives of the log likelihood function), as the latter may produce biased standard errors [2].

References

- [1] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
- [2] Kenward, M.G. & Molenberghs, G. (1988). Likelihood based frequentist inference when data are missing at random, *Statistical Science* **13**, 236–247.
- [3] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models*, 2nd Edition, Sage, Thousand Oaks.
- [4] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.

CRAIG K. ENDERS

Directed Alternatives in Testing

ARTHUR COHEN

Volume 1, pp. 495–496

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Directed Alternatives in Testing

A common statistical testing situation is the consideration of a null hypothesis, which specifies that the means of populations are the same (or are homogeneous). For example, three different types of diet plans may be independently administered to three samples from populations of obese individuals of similar composition. The data consist of the amount of weight loss for each individual in the study. In such a case, the typical alternative is that the true 3 average (mean) weight losses for the populations are not the same. Such an alternative would be regarded as a nonrestricted alternative or one which has parameters, in a sense, in every direction. On the other hand, suppose one diet consisted of taking an innocuous pill (placebo), a second diet plan involved calorie reduction plus a bonafide diet pill, and a third diet regimen included everything in the second diet plan plus counseling. The null hypothesis is still that the true average weight loss is the same for the three diet plans, but a more realistic and more useful alternative is that the average weight loss is nondecreasing as we vary from the first to second to third diet plan. Such an alternative would be a directed alternative of the simple order type. Formally, if μ_i , $i = 1, 2, 3$ represents the true mean weight loss for diets 1, 2, and 3 respectively, the null hypothesis is $H : \mu_1 = \mu_2 = \mu_3$. The unrestricted alternative denoted by K_U is all parameter points except those in H . A directed alternative called the *simple order alternative* is $K_{S0} : \mu_1 \leq \mu_2 \leq \mu_3$, but not $\mu_1 = \mu_2 = \mu_3$.

The advantage of posing a directed alternative when it is appropriate to do so is that the appropriate statistical procedure has a much greater ability to detect such an alternative. In statistical jargon, this means that the power of the test for the restricted alternative can be decidedly greater than the **power** for the test of K_U .

When the directed alternative is simple order and normality assumptions for the data are reasonable, likelihood ratio tests (LRTs) can be recommended (see **Maximum Likelihood Estimation**). These are outlined in [6, Chapter 2]. This reference also contains a table on p. 95 from which one can measure the huge gains in power by subscribing to a test suitable

for the simple order alternative. Robertson, Wright, and Dykstra [6] also offer tests for the simple order alternative for binomial, multinomial, and Poisson data (see **Catalogue of Probability Density Functions**). Usually, large samples are required in these latter cases. For small samples and discrete data, the method of directed vertex peeling, DVP, as outlined in [2], is effective. See Example 4.1 of that reference for details. For nonparametric models, one can use rank test methodology as given by [1].

Another directed alternative commonly encountered in practical situations is the tree order alternative. Such an alternative is often appropriate when comparing k different treatment regimens with a control. Formally, if μ_i denotes the true mean for the i th treatment, $i = 1, \dots, k$, and μ_0 denotes the true mean for the control then the tree order alternative is $K_T : \mu_i - \mu_0 \geq 0, i = 1, \dots, k$, with strict inequality for some i .

For the simple order directed alternative, the LRTs can unhesitatedly be recommended. For the tree order directed alternative, a variety of tests have different types of advantageous properties. Dunnett [5] proposes a test for normal data with precise critical values. Cohen and Sackrowitz [3] recommend both a modification of Dunnett's test and a second test procedure with desirable monotonicity properties. That same reference contains a test appropriate for the nonparametric case. The DVP methodology is appropriate for discrete data and small sample sizes.

A third type of directed alternative that is important and frequently encountered is stochastic ordering. We describe this alternative for a $2 \times k$ **contingency table** with ordered categories. The rows of the table correspond to, say, control and treatment, whereas the columns of the table correspond to ordered responses to treatment. An example is a 2×3 table where the first row represents a placebo treatment and the second row represents an active treatment. The three columns represent respectively, no improvement, slight improvement, and substantial improvement.

The 2×3 table has cell frequencies as follows in Table 1:

Table 1

	No Improvement	Slight Improvement	Substantial Improvement	
Placebo	X_1	X_2	X_3	n_1
Treatment	Y_1	Y_2	Y_3	n_2

2 Directed Alternatives in Testing

Table 2

	No Improvement	Slight Improvement	Substantial Improvement
Placebo	p_1	p_2	p_3
Treatment	q_1	q_2	q_3

Here, X_1 , for example represents the number of n_1 individuals who received the placebo and had no improvement. The corresponding table of probabilities is as follows in Table 2:

Here, for example, p_1 represents the probability of having no improvement given the placebo was taken. Also $p_1 + p_2 + p_3 = q_1 + q_2 + q_3 = 1$.

The null hypothesis of interest is that the placebo probability distribution is the same as the treatment distribution. That is $H : p_i = q_i, i = 1, \dots, k$. The directed alternative is $K_{ST} : \sum_{j=1}^i p_j \geq \sum_{j=1}^i q_j, i = 1, \dots, k$, with at least one inequality. Thus, for $k = 3$, $K_{ST} : p_1 \geq q_1, p_1 + p_2 \geq q_1 + q_2$, with at least one strict inequality. The idea behind this directed alternative is that the probability distribution for the treatment population is more heavily concentrated on the higher-ordered categories.

To test H versus K_{ST} in a $2 \times k$ table, the methodology called *directed chi-square* is recommended. This methodology appears in [4]. One can simply input the data of a $2 \times k$ table into the following website: <http://stat.rutgers.edu/~madigan/dvp.html>. A conditional P value is quickly provided, which can be used to accept or reject H .

Some concluding and cautionary remarks are needed. We have offered three types of directed alternatives. There are many others. An excellent compilation of such alternatives is offered in [6].

By specifying a directed alternative, great gains in the power of the testing procedures are realized.

However, one needs to be confident, based on the practicality of the problem or based on past experience that the directed alternative is appropriate. Equally important is the correct specification of the null hypothesis. In specifying a null and a directed alternative, oftentimes many parameters in the original space are ignored. Recall in our first example, $H : \mu_1 = \mu_2 = \mu_3, K_{SO} : \mu_1 \leq \mu_2 \leq \mu_3$. Here, parameter points where $\mu_1 > \mu_2 > \mu_3$ are ignored. One should be confident that those parameter points that are left out can safely be ignored. Should the ignored parameters be relevant, the advantage of using a directed alternative is considerably diminished.

References

- [1] Chacko, V.J. (1963). Testing homogeneity against ordered alternatives, *Annals of Mathematical Statistics* **34**, 945–956.
- [2] Cohen, A. & Sackrowitz, H.B. (1998). Directional tests for one-sided alternatives in multivariate models, *Annals of Statistics* **26**, 2321–2338.
- [3] Cohen, A. & Sackrowitz, H.B. (2002). Inference for the model of several treatments and a control, *Journal of Statistical Planning and Inference* **107**, 89–101.
- [4] Cohen, A. & Sackrowitz, H.B. (2003). Methods of reducing loss of efficiency due to discreteness of distribution, *Statistical Methods in Medical Research* **12**, 23–36.
- [5] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control, *Journal of the American Statistical Association* **50**, 1096–1121.
- [6] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.

(See also **Monotonic Regression**)

ARTHUR COHEN

Direction of Causation Models

NATHAN A. GILLESPIE AND NICHOLAS G. MARTIN

Volume 1, pp. 496–499

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Direction of Causation Models

In the behavioral sciences, experimental manipulation is often not an option when investigating direction of causation (DOC) and so alternative statistical approaches are needed. Longitudinal or two-wave data designs (*see Longitudinal Designs in Genetic Research*), while potentially informative, are not without their disadvantages. In addition to the cost and time required for data collection, these include stringent methodological requirements (*see* [3, 6]). When only cross-sectional data are available, a novel approach is to model direction of causation on the basis of pairs of relatives, such as twins measured on a single occasion (*see Twin Designs*) [1, 3, 6].

The pattern of cross-twin cross-trait correlations can, under certain conditions, falsify strong hypotheses about the direction of causation, provided several

assumptions are satisfied (*see* [6]). One of these is that twin pair correlations are different between target variables, which is critical, because the power to detect DOC will be greatest when the target variables have very different modes of inheritance [3].

Figure 1 provides an illustrative example of DOC modeling based on cross-sectional data. Let us assume that variable A is best explained by shared (C) and nonshared (E) environmental effects, while variable B is best explained by additive genetic (A), dominant genetic (D), and nonshared (E) environment effects (*see ACE Model*). Under the “A causes B” hypothesis (a), the cross-twin cross-trait correlation (i.e., A_{t1} to B_{t2} or A_{t2} to B_{t1}) is $c_{AS}^2 i_B$ for MZ and DZ twin pairs alike. However, under the ‘B causes A’ hypothesis (b), the cross-twin cross-trait correlation would be $(a_{BS}^2 + d_{BS}^2) i_A$ for MZ and $(1/2 a_{BS}^2 + 1/4 d_{BS}^2) i_A$ for DZ twin pairs. It is apparent that if variables A and B have identical modes of inheritance, then the cross-twin cross-trait correlations will be equivalent for MZ and DZ twin

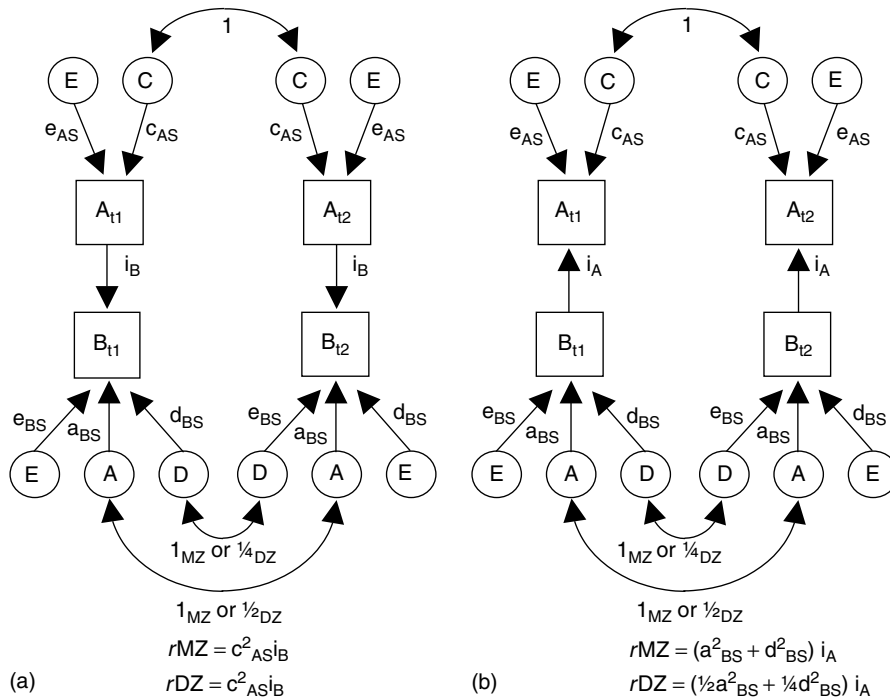


Figure 1 Unidirectional causation hypotheses between two variables A and B measured on a pair of twins. (a) Trait A causes Trait B and (b) Trait B causes Trait A. The figure also includes the expected cross-twin cross-trait correlations for MZ and DZ twins under each unidirectional hypothesis. Example based on simplified model of causes of twin pair resemblance in Neale and Cardon [5] and is also reproduced from Gillespie and colleagues [2]

2 Direction of Causation Models

pairs alike, regardless of the direction of causation, and the power to detect the direction of causation will vanish.

Neale and colleagues [7] have modeled direction of causation on the basis of the cross-sectional data between symptoms of depression and parenting, as measured by the dimensions of Care and Overprotection from the Parental Bonding Instrument [8]. They found that models that specified parental rearing as the cause of depression (parenting $\rightarrow$ depression) fitted the data significantly better than did a model that specified depression as causing

parental rearing behavior (depression $\rightarrow$ parenting). Yet, when a term for **measurement error** (omission of which is known to produce biased estimates of the causal parameters [5]) was included, the fit of the ‘parenting $\rightarrow$ depression’ model improved, but no longer explained the data as parsimoniously as a common additive genetic effects model (*see Additive Genetic Variance*) alone (i.e., implying indirect causation).

Measurement error greatly reduces the statistical **power** for resolving alternative causal hypotheses [3]. One remedy is to model DOC using multiple

Table 1 Results of fitting direction of causation models to the psychological distress and parenting variables. Reproduced from Gillespie and colleagues [2]

Model	Goodness of fit					
	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	<i>p</i>	<i>AIC</i>
Full bivariate	141.65	105				−68.35
Reciprocal causation	142.12	106	0.47	1	0.49	−69.88
Distress ^a $\rightarrow$ Parenting ^b	152.28	107	10.63	2	**	−61.72
Parenting $\rightarrow$ Distress	143.13	107	1.48	2	0.48	−70.87
No correlation	350.60	108	208.95	3	***	134.60

Results based on 944 female MZ twin pairs and 595 DZ twin pairs aged 18 to 45.

^aDistress as measured by three indicators: depression, anxiety, and somatic distress.

^bParenting as measured by three indicators: coldness, overprotection, and autonomy.

* $p < .05$, ** $p < .01$, *** $p < .001$.

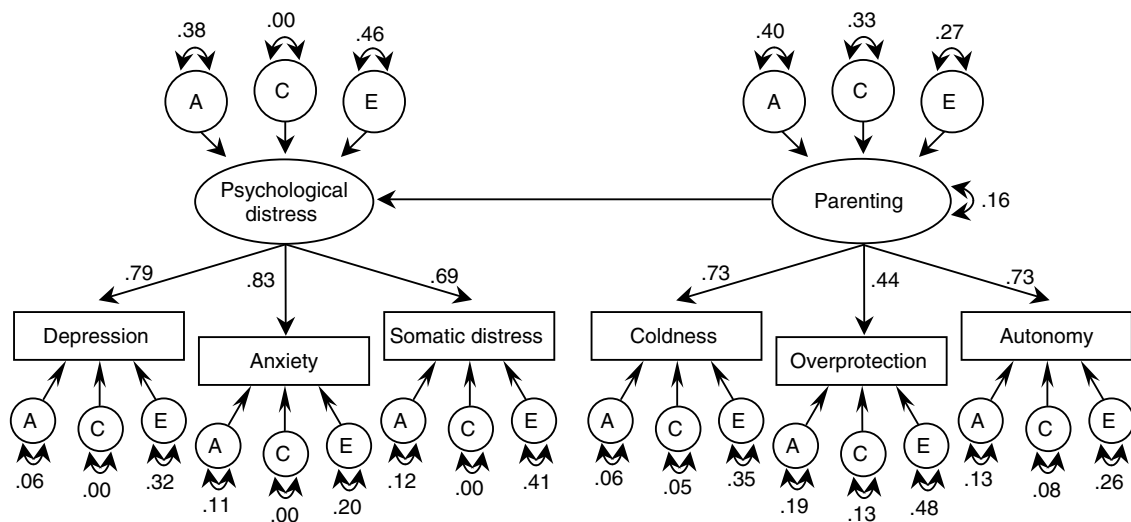


Figure 2 The best-fitting unidirectional causation model for the psychological distress and PBI parenting dimensions with standardized variance components (double-headed arrows) and standardized path coefficients. Circles represent sources of latent additive genetic (A), shared (C), and nonshared (E) environmental variance. Ellipses represent common pathways psychological distress and parenting. Reproduced from Gillespie and colleagues [2]

indicators [3–5]. This method assumes that measurement error occurs, not at the latent variable level but at the level of the indicator variables, and is uncorrelated across the indicator variables [5]. Gillespie and colleagues [2] have used this approach to model the direction of causation between multiple indicators of parenting and psychological distress. Model-fitting results are shown in Table 1.

The ‘parenting → distress’ model, as illustrated in Figure 2, provided the most parsimonious fit to the data. Unfortunately, there was insufficient statistical power to reject a full bivariate model. Therefore, it is possible that the parenting and psychological distress measures were correlated because of shared genetic or environmental effects (bivariate model), or simply arose via a reciprocal interaction between parental recollections and psychological distress. Despite this limitation, the chief advantage of this model-fitting approach is that it provides a clear means of rejecting the ‘distress → parenting’ and ‘no causation’ models, because these models deteriorated significantly from the full bivariate model. The correlations between the parenting scores and distress measures could not be explained by the hypothesis that memories of parenting were altered by symptoms of psychological distress.

Despite enthusiasm in the twin and behavior genetic communities, DOC modeling has received little attention in the psychological literature, which is a shame because it can prove exceedingly useful in illuminating the relationship between psychological constructs. However, as stressed by Duffy and Martin [1], these methods are not infallible or invariably informative, and generally require

judgment on the part of the user as to their interpretation.

References

- [1] Duffy, D.L. & Martin, N.G. (1994). Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations [see comments], *Genetic Epidemiology* **11**, 483–502.
- [2] Gillespie, N.G., Zhu, G., Neale, M.C., Heath, A.C. & Martin, N.G. (2003). Direction of causation modeling between measures of distress and parental bonding, *Behavior Genetics* **33**, 383–396.
- [3] Heath, A.C., Kessler, R.C., Neale, M.C., Hewitt, J.K., Eaves, L.J. & Kendler, K.S. (1993). Testing hypotheses about direction of causation using cross-sectional family data, *Behavior Genetics* **23**, 29–50.
- [4] McArdle, J.J. (1994). Appropriate questions about causal inference from “Direction of Causation” analyses [comment], *Genetic Epidemiology* **11**, 477–482.
- [5] Neale, M.C. & Cardon, L.R. (1992). Methodology for genetic studies of twins and families, *NATO ASI Series*, Kluwer Academic Publishers, Dordrecht.
- [6] Neale, M.C., Duffy, D.L. & Martin, N.G. (1994a). Direction of causation: reply to commentaries, *Genetic Epidemiology* **11**, 463–472.
- [7] Neale, M.C., Walters, E., Heath, A.C., Kessler, R.C., Perusse, D., Eaves, L.J. & Kendler, K.S. (1994b). Depression and parental bonding: cause, consequence, or genetic covariance? *Genetic Epidemiology* **11**, 503–522.
- [8] Parker, G., Tupling, H. & Brown, L.B. (1979). A parental bonding instrument, *British Journal of Medical Psychology* **52**, 1–10.

NATHAN A. GILLESPIE AND NICHOLAS
G. MARTIN

Discriminant Analysis

CARL J. HUBERTY

Volume 1, pp. 499–505

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Discriminant Analysis

There are many data analysis methods that involve multiple response variables, one of which is discriminant analysis. This analysis method was initiated by **Sir Ronald A. Fisher** in the 1930s in the context of classifying a plant into one of two species using four flower measurements as predictor scores. It is the context of classification/prediction that researchers in the natural sciences (e.g., genetics, biology) have associated discriminant analysis. In the behavioral sciences, however, applied researchers have typically associated discriminant analysis with the study of group differences. This latter association appears to have been initiated by methodologists at Harvard University in the 1950s. (See [1, pp. 25–26] for a little more detail on the history of discriminant analysis and *see also* **History of Discrimination and Clustering**.)

The typical design that might suggest the use of discriminant analysis involves two or more (say, k) groups of analysis units – or, subjects, such as students. A collection of two or more (say, p) response variable scores is available for each unit. One research question with such a design is: Are the k groups different with respect to means on the p outcome variables? (Or, equivalently, does the grouping variable have an effect on the collection of p outcome variables?) To answer this question, the typical analysis is **multivariate analysis of variance (MANOVA)**. Assuming that the answer to this research question is ‘yes,’ then one proceeds to describing the group differences – or, the effect(s) of the grouping variable. To address the description, one uses *descriptive discriminant analysis* (DDA). A second research question with the above design is: How well can group membership be predicted using scores on the p response variables? To answer this question, one would use *predictive discriminant analysis* (PDA). It should be noted that in DDA the response variables are *outcome* variables, while in PDA the response variables are *predictor* variables.

Description of Grouping-variable Effects

To repeat, the basic design considered here involves k groups of analysis units and p outcome variables. It is assumed at the outset that the collection of p outcome variables constitutes some type of ‘system.’

That is, the whole collection, or subsets, somehow ‘hang together’ in some substantive or theoretical sense. For an example, consider Data Set B in [1, p. 278]. The grouping variable is post-high-school education with levels: teacher college ($n_1 = 89$ students), vocational school ($n_2 = 75$), business or technical school ($n_3 = 78$), and university ($n_4 = 200$). The $N = 442$ students – the analysis units – were randomly selected from a nationwide stratified sample of nearly 26 000 eleventh-grade students. The $p = 15$ outcome variables are as follows.

Cognitive:

Literature information (LINFO),
Social Science information (SINFO),
English proficiency (EPROF),
Mathematics reasoning (MRSNG),
Visualization in three dimensions (VTDIM),
Mathematics information (MINFO),
Clerical-perceptual speed (CPSPD),

Interest:

Physical science (PSINT),
Literary-linguistic (LLINT),
Business management (BMINT),
Computation (CMINT),
Skilled trade (TRINT),

Temperament:

Sociability (SOCBL),
Impulsiveness (IMPLS),
Mature personality (MATRP).

(Whether or not these 15 outcome variables constitute a single ‘system’ is a judgment call.)

For this data set, the primary research question is: Are the $k = 4$ groups different with respect to the means on the $p = 15$ variables? To address this question, a MANOVA is conducted – assuming data conditions are satisfactory (see [6]). The MANOVA (omnibus) null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4, \quad (1)$$

(μ_j denotes the vector of 15 means for the j th population) may be tested using the Wilks lambda (Λ) criterion (*see* **Multivariate Analysis of Variance**), which is transformed to an F statistic – see [1, pp. 183–185]. For Data Set B, $\Lambda \doteq 0.5696$, $F(45, 1260.4) \doteq 5.841$, $P \doteq .0001$ and η_{adj}^2 . With these results, it is reasonable to reject H_0 , and conclude that the four group mean vectors are statistically different. Now the more interesting research question

2 Discriminant Analysis

is: Different with respect to what? (Or, on *what* does the grouping variable have an effect?) Here is where DDA comes into play.

DDA is used to determine what outcome variable *constructs* underlie the resultant group differences. The identification of the constructs is based on what are called *linear discriminant functions* (LDFs). The LDFs are linear combinations (or, composites) of the p outcome variables. (A linear combination/composite is a sum of products of variables and respective ‘weights.’) Derivation of the LDF weights is based on a mathematical method called an *eigen-analysis* (see [1, pp. 207–208]). This analysis yields numbers called **eigenvalues**. The number of eigenvalues is, in an LDF context, the minimum of p and $k - 1$, say, m . For Data Set B, $m = \min(15, 3) = 3$. With each eigenvalue is associated an **eigenvector**, numerical elements of which are the LDF weights. So, for Data Set B, there are three sets of weights (i.e., three LDFs) for the 15 outcome variables. The first LDF is defined by

$$Z_1 = 0.54Y_1 + 0.59Y_2 + \dots + 0.49Y_{15}. \quad (2)$$

The first set of LDF weights is mathematically derived so as to maximize, for the data on hand, the (canonical) correlation between Z_1 and the grouping variable (see **Canonical Correlation Analysis**). Weights for the two succeeding LDFs are determined so as to maximize, for the data on hand, the correlation between the linear combination/composite and the grouping variable that is, successively, independent of the preceding correlation.

Even though there are, for Data Set B, $m = 3$ LDFs, not all three need be retained for interpreting the resultant group differences. Determination of the ‘LDF space dimension’ may be done in two ways (see [1, pp. 211–214]). One way is to conduct three statistical tests:

- H_{01} : no separation on any dimension,
- H_{02} : separation on at most one dimension, and
- H_{03} : separation on at most two dimensions.

(Note that these are NOT hypotheses for significance of *individual* LDFs.) For Data Set B, results of the three tests are: $F_1(45, 1260.4) \doteq 5.84$, $P_1 \doteq 0.0001$; $F_2(28, 850) \doteq 1.57$, $P_2 \doteq 0.0304$ and $F_3(13, 426) \doteq 0.73$, $P_3 \doteq 0.7290$. On the basis of these results, *at most* two LDFs should be retained. The second way to address the dimension issue is a

proportion-of-variance approach. Each derived eigenvalue is a squared (canonical) correlation (between the grouping variable and the linear combination of the outcome variables), and, thus, reflects a proportion of shared variance. There is a shared-variance proportion associated with each LDF. For Data Set B, the proportions are: LDF₁, 0.849; LDF₂, 0.119; and LDF₃, 0.032. From this numerical information it may be concluded, again, that *at most* two LDFs need be retained.

Once the number of LDFs to be retained for interpretation purposes is determined, it may be helpful to get a ‘picture’ of the results. This may be accomplished by constructing an *LDF plot*. This plot is based on outcome variable means for each group on each LDF. For Data Set B and for LDF₁, the group 1 mean vector value is determined, from (1), to be

$$\bar{Z}_1 = 0.54\bar{Y}_1 + 0.59\bar{Y}_2 + \dots + 0.49\bar{Y}_{15} \doteq -0.94 \quad (3)$$

The group 1 mean vector (i.e., *centroid*) used with LDF₂ yields an approximate value of -0.21 . The two centroids for group 2, for group 3, and for group 4 are similarly calculated. The proximity of the group centroids is reflected in the LDF plot. (The typical plot used is that with the LDF axes at a right angle.) With the four groups in Data Set B, each of the four plotted points reflects the two LDF means. The two-dimensional plot for Data Set B is given in Figure 1. By projecting the centroid points, – for example, $(-0.94, -0.21)$ – onto the respective axes, one gets a general idea of group separation that may be attributed to each LDF. From Figure 1, it appears

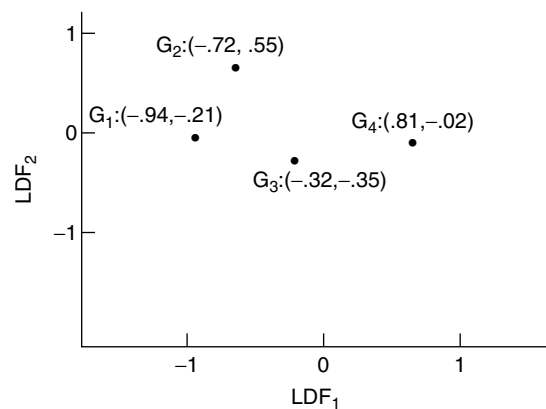


Figure 1 Plot of group centroids for Data Set B

that LDF_1 may account for separation of G_4 on one hand versus G_1 , G_2 , and G_3 on the other. Also, it may be concluded that LDF_2 may account for separation of G_2 versus G_1 , G_3 , and G_4 . The latter separation may not appear to be as clear-cut as that associated with LDF_1 .

All this said and done, the researcher proceeds to making a substantive interpretation of the two retained LDFs. To accomplish this, one determines the correlations between each of the two LDFs and each of the 15 outcome variables. Thus, there are two sets of *structure r*'s, 15 in each set (see **Canonical Correlation Analysis**). The 'high' values are given in Table 1. The *structure r*'s for LDF_1 indicate that the difference of university students versus teacher college, vocational school, and business or technical school students may be attributed to the 'construct' of skill capability in mathematics, social science, and literature, along with maturity and interest in physical science. (It is left to the reader to arrive at a more succinct and concise 'name' for the construct reflected by LDF_1 .) The second construct is a 'combination' of mathematics reasoning and 3D visualization. (It should be recognized that researcher judgment is needed to determine the number of LDFs to retain, as well as to name them.) It is these constructs that describe/explain the group differences found via the MANOVA and illustrated with the LDF plot discussed above. The constructs may, alternatively, be viewed as on what the grouping variable has an effect.

It may be of further interest to the researcher to determine an 'ordering' of the outcome variables. This interest may also be viewed as determining the 'relative importance' of the outcome variables. Now there are two ways of viewing the variable ordering/relative importance issue. One way pertains to an ordering with respect to differences between/among

the groups. This ordering may be determined by conducting *all-but-one-variable MANOVAs*. With the above example, this would result in 15 14-variable analyses. What are examined, then, are the 15 14-variable MANOVA F values, and comparing each with the F value yielded by the overall 15-variable MANOVA. The variable not in the 14-variable subset that yields the largest F -value drop, relative to the 15-variable F value, is considered the most important variable with respect to the contribution to group differences. The remaining 14 variables would be similarly ordered using, of course, some research judgment – there usually would be some rank ties. It turns out that an equivalent way to accomplish an outcome variable ordering with respect to group differences is to examine ' F -to-remove' values via the use of the SPSS DISCRIMINANT program (see **Software for Statistical Analyses**). For Data Set B, the F -to-remove values are given in Table 2 – these are the F -transformations of Wilks lambda values obtained in the 15 14-variable analyses.

A second way to order the 15 outcome variables is to examine the absolute values of the two sets of 15 *structure r*'s. Such an ordering would reflect the relative contributions of the variables to the definitions of the respective resulting constructs. From Table 1 – trusting that all other *structure r*'s are 'low' – the relative importance of the five variables used to name the first construct is

Table 1 Selected *structure r*'s for Data Set B

	LDF_1	LDF_2
MINFO	0.68	
SINFO	0.50	
LINFO	0.46	
MATRP	0.41	
PSINT	0.39	
MRSNG		0.71
VTDIM		0.58

Table 2 F -to-remove values for Data Set B

Variable	F -to-remove	Rank
TRINT	14.21	1.5
MINFO	13.04	1.5
BMINT	4.58	4
EPROF	4.50	4
LLINT	4.30	4
PSINT	3.28	7
MRSNG	3.05	7
SINFO	2.27	7
CMINT	1.93	11
MATRP	1.74	11
VTDIM	1.30	11
LINFO	1.24	11
CPSPD	1.12	11
IMPLS	0.61	14.5
SOCBL	0.09	14.5

obvious; similarly for the second construct. One could sum the squares of the two structure r 's for each of the 15 outcome variables and order the 15 variables with respect to the 15. Such an ordering would indicate the 'collective relative importance' to the definition of the pair of constructs. This latter ordering is rather generic and is judged to be of less interpretive value than the two separate orderings.

As illustrated above with Data Set B, what was discussed were the testing, construct identification, and variable ordering for the *omnibus* effects. That is, what are the effects of the grouping variable on the collection of 15 outcome variables? In some research, more specific questions are of interest. That is, there may be interest in group *contrast* effects (*see Analysis of Variance*). With Data Set B, for example, one might want to compare group 4 with any one of the other three groups or with the other three groups combined. With any contrast analysis – accomplished very simply with any computer package MANOVA program – there is only one LDF, which would be examined as above, assuming it is judged that the tested contrast effect is 'real.'

It is important to note that DDA methods are also applicable in the context of a **factorial design**, say, A-by-B. One may examine the LDF(s) for the **interaction effects**, for simple A effects, for simple B effects, or (if the interaction effects are not 'real') for main A effects and main B effects (*see Analysis of Variance*). Here, too, there may be some interest in contrast effects.

Whether a one-factor or a multiple-factor design is employed, the initial choice of the outcome variable set for a DDA is *very* important. Unless the variable set is, in some way, a substantive collection of analysis unit attributes, little, if any, meaningful interpretation can be made of the DDA results. If one has just a hodgepodge of p outcome variables, then what is suggested is that p univariate analyses be conducted – see [4].

Finally, with respect to DDA, there may be research situations in which multiple MANOVAs, along with multiple DDAs, may be conducted. Such a situation may exist when the collection of outcome variables constitutes multiple 'systems' of variables. Of course, each system would be comprised of a respectable number of outcome variables so that at least one construct might be meaningfully identified. (Should this have been considered with Data Set B?)

Group Membership Prediction

Suppose an educational researcher is interested in predicting student post-high-school experience using (a hodgepodge of) nine predictor variables. Let there be four criterion groups determined four years after ninth grade enrollment: college, vocational school, full-time job, and other. The nine predictor variable scores would be obtained prior to, or during, the ninth grade: four specific academic achievements, three family characteristics, and two survey-based attitudes. The analysis to be used with this $k = 4$, $p = 9$ design is PDA. The basic research question is: How well can the four post-high-school experiences be predicted using the nine predictors?

Another PDA example is that based on Data Set A in [1, p. 227]. The grouping variable is level of college French course – beginning ($n_1 = 35$), intermediate ($n_2 = 81$), and advanced ($n_3 = 37$). The $N = 153$ students were assessed on the following 13 variables prior to college entry:

- Five high school cumulative grade-point averages:
 - English (EGPA),
 - Mathematics (MGPA),
 - Social Science (SGPA),
 - Natural Science (NGPA),
 - French (FGPA);
- The number of semesters of high school French (SHSF);
- Four measures of academic aptitude
 - ACT English (ACTE),
 - ACT Mathematics (ACTM),
 - ACT Social Studies (ACTS),
 - ACT Natural Sciences (ACTN);
- Two scores on a French test:
 - ETS Aural Comprehension (ETSA),
 - ETS Grammar (ETSG); and
- The number of semesters since the last high school French course (SLHF).

It is assumed that French course enrollment was initially 'appropriate' – that is, the grouping variable is well defined. The purpose of the study, then, is to determine how well membership in $k = 3$ levels of college French can be predicted using scores on the $p = 13$ predictors. (Note that in PDA, the response variables are predictor variables, whereas in DDA the response variables are outcome variables. Also, in PDA, the grouping variable is an outcome

variable, whereas in DDA the grouping variable is a predictor variable.)

Assuming approximate multivariate normality (*see Multivariate Normality Tests*) of predictor variable scores in the k populations (see [1, Chs. IV & X]), there are two types of PDAs, linear and quadratic. A *linear prediction/classification rule* is appropriately used when the k group **covariance matrices** are in ‘the same ballpark.’ If so, a linear composite of the p predictors is determined for *each* of the k groups. (These linear composites are not the same as the LDFs determined in a DDA – different in number and different in derivation.) The linear combination/composite for each group is of the same general form as that in (1). For Data Set A, the first *linear classification function* (LCF) is

$$99.33 + 0.61 X_1 - 4.47 X_2 + 12.73 X_3 \\ + \cdots + 2.15 X_{13}. \quad (4)$$

The three sets of LCF weights, for this data set, are mathematically derived so as to maximize correct group classification for the data on hand. If it is to be concluded that the three population covariance matrices are not equal (see [6]), then a *quadratic prediction/classification rule* would be used. With this PDA, three quadratic classification functions (QCFs) involving the 13 predictors are derived (with the same mathematical criterion as for the linear prediction rule). The QCFs are rather complicated and lengthy, including weights for X_j , for X_j^2 and for $X_j X_{j'}$.

Whether one uses a linear rule or a quadratic rule in a group-membership prediction/classification study, there are two bases for group assignment. One basis is the *linear/quadratic composite score* for each analysis unit for each group – for each unit, there are k composite scores. A unit, then, is assigned to that group with which the larger(est) composite score is associated. The second basis is, for each unit, an estimated probability of group membership, given the unit’s vector of predictor scores; such a probability is called a *posterior probability* (*see Bayesian Statistics*). A unit, then, is assigned to that group with which the larger (or largest) posterior probability is associated. (For each unit, the sum of these probability estimates is 1.00.) The two bases will yield identical classification results.

Prior to discussing the summary of the group-membership prediction/classification results, there is another probability estimate that is *very* important

in the LCFs/QCFs and posterior probability calculations. This is a *prior* (or, *a priori*) *probability* (*see Bayesian Statistics*). The k priors reflect the relative sizes of the k populations and must sum to 1.00. The priors are included in both the composite scores and the posterior probabilities, and, thus, have an impact on group assignment. The values of the priors to be used may be based on theory, on established knowledge, or on expert judgment. The priors used for Data Set A are, respectively, 0.25, 0.50, and 0.25. (This implies that the proportion of students who enroll in, for example, the intermediate-level course is approximately 0.50.)

The calculation of LCF/QCF scores, for the data set on hand, are based on predictor weights that are determined from the very same data set (Similarly, calculation of the posterior probability estimates is based on mathematical expressions derived from the data set on hand.) In other words, the prediction rule is derived from the very data on which the rule is applied. Therefore, these group-membership prediction/classification results are (internally) biased – such a rule is considered an *internal rule*. Using an internal rule is NOT to be recommended in a PDA. Rather, an *external rule* should be employed. The external rule that I suggest is the *leave-one-out* (L-O-O) rule. The method used with the L-O-O approach involves N repetitions of the following two steps.

1. Delete one unit and derive the rule of choice (linear or quadratic) on the remaining $N-1$ units; and
2. Apply the rule of choice to the deleted unit.

(Note: At the time of this writing, the quadratic external (i.e., L-O-O) results yielded by SPSS are NOT correct. Both linear and quadratic external results are correctly yielded by the SAS package, while the SPSS package only yields correct *linear* external results.)

A basic summary of the prediction/classification results is in the form of a *classification table*. For Data Set A, the L-O-O linear classification results are given in Table 3. For this data set, the *separate-group hit rates* are $29/35 \doteq 0.83$ for group 1, $68/81 \doteq 0.84$ for Group 2, and $30/37 \doteq 0.81$ for group 3. The *total-group hit rate* is $(29 + 68 + 30)/153 \doteq 0.83$. (All four of these hit rates are inordinately ‘high.’) It is advisable, in my opinion, to assess the hit rates relative to chance. That is: Is an observed hit rate better than a hit rate that can be obtained by chance?

6 Discriminant Analysis

Table 3 Classification table for Data Set A

		Predicted group			
		1	2	3	
Actual group	1	29	6	0	35
	2	8	68	5	81
	3	0	7	30	37
		37	81	35	153

To address this question, one can use a *better-than-chance* index:

$$I = \frac{H_o - H_e}{H_e}, \quad (5)$$

where H_o is the observed hit rate, and H_e is the hit rate expected by chance. For the total-group hit rate using Data Set A, $H_o \doteq 0.83$ and $H_e \doteq (0.25 \times 35 + 0.50 \times 81 + .25 \times 37)/153 \doteq 0.38$; therefore, $I \doteq 0.72$. Thus, by using a linear external rule, about 72% fewer classification errors across all three groups would be made than if classification was done by chance. For group 3 alone, $I \doteq (0.81 - 0.25)/(1 - 0.25) \doteq 0.75$.

A researcher may want to ask two more specific PDA questions:

1. May some predictor(s) be deleted?
2. What are the more important predictors (with respect to some specific group hit rate, or to the total-group hit rate)?

The first question is a very important one – for practical purposes. (My experience has been that in virtually every PDA, at least one predictor may be discarded to result in a *better* prediction rule.) There are two reasonable analysis approaches for this question. One is the *p all-but-one-predictor analyses*. By examining the results, the predictor, when deleted, dropped the hit rate(s) of interest the least – or, which is more usual, increased the hit rate(s) of interest – could be deleted. After that is done, the $p - 1$ all-but-one-predictor analyses could be conducted. When to stop this process is a researcher judgment call. The second approach is to determine a ‘best’ subset of predictors for the hit rate(s) of interest. This may be done using an *all-subsets analysis*; that is, all subsets of size 1, of size 2, ..., of size $p - 1$. (J. D. Morris at Florida Atlantic University has written a computer program to conduct the all-subset L-O-O analyses.) Again, as

usual, judgment calls will have to be made about the retention of the final subset. For Data Set A, the best subset of the 13 predictors is comprised of six predictors (EGPA, MGPA, SGPA, NGPA, ETSA, and ETSG) with a total-group L-O-O hit rate (using the priors of 0.25, 0.50, and 0.25) of 0.88 (as compared to the total-group hit rate of 0.83 based on all 13 predictors).

The second question pertains to predictor ordering/relative importance. This may be simply addressed by conducting the *p* all-but-one-predictor analyses. The predictor, when deleted, that leads to the largest drop in the hit rates of interest, may be considered the most important one, and so on. For the $p = 13$ analyses with Data Set A, it is found that when variable 12 is deleted, the total-group hit drops the most, from 0.83 (with all 13 variables) to 0.73. Therefore, variable 12 is considered most important (with respect to the total-group hit rate). There are four variables, which when singly deleted, actually *increase* the total-group hit rate.

There are some other specific PDA-related aspects in which a researcher might have some interest. Such interest may arise when the developed prediction rule (in the form of a set of k linear or quadratic composites) is to be used with another, comparable sample. Four such aspects are listed here but will not be discussed: **outliers**, in-doubt units, nonnormal rules, and posterior probability threshold (see [1] for details).

Summary

The term ‘discriminant analysis’ may be viewed in two different ways. One, it is an analysis used to describe differences between/among groups of analysis units on the basis of scores on a ‘system’ of outcome variables. In a factorial-design context, this view would pertain to analyzing the interaction effects, main effects, and simple effects. This is when DDA is applicable. The second view is an analysis used to predict group membership on the basis of scores on a collection of predictor variables. This is when PDA is applicable. DDA and PDA are *different* analyses with *different* purposes, *different* computations, *different* interpretations, and *different* reporting [1, 5]. (With regard to the latter, I have witnessed many, many problems – see [3].) From my perspective, another view of DDA versus

PDA pertains to research context. DDA is applicable to ‘theoretical’ research questions, while PDA is applicable to ‘applied’ research questions. This relationship reminds me of an analogy involving multiple correlation and **multiple linear regression** (see [2]).

References

- [1] Huberty, C.J. (1994). *Applied Discriminant Analysis*, Wiley, New York.
- [2] Huberty, C.J. (2003). Multiple correlation versus multiple regression, *Educational and Psychological Measurement* **63**, 271–278.
- [3] Huberty, C.J. & Hussein, M.H. (2001). Some problems in reporting use of discriminant analyses, *Journal of Experimental Education* **71**, 177–191.
- [4] Huberty, C.J. & Morris, J.D. (1989). Multivariate analysis versus multiple univariate analyses, *Psychological Bulletin* **105**, 302–308.
- [5] Huberty, C.J. & Olejnik, S.O. (in preparation). *Applied MANOVA and Discriminant Analysis*, 2nd Edition, Wiley, New York.
- [6] Huberty, C.J. & Petoskey, M.D. (2000). Multivariate analysis of variance and covariance, in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, H.E.A. Tinsley & S.D. Brown, eds, Academic Press, New York, pp. 183–208.

(See also **Hierarchical Clustering; k-means Analysis**)

CARL J. HUBERTY

Distribution-free Inference, an Overview

CLIFFORD E. LUNNEBORG

Volume 1, pp. 506–513

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Distribution-free Inference, an Overview

Introduction

Nearly fifty years have passed since the 1956 publication of social psychologist Sidney Siegel's *Nonparametric Statistics for the Behavioral Sciences* [23] provided convenient access to a range of nonparametric or distribution-free procedures for researchers in psychology and other disciplines. The first edition of his book still actively circulates in no fewer than eight branches of my university's library, including the specialized collections in architecture, business administration, and marine biology. What was Siegel's motivation? What techniques did he provide the researchers of the 1950s? How has the nonparametric landscape evolved in recent years?

Importance of Level of Measurement

Siegel was motivated primarily by the mismatch between parametric statistics and the level of measurement (*see Scales of Measurement*) associated with many psychological scales. The *t* Test (*see Catalogue of Parametric Tests*) and the *analysis of variance* require scales that are not only numeric but at the interval or ratio level. Here are some representative exceptions:

- Potential car buyers report that they intend to purchase one of the following: a SUV, a minivan, a sedan, or a sports car. The four possible responses are nonnumeric and have no natural order. They form a set of nominal categories.
- Students rate their instructors on a five-point scale, ranging from poor through fair, good, and very good to excellent. The five possibilities form an ordinal scale of instructor performance. We could replace the verbal labels with numbers, but how big is the difference between fair and good compared with the difference between very good and excellent?
- The difference in intelligence between two 12-year old children with WISC IQ scores of 85 and 95 may or may not correspond to the difference in intelligence between another pair of twelve-year

olds with IQ scores of 135 and 145. Although the WISC scores are numeric, a researcher may want to treat the scores as providing a ranking or numeric ordering of the intelligences of tested children rather than as measuring those intelligences on an interval or, were zero IQ to be well-defined, a ratio scale.

The centrality to distribution-free statistics of scales of measurement is reinforced in a more recent standard text, the 1999 revision of Conover's *Practical Nonparametric Statistics* [4]. Conover offers this definition of nonparametric: 'A statistical method is nonparametric if it satisfies at least one of these criteria:

1. The method may be used on data with a nominal scale of measurement.
2. The method may be used on data with an ordinal scale of measurement.
3. The method may be used on data with an interval or ratio scale of measurement, where the distribution function is either unspecified or specified except for an infinite number of parameters.'

The last criterion includes, among others, instances in which (a) the normality of a distribution is in question, (b) that distribution is contaminated by **outliers**, or (c) observations are not identically distributed but are drawn from multiple parameterized distributions, perhaps as many distributions as there are observations.

Basic Distribution-free Techniques

Both [4] and [23] employ a 'factorial design' in presenting what I will call the 'basic' nonparametric techniques, crossing levels of measurement with a common set of research designs. Notably, the fly leaves of the two texts feature the resulting two-way table. I have utilized this structure in Table 1 to provide an overview of the basic distribution-free techniques and to show the extent of overlap between the 1956 and 1999 coverages. It should be noted that each of the two texts does mention techniques other than the ones listed here. Table 1 is limited to those statistics elevated to inclusion in the fly leaf table by one or other of the two authors.

In an important sense, the heart of classical distribution-free statistics is contained in footnote c

2 Distribution-free Inference, an Overview

Table 1 Basic distribution-free techniques, [4] and [23]^a

	Nominal	Ordinal ^b	Interval ^c
One sample	Binomial Test Chi-squared Test <i>CI for p</i>	(Runs Test) Kolmogorov–Smirnov <i>Quantile Test</i> <i>Cox–Stuart Test</i> <i>Daniels Test</i> <i>CI for Quantile</i>	<i>Wilcoxon Test</i> <i>Lilliefors</i> <i>Shapiro–Wilk</i> <i>CI for Median</i>
Two related samples	McNemar Change <i>Fisher Exact Test</i> <i>Chi-squared Test</i> <i>CI for p</i>	Sign Test <i>Quantile Test</i> <i>CI for Quantile</i>	(Walsh Test) Permutation Test Wilcoxon Test ^d <i>Normal Scores</i> <i>CI for mdn diff</i>
Two independent samples	(Fisher Exact Test) Chi-squared Test <i>CI for $p_1 - p_2$</i>	(Median Test) (Wald–Wolfowitz) (Moses Extremes) Mann–Whitney Kolmogorov–Smirnov <i>Cramer-von Mises</i> <i>Mann–Whitney CI</i>	Permutation Test <i>Squared Ranks</i> ^e <i>Klotz Test</i> ^e
<i>k</i> related samples	Cochran Q	Friedman Two-way <i>Page Test</i>	<i>Quade Test</i>
<i>k</i> independent samples	Chi-squared Test <i>Mantel–Haenszel</i>	Median Test Kruskal–Wallis <i>Jonckheere–Terpstra</i>	<i>Normal Scores</i> ^e <i>Squared Ranks</i> ^e
Correlation and regression	Contingency coeff, C <i>Alternative contingency coeffs</i> <i>Phi Coefficient</i>	Spearman ρ Kendall τ Kendall concordance, W	<i>Slope Test</i> <i>Monotone regression</i> ^e

^aParenthesized entries feature only in [23] and italicized entries feature only in [4].

^bTechniques identified with the Nominal level can be used with Ordinal measures as well.

^cTechniques identified with either the Nominal or Ordinal levels can be used with Interval measures as well.

^dActually classified as Ordinal in [23].

^eActually classified as Ordinal in [4].

to Table 1. Measures that are numeric, apparently on an interval scale, can be treated as ordinal data. We simply replace the numeric scores with their ranks. I will describe, however, the whole of Table 1 column by column.

In the first column of the table, data are categorical and the *Chi-squared Test*, introduced in a first statistics course, plays an important role (see **Contingency Tables**). In one-sample designs, it provides for testing a set of observed frequencies against a set of theoretical expectations (e.g., the proportions of blood types A, B, AB, and O) and in two-sample designs it provides for testing the equivalence of two distributions of a common set of categories. An accurate probability of incorrectly rejecting the null hypothesis is assured only asymptotically and is compromised where parameters are estimated in the first case.

The *Fisher Exact Test* (see **Exact Methods for Categorical Data**) provides, as its name implies, an exact test of the independence of row and column classifications in a 2×2 table of frequencies (see **Two by Two Contingency Tables**). Typically, the rows correspond to two treatments and the columns to two outcomes of treatment, so the test of independence is a test of the equality of the proportion of successes in the two treatment populations. Under the null hypothesis, the cell frequencies are regulated by a hypergeometric distribution (see **Catalogue of Probability Density Functions**) and that distribution allows the computation of an exact *P* value for either directional or nondirectional alternatives to independence. The chi-squared test is used frequently in this situation but yields only approximate probabilities and does not provide for directionality in the alternative hypothesis.

The **Binomial Test** invokes the family of binomial distributions (see **Catalogue of Probability Density Functions**) to test a hypothesis about the proportion of ‘successes’ in a distribution of successes and failures. One difference between [4] and [23] is that the former emphasizes hypothesis tests while the later reflects the more recent interest in the estimation of **confidence intervals** (CIs) for parameters such as the proportion of successes or, in the two-sample case, the difference in the proportions of success (see **Confidence Intervals: Nonparametric**).

The **Maentel–Haenszel Test** extends Fisher’s exact test to studies in which the two treatments have been evaluated, independently, in two or more populations. The null hypothesis is that the two treatments are equally successful in all populations. The alternative may be directional (e.g., that treatment ‘A’ will be superior to treatment ‘B’ in at least some populations and equivalent in the others) or nondirectional. *Cochran’s Q* is useful as an omnibus test of treatments in **randomized complete block designs** where the response to treatment is either a success or failure. As *Q* is a transformation of the usual Pearson chi-squared statistic, the suggested method for finding a *P* value is to refer the statistic to a chi-squared distribution. As noted earlier, the result is only approximately correct. Subsequent pairwise comparisons can be made with *McNemar’s Test for Significance of Change* (see **Matching**). The latter procedure can be used as well in matched pair designs for two treatments with dichotomous outcomes. Although significance for McNemar’s test usually is approximated by reference to a chi-squared distribution, Fisher’s exact test could be used to good effect.

The *Phi Coefficient* (see **Effect Size Measures**) expresses the association between two dichotomous classifications of a set of cases on a scale not unlike the usual correlation coefficient. As it is based on a 2×2 table of frequencies, significance can be assessed via Fisher’s exact test or approximated by reference to a chi-squared distribution. The several contingency coefficients that have been proposed transform the chi-squared statistic for a two-way table to a scale ranging from 0 for independence to some positive constant, sometimes 1, for perfect dependence. The underlying chi-squared statistic provides a basis for approximating a test of significance of the null hypothesis of independence.

The second column of Table 1 lists techniques requiring observations that can be ordered from smallest to largest, allowing for ties. The one sample **Runs Test** in [23] evaluates the randomness of a sequence of occurrences of two equally likely events (such as, heads and tails in the flips of a coin). The examples in [23] are based on numeric data, but as these are grouped into two events, the test could as well have appeared in the Nominal column. Exact *P* values are provided by the appropriate Binomial random variable with $p = 0.5$. Runs tests are deprecated in [4] as having less power than alternative tests.

The **Kolmogorov–Smirnov test** compares two cumulative distribution functions. The one-sample version compares an empirical distribution function with a theoretical one. The two-sample version compares two empirical distribution functions. The **Cramer–von Mises Test** is a variation on the two-sample Kolmogorov–Smirnov, again comparing two empirical cumulative distribution functions. *P* values for both two-sample tests can be obtained by permuting observations between the two sources.

The *Quantile Test* uses the properties of a Binomial distribution to test hypotheses about (or find CIs for) **quantiles**, such as the median or the 75th percentile, of a distribution. The Cox–Stuart Test groups a sequence of scores into pairs and then applies the **Sign Test** to the signs (positive or negative) of the pairwise differences to detect a trend in the data. The Sign Test itself is noted in [23] as the oldest of all distribution-free tests. The null hypothesis is that the two signs, + and –, have equal probability of occurring and the binomial random variable with $p = 0.5$ is used to test for significance. The *Daniels Test for Trend* is an alternative to the Cox–Stuart test. It uses **Spearman’s rho**, computed between the ranks of a set of observations and the order in which those observations were collected to assess trend. Below I mention how rho can be tested for significance.

The **Wilcoxon–Mann–Whitney Test** (WMW, it has two origins) has become the distribution-free rival to the *t* Test for comparing the magnitudes of scores in two-sampled distributions. The observations in the two samples are pooled and then ranked from smallest to largest. The test statistic is the sum of ranks for one of the samples and significance is evaluated by comparing this rank sum with those computed from all possible permutations of the ranks

between treatments. The rank sums can be used as well to find a CI for the difference in medians [4].

The two-sample **Median Test** evaluates the same hypothesis as the WMW by applying Fisher's exact test to the fourfold table created by noting the number of observations in each sample that are larger than or are smaller than the median for the combined samples. The *Wald–Wolfowitz Runs Test* approaches the question of whether the two samples were drawn from identical distributions by ordering the combined samples, smallest to largest, and counting the number of runs of the two sources in the resulting sequence. A Binomial random variable provides the reference distribution for testing significance.

The *Moses Test of Extreme Reactions* is tailored to a particular alternative hypothesis, that the 'active' treatment will produce extreme reactions, responses that are either very negative (small) or very positive (large). The combined samples are ranked as for the runs test and the test statistic is the span in ranks of the 'control' sample. Exact significance can be assessed by referring this span to the distribution of spans computed over all possible permutations of the ranks between the active and control treatments.

Just as the WMW test is the principal distribution-free alternative to the t Test, **Friedman's Test** is the nonparametric choice as an omnibus treatment test in the complete randomized block design (see **Randomized Block Design: Nonparametric Analyses**). Responses to treatment are ranked within each block and these ranks summed over blocks, separately for each treatment. The test statistic is the variance of these sums of treatment ranks. Exact significance is evaluated by comparing this variance with the distribution of variances resulting from all possible combinations of permutations of the ranks within blocks.

Friedman's test is an omnibus one. Where the alternative hypothesis specifies the order of effectiveness of the treatments, the *Page Test* can be used. The test statistic is the Spearman rank correlation between this order and the rank of treatment sums computed as for Friedman's test. Exact significance is assessed by comparing this correlation with those resulting, again, from all possible combinations of permutations of the ranks within blocks (see **Page's Ordered Alternatives Test**).

The k -sample *Median Test* tests the null hypothesis that the k -samples are drawn from populations

with a common median. The test statistic is the Pearson chi-squared statistic computed from the $2 \times k$ table of frequencies that results from counting the number of observations in each of the samples that are either smaller than or greater than the median of the combined samples. Approximate P values are provided by referring the statistic to the appropriate chi-squared distribution.

The **Kruskal–Wallis Test** is an extension of the WMW test to k independent samples. All observations are pooled and a rank assigned to each. These ranks are then summed separately for each sample. The test statistic is the variance of these rank sums. Exact significance is assessed by comparing this variance against a reference distribution made up of similar variances computed from all possible permutations of the ranks among the treatments.

The Kruskal–Wallis test is an omnibus test for equivalence of k treatments. By contrast, the **Jonckheere–Terpstra Test** has as its alternative hypothesis an ordering of expected treatment effectiveness. The test statistic can be evaluated for significance exactly by referring it to a distribution of similar values computed over all possible permutations of the ranks of the observations among treatments. In practice, the test statistic is computed on the basis of the raw observations, but as the computation is sensitive only to the ordering of these observations, ranks could be used to the same result.

Two distribution-free measures of association between a pair of measured attributes, **Spearman's Rho** and **Kendall's Tau**, are well known in the psychological literature. The first is simply the product–moment correlation computed between the two sets of ranks. Tau, however, is based on an assessment of the concordance or not of each of the $[n \times (n - 1)]/2$ pairs of bivariate observations. Though computed from the raw observations, the same value of τ would result if ranks were used instead. Significance of either can be assessed by comparing the statistic against those associated with all possible permutations of the Y scores (or their ranks) paired with the X scores (or their ranks).

To assess the degree of agreement among b raters, when assessing (or, ordering) k stimulus objects, **Kendall's Coefficient of Concordance** has been employed. Although W has a different computation, it is a monotonic function of the statistic used in Friedman's test for balanced designs with b blocks

and k treatments and can be similarly evaluated for significance.

The final column of Table 1 lists techniques that, arguably, require observations on an interval scale of measurement. There is some disagreement on the correct placement among [4, 23], and myself. I'll deal first, and quite briefly, with six techniques that quite clearly require interval measurement.

The *Lilliefors* and *Shapiro–Wilks* procedures are used primarily as tests of normality. Given a set of measurements on an interval scale, should we reject the hypothesis that it is a sample from a normal random variable? The *Squared Ranks* and *Klotz* tests are distribution-free tests of the equivalence of variances in two-sampled distributions. Variance implies measurement on an interval scale. While the **Slope Test** and CI Estimate employ Spearman's Rho and Kendall's Tau, the existence of a regression slope implies interval measurement. Similarly, *Monotonic Regression* uses ranks to estimate a regression curve, again defined only for interval measurement. The regression curve, whether linear or not, tracks the dependence of the mean of Y on the value of X . Means require interval measurement.

Usually, **Wilcoxon's Signed Ranks Test** is presented as an improvement on the Sign test, an Ordinal procedure. While the latter takes only the signs of a set of differences into account, Wilcoxon's procedure attaches those signs to the ranks of the absolute values of the differences. Under the null hypothesis, the difference in the sums of positive and negative ranks ought to be close to zero. An exact test is based on tabulating these sums for all of the 2^n possible assignments of signs to the ranks. Although only ranks are used in the statistic, our ability to rank differences, either differences between paired observations or differences between observations and a hypothesized median, depends upon an interval scale for the original observations. Wilcoxon's procedure can be used to estimate CIs for the median or median difference.

The *Walsh Test* is similar in purpose to the Signed Rank test but uses signed differences, actually pairwise averages of signed differences, rather than signed ranks of differences. It is a small sample procedure; [23] tables significant values only for sample sizes no larger than 15. The complete set of $[n \times (n - 1)]/2$ pairwise averages, known as **Walsh Averages** can be used to estimate the median

or median difference and to find a CI for that parameter.

The *Quade Test* extends Friedman's test by differentially weighting the contribution of each of the blocks. The weights are given by the ranks of the ranges of the raw observations in the block. the use of the range places this test in the Interval, rather than Ordinal column, of my Table 1. Quade's test statistic does not have a tractable exact distribution, so an approximation is used based on the parametric family of F random variables. It appears problematic whether this test does improve on Friedman's approach, which can be used as well with ordinal measures.

Both [4] and [23] list the *Permutation Test* (see **Permutation Based Inference**) as a distribution-free procedure for interval observations, though not under that name. It is referred to as the Randomization Test by [23] and as Fisher's Method of Randomization by [4]. I refer to it as the permutation test or, more explicitly, as the Raw Score Permutation Test and reserve the name Randomization Test for a related, but distinct, inferential technique. Incidentally, the listing of the permutation test in the Interval column of Table 1 for two independent samples has, perhaps, more to do with the hypothesis most often associated with the test than with the logic of the test. The null hypothesis is that the two samples are drawn from identical populations. Testing hypotheses about population means implies interval measurement. Testing hypotheses about population medians, on the other hand, may require only ordinal measurement.

We have already encountered important distribution-free procedures, including the Wilcoxon–Mann–Whitney and Kruskal–Wallis tests, for which significance can be assessed exactly by systematically permuting the ranks of observations among treatments. These tests can be thought of as Rank Permutation Tests. Raw score permutation test P values are obtained via the same route; we refer a test statistic computed from raw scores to a reference distribution made up of values of that statistic computed for all possible permutations of the raw scores among treatments.

I have identified a third class of permutation tests in the Interval column of Table 1 as *Normal Scores* tests (see **Normal Scores and Expected Order Statistics**). Briefly, these are permutation tests that are carried out after the raw scores have been replaced, not by their ranks, but by scores that

inherit their magnitudes from the standard Normal random variable while preserving the order of the observations. The ‘gaps’ between these normal scores will vary, unlike the constant unit difference between adjacent ranks. There are several ways of finding such normal scores. The Normal Scores Permutation Test is referred to as the *van der Waerden Test* by [4]. This name derives from the use as normal scores of quantiles of the Standard Normal random variable (mean of zero, variance of one). In particular, the k th of n ranked scores is transformed to the $q(k) = [k/(n + 1)]$ quantile of the Standard Normal, for example, for $k = 3, n = 10, q(k) = 3/11$. The corresponding van der Waerden score is that z score below which fall 3/11 of the distribution of the standard normal, that is, $z = -0.60$.

Normal scores tests have appealing power properties [4] although this can be offset somewhat by a loss in accuracy if normal theory approximations, rather than actual permutation reference distributions, are employed for hypothesis testing.

Had [23] and, for that matter, [4] solely described a set of distribution-free tests, the impact would have been minimal. What made the techniques valuable to researchers was the provision of tables of significant values for the tests. There are no fewer than 21 tables in [23] and 22 in [4]. These tables enable exact inference for smaller samples and facilitate the use of normal theory approximations for larger studies.

In addition to [4], other very good recent guides to these techniques include [13, 14, 19] and [25].

Growth of Distribution-free Inference

The growth of distribution-free inference beyond the techniques already surveyed has been considerable. Most of this growth has been facilitated, if not stimulated, by the almost universal availability of inexpensive, fast computing. These are some highlights.

The analysis of frequencies, tabulated by two or more sets of nominal categories, now extends far beyond the chi-squared test of independence thanks to the development [3] and subsequent popularization [2] of **Log Linear Models**. The flavor of these analyses is not unlike that of the analysis of factorial designs for measured data; what higher order interactions are needed to account for the data?

A graphical descriptive technique for cross-classified frequencies, as yet more popular with

French than with English or US researchers, is **Correspondence Analysis** [12, 15]. In its simplest form, the correspondence referred to is that between row and column categories. In effect, correspondence analysis decomposes the chi-squared lack of fit of a model to the observed frequencies into a number of, often interpretable, components.

Regression models for binary responses, known as **Logistic Regression**, pioneered by [5] now see wide usage [2]. As with linear regression, the regressors may be a mix of measured and categorical variables. Unlike linear regression, the estimation of model parameters must be carried out iteratively, as also is true for the fitting of many log linear models. Thus, the adoption of these techniques has required additional computational support. Though originally developed for two-level responses, logistic regression has been extended to cover multicategory responses, either nominal or ordinal [1, 2].

Researchers today have access to measures of association for cross-classified frequencies, such as the *Goodman–Kruskal Gamma* and *Tau* coefficients [11], that are much more informative than contingency coefficients.

Earlier, I noted that both [4] and [23] include (raw score) Permutation Tests among their distribution-free techniques. Though they predate most other distribution-free tests [22], their need for considerable computational support retarded their wide acceptance. Now that sufficient computing power is available, there is an awakening of interest in permutation inference, [10, 18] and [24], and the range of hypotheses that can be tested is expanding [20, 21].

Important to the use of permutation tests has been the realization that it is not necessary to survey all of the possible permutations of scores among treatments. Even with modern computing power, it remains a challenge to enumerate all the possible permutations when, for example, there are 16 observations in each of two samples: $[32!/(16! \times 16!)] = 601,080,390$. A Monte Carlo test (*see Monte Carlo Goodness of Fit Tests; Monte Carlo Simulation*) based on a reference distribution made up of the observed test statistic plus those resulting from an additional $(R - 1)$ randomly chosen permutations also provides an exact significance test [8, 18]. The power of this test increases with R , but with modern desktop computing power, an R of 10 000 or even larger is a quite realistic choice.

In his seminal series of papers, Pitman [22] noted that permutation tests were valid even where the samples ‘exhausted’ the population sampled. Though the terminology may seem odd, the situation described is both common and very critical. Consider the following.

A psychologist advertises, among undergraduates, for volunteers to participate in a study of visual perception. The researcher determines that 48 of the volunteers are qualified for the study and randomly divides those students into two treatment groups, a Low Illumination Level group and a High Illumination Level group. Notably, the 48 students are not a random sample from any larger population; they are a set of available cases. However, the two randomly formed treatment groups are random samples from that set and, of course, together they ‘exhaust’ that set. This is the situation to which Pitman referred. The set of available cases constitutes what I call a local population.

Parametric inference, for example, a t Test, assumes the 48 students to be randomly chosen from an essentially infinitely large population and is clearly inappropriate in this setting [8, 16, 17]. The permutation test mechanics, however, provide a valid test for the local population and Edgington [8] advocates, as do I, the use of the term *Randomization Test* when used in this situation (see **Randomization Based Tests**). The distinctive term serves to emphasize that the inference (a) is driven by the randomization rather than by random sampling and (b) that inference is limited to the local population rather than some infinitely large one.

Truly random samples remain a rarity in the behavioral sciences. Randomization, however, is a well-established experimental precaution and randomization tests ought to be more widely used than they have been [8, 16, 17]. In the preface to [4], the author notes that, in 1999, distribution-free methods are ‘essential tools’ for researchers doing statistical analyses. The authors of [14] go even further, declaring distribution-free tests to be the ‘preferred methodology’ for data analysts. There is some evidence, however, that psychologists may be reluctant to give up parametric techniques; in 1994, de Leuw [7] noted that there remained ‘analysis of variance oriented programs in psychology departments’.

Arguably, applications of the **Bootstrap** have had the greatest recent impact on distribution-free inference. The bootstrap provides a basis for estimating

standard errors and confidence intervals and for carrying out hypothesis tests on the basis of samples drawn by resampling from an initial random sample. The approach is computer intensive but has wide applications [6, 9, 17, 18, 24].

Fast computing has changed the statistical landscape forever. Parametric methods thrived, in large part, because their mathematics led to easy, albeit approximate and inaccurate, computations. That crutch is no longer needed.

References

- [1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, Wiley, New York.
- [2] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.
- [3] Bishop, Y.V.V., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*, MIT Press, Cambridge.
- [4] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [5] Cox, D.R. (1970). *The Analysis of Binary Data*, Chapman & Hall, London.
- [6] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press, Cambridge.
- [7] de Leuw, J. (1994). Changes in *JES*, *Journal of Educational and Behavioral Statistics* **19**, 169–170.
- [8] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [9] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [10] Good, P. (2000). *Permutation Tests*, 2nd Edition, Springer, New York.
- [11] Goodman, L.A. & Kruskal, W.H. (1979). *Measures of Association for Cross Classifications*, Springer, New York.
- [12] Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, London.
- [13] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*, Wiley, New York.
- [14] Hollander, M. & Wolfe, C.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.
- [15] Lebart, L., Morineau, A. & Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*, Wiley, New York.
- [16] Ludbrook, J. & Dudley, H.A.F. (1998). Why permutation tests are superior to t - and F -tests in biomedical research, *The American Statistician* **52**, 127–132.
- [17] Lunneborg, C.E. (2000). *Data Analysis by Resampling*, Duxbury, Pacific Grove.
- [18] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.
- [19] Maritz, J.S. (1984). *Distribution-free Statistical Methods*, Chapman & Hall, London.

8 Distribution-free Inference, an Overview

- [20] Mielke, P.W. & Berry, K.J. (2001). *Permutation Methods: A Distance Function Approach*, Springer, New York.
- [21] Pesarin, F. (2001). *Multivariate Permutation Tests*, Wiley, Chichester.
- [22] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, *Journal of the Royal Statistical Society B* **4**, 119–130.
- [23] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, New York.
- [24] Sprent, P. (1998). *Data Driven Statistical Methods*, Chapman & Hall, London.
- [25] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Edition, Chapman & Hall/CRC, London.

CLIFFORD E. LUNNEBORG

Dominance

DAVID M. EVANS

Volume 1, pp. 513–514

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Dominance

Dominance refers to the nonlinear interaction between alleles *within* a locus. In the case of a discrete trait, when the effect of one allele masks the effect of the other allele at a single locus, we say that the first allele exhibits dominance over the second allele. In the case of a quantitative trait, dominance is best illustrated in reference to the standard biometrical model. Consider a single autosomal biallelic locus (Figure 1). Let the genotypic value of the homozygote A_1A_1 be $+a$ and the genotypic value of the homozygote A_2A_2 be $-a$. The genotypic value of the heterozygote A_1A_2 depends upon the degree of dominance at the locus and is quantified by the parameter d . When there is no dominance ($d = 0$), then alleles A_1 and A_2 are said to act additively in that the genotypic value of the heterozygote is exactly half the sum of the genotypic values of the two homozygotes. When $d > 0$, allele A_1 displays dominance over allele A_2 . Conversely, when $d < 0$, allele A_2 displays dominance over allele A_1 . When dominance is complete, d is equal to $+a$ or $-a$ [3–5]. Note that the concept of dominance rests critically on the choice of scale used to measure the trait of interest, in that a trait may exhibit dominance when measured on one scale, but not when the trait is measured on a different transformed (e.g., logarithmic) scale [2, 9].

If one regresses the number of copies of an allele (say A_1) against genotypic value, it is possible to partition the genotypic value into an expected value based on additivity at the locus, and a deviation based on dominance. The proportion of variance in the genotypic value explained by the regression is the additive genetic variance (*see Additive Genetic Variance*). The residual variation, which is not explained by the regression, is referred to as

the *dominance genetic variance* and arises because of the nonlinear interaction between alleles at the same locus.

In the case of the biallelic locus above, the dominance (σ_D^2) variance is given by the formula:

$$\sigma_D^2 = (2pqd)^2 \quad (1)$$

Thus, the dominance variance is a function of both the allele frequencies in the population and the dominance parameter d . Note that a low proportion of dominance variance does not necessarily imply the absence of dominant gene action, but rather may be a consequence of the particular allele frequencies in the population [9].

The classical **twin design** which compares the similarity between monozygotic and dizygotic twins reared together may be used to estimate the amount of dominance variance contributing to a trait, although the power to do so is quite low [1, 8]. However, it is important to realize that it is not possible to estimate dominant genetic and shared environmental components of variance simultaneously using this design. This is because both these variance components are negatively confounded in a study of twins reared together. That is not to say that these components cannot contribute simultaneously to the trait variance, but rather they cannot be estimated from data on twins alone [6, 8]. When the correlation between monozygotic twins is greater than half the correlation between dizygotic twins, it is assumed that shared environmental factors do not influence the trait, and a dominance genetic component is estimated. In contrast, when the correlation between monozygotic twins is less than half the correlation between dizygotic twins, it is assumed that dominance genetic factors do not influence the trait and a shared environmental variance component is estimated. The consequence of this confounding is

Genotype	A_2A_2		A_1A_2	A_1A_1
Genotypic value	$-a$	0	d	$+a$
Genotypic frequency	q^2		$2pq$	p^2

Figure 1 A biallelic autosomal locus in Hardy–Weinberg equilibrium. The genotypic values of the homozygotes A_1A_1 and A_2A_2 are $+a$ and $-a$ respectively. The genotypic value of the heterozygote A_1A_2 is d , which quantifies the degree of dominance at the locus. The gene frequencies of alleles A_1 and A_2 are p and q respectively, and the frequencies of the genotypes are as shown

2 Dominance

that variance component estimates will be biased when dominance genetic and shared environmental components simultaneously contribute to trait variation [6–8].

References

- [1] Eaves, L.J. (1988). Dominance alone is not enough, *Behaviour Genetics* **18**, 27–33.
- [2] Eaves, L.J., Last, K., Martin, N.G. & Jinks, J.L. (1977). A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences, *The British Journal of Mathematical and Statistical Psychology* **30**, 1–42.
- [3] Evans, D.M., Gillespie, N.G. & Martin, N.G. (2002). Biometrical genetics, *Biological Psychology* **61**, 33–51.
- [4] Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, Longman, Burnt Mill.
- [5] Fisher, R.A. (1918). The correlation between relatives on the supposition of mendelian inheritance, *Transaction of the Royal Society of Edinburgh* **52**, 399–433.
- [6] Grayson, D.A. (1989). Twins reared together: minimizing shared environmental effects, *Behavior Genetics* **19**, 593–604.
- [7] Hewitt, J.K. (1989). Of biases and more in the study of twins reared together: a reply to Grayson, *Behavior Genetics* **19**, 605–608.
- [8] Martin, N.G., Eaves, L.J., Kearsley, M.J. & Davies, P. (1978). The power of the classical twin study, *Heredity* **40**, 97–116.
- [9] Mather, K. & Jinks, J.L. (1982). *Biometrical Genetics*, Chapman & Hall, New York.

DAVID M. EVANS

Dot Chart

BRIAN S. EVERITT

Volume 1, pp. 514–515

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Dot Chart

Many data sets consist of measurements on some continuous variable of interest recorded within the categories of a particular categorical variable. A very simple example would be height measurements for a sample of men and a sample of women. The dot chart, in which the position of a dot along a horizontal line indicates the value of the continuous measurement made within each of the categories involved, is often a useful graphic for making comparisons and identifying possible ‘outlying’ categories. An example of a dot chart is shown in Figure 1. The plot represents standardized mortality rates for lung cancer in 25 occupational groups; to enhance the usefulness of the graphic, the categories are ordered according to their mortality rates.

A dot chart is generally far more effective in communicating the pattern in the data than a **pie chart** or a **bar chart**.

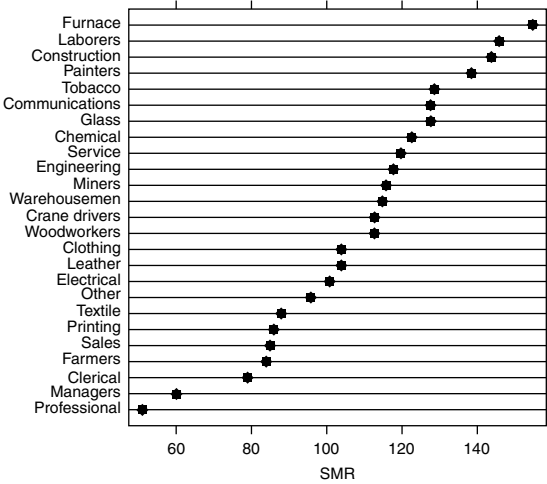


Figure 1 Dot chart of standardized mortality rates for lung cancer in 25 occupational groups

BRIAN S. EVERITT

Dropouts in Longitudinal Data

EDITH D. DE LEEUW

Volume 1, pp. 515–518

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Dropouts in Longitudinal Data

In longitudinal studies, research units (e.g., households, individual persons, establishments) are measured repeatedly over time (*see* **Longitudinal Data Analysis; Repeated Measures Analysis of Variance**). Usually, a limited number of separate measurement occasions or waves is used. The minimum number of waves is two, as in the classical pretest–posttest designs, that are well-known in intervention studies and experiments (*see* **Clinical Trials and Intervention Studies**). But, longitudinal studies can have any number of measurement occasions (waves) in time. If the number of occasions is very large this is called a **time series**. In a time series, a small number of research units is followed through time and measured on many different occasions on a few variables only. Examples of time series can be found in psychological studies, educational research, econometrics, and medicine. In social research and official statistics, a common form of longitudinal study is the panel survey. In a panel, a well-defined set of participants is surveyed repeatedly. In contrast to time series, panel surveys use a large number of research units and a large number of variables, while the number of time points is limited. Examples are budget surveys, election studies, socioeconomic panels, and general household panels (*see* **Panel Study**). In the following sections, most examples will come from panel surveys and survey methodology. However, the principles discussed also apply to other types of longitudinal studies and other disciplines.

The validity of *any* longitudinal study can be threatened by dropout (*see* **Dropouts in Longitudinal Studies: Methods of Analysis**). If the dropout is *selective*, if the missing data are not missing randomly, than the results may be biased. For instance, if in a panel of elderly, the oldest members and those in ill-health drop out more often, or if in a clinical trial for premature infants, the lightest infants are more likely to stay in the intervention group, while the more healthy, heavier babies drop out over time. When one knows who the dropouts are and why the dropout occurs, one can statistically adjust for dropout (*see* **Dropouts in Longitudinal Studies: Methods of Analysis; Missing Data**). But this is far

from simple, and the more one knows about the missing data, the better one can adjust. So, the first step in good adjustment is to prevent dropout as much as possible, and collect as much data as possible of people who may eventually drop out. But even if the dropout is not selective, even if people are missing completely at random, this may still cause problems in the analysis. The smaller number of cases will result in less statistical power and increased variance. Furthermore, in subgroup comparisons, dropout may lead to a very small number of persons in a particular subgroup. Again the best strategy is to limit the problem by avoiding dropout as far as possible.

Nonresponse in longitudinal studies can occur at different points in time. First of all, not everyone who is invited to participate in a longitudinal study will do so. This is called *initial nonresponse*. Especially when the response burden is heavy, initial nonresponse at recruitment may be high. Initial nonresponse threatens the representativeness of the entire longitudinal study. Therefore, at the beginning of each longitudinal study one should first of all try to reduce the initial nonresponse as much as possible, and secondly collect as much data as possible on the nonrespondents to be used in statistical adjustment (e.g., weighting). Initial nonresponse is beyond the scope of this entry, but has been a topic of great interest for survey methodologist (*see* **Nonresponse in Sample Surveys**), and in the past decade much empirical knowledge on nonrespondents and reduction of nonresponse has been collected [1].

After the initial recruitment, when research participants have agreed to cooperate in the longitudinal study, nonresponse can occur at every time point or wave. This is called *dropout*. Dropout or wave nonresponse occurs when a participant in the study does not produce a completed questionnaire or interview at a specific time point, or fails to appear at a scheduled appointment in an experiment. If after a certain time point, research participants stop responding to all subsequent questionnaires or interviews, this is called *attrition* or *panel mortality*.

Finally, besides dropout, there is another source of nonresponse that may threaten the validity of longitudinal data and should be taken into account: item nonresponse. When item nonresponse occurs a unit (e.g., research participant, respondent) provides data, but for some reason data on particular questions or measurements are not available for analysis. Item

2 Dropouts in Longitudinal Data

nonresponse is beyond the scope of this entry; for an introductory overview on prevention and treatment of item nonresponse, see [2].

Starting at the initial recruitment, the researcher has to take steps to reduce future nonresponse. This needs careful planning and a total design approach. As research participants will be contacted over time, it is extremely important that the study has a well-defined image and is easily recognized and remembered at the next wave. A salient title, a recognizable logo, and graphical design are strong tools to create a positive study identity, and should be consistently used on all survey materials. For instance, the same logo and graphical style can be used on questionnaires, interviewer identity cards, information material, newsletters, and thank-you cards. When incentives are used, one should try to tie these in with the study. A good example comes from a large German study on exposure to printed media. The logo and mascot of this study is a little duckling, Paula. In German, the word 'Ente' or duck has the same meaning as the French word 'canard': a false (newspaper) report. Duckling Paula appears on postcards for the panel members, as a soft toy for the children, as an ornament for the Christmas tree, printed on aprons, t-shirts and so on, and has become a collector's item.

Dropout in longitudinal studies originates from three sources: failure to locate the research unit, failure to contact the potential respondent, and failure to obtain cooperation from the response unit [3].

Thus, the first task is limiting problems in *locating* research participants. At the recruitment phase or during the base-line study, the sample is fresh and address information is up-to-date. As time goes by, people move, and address, phone, and e-mail information may no longer be valid. It is of the utmost importance, that from the start at each consecutive time point, special locating information is collected. Besides the full name, also the maiden name should be recorded to facilitate follow-up after divorce. It is advisable to collect full addresses and phone numbers of at least three good friends or relatives as 'network contacts.' Depending on the study, names and addresses of parents, school-administration, or employers may be asked too. One should always provide 'change-of-address-cards' and if the budget allows, print on this card a message conveying that if one sends in a change of address, the researchers

will send a small 'welcome in your new home-gift' (e.g., a flower token, a DIY-shop token, a monetary incentive). It goes without saying, that the change-of-address cards are preaddressed to the study administration and that no postage is needed.

When the waves or follow-up times are close together, there is opportunity to keep locating-information up-to-date. If this is not the case, for instance in an annual or biannual study, it pays to incorporate between-wave locating efforts. For instance, sending a Christmas card with a spare 'change-of-address card', birthday cards for panel-members, and sending a newsletter with a request for address update. Additional strategies are to keep in touch and follow-up at known life events (e.g., pregnancy, illness, completion of education). This is not only motivating for respondents; it also limits loss of contact as change-of-address cards can be attached. Any mailing that is returned as undeliverable should be tracked immediately. Again, the better the contact ties in with the goal and topic of the study, the better it works. Examples are mother's day cards in a longitudinal study of infants, and individual feedback and growth curves in health studies. A total design approach should be adopted with material identifiable by house style, mascot, and logo, so that it is clear that the mail (e.g., child's birthday card) is coming from the study. Also ask regularly for an update, or additional network addresses. This is extremely important for groups that are mobile, such as young adults.

If the data are collected by means of face-to-face or telephone interviews, the interviewers should be clearly instructed in procedures for locating respondents, both during training and in a special tracking manual. Difficult cases may be allocated to specialized 'trackers'. Maintaining interviewer and tracker morale, through training, feedback, and bonuses helps to attain a high response. If other data collection procedures are used (e.g., mail or internet survey, experimental, or clinical measurements), staff members should be trained in tracking procedures. Trackers have to be trained in use of resources (e.g., phone books, telephone information services), and in the approach of listed contacts. These contacts are often the only means to successfully locate the research participant, and establishing rapport and maintaining the conversation with contacts are essential.

The second task is limiting the problems in *contacting* research participants. The first contact in a longitudinal study takes effort to achieve, just like establishing contact in a cross-sectional, one-time survey. Interviewers have to make numerous calls at different times, leave cards after a visit, leave messages on answering machines, or contact neighbors to extract information on the best time to reach the intended household. However, after the initial recruitment or base-line wave, contacting research participants is far less of a problem. Information collected at the initial contact can be fed to interviewers and used to tailor later contact attempts, provided, of course, that good locating-information is also available. In health studies and experimental research, participants often have to travel to a special site, such as a hospital, a mobile van, or an office. Contacts to schedule appointments should preferably be made by phone, using trained staff. If contact is being made through the mail, a phone number should always be available to allow research participants to change an inconvenient appointment, and trained staff members should immediately follow-up on 'no-shows'.

The third task is limiting dropout through lost willingness to cooperate. There is an extensive literature on increasing the cooperation in cross-sectional surveys. Central in this is reducing the cost for the respondent, while increasing the reward, motivating respondents and interviewers, and personalizing and tailoring the approach to the respondent [1, 4, 5]. These principles can be applied both during recruitment and at subsequent time points. When interviewers are used, it is crucial that interviewers are kept motivated and feel valued and committed. This can be done through refresher training, informal interviewer meetings, and interviewer incentives. Interviewers can and should be trained in special techniques to persuade and motivate respondents, and learn to develop a good relationship [1]. It is not strictly necessary to have the same interviewers revisit the same respondents at all time points, but it is necessary to feed interviewers information about previous contacts. Also, personalizing and adapting the wording of the questions by incorporating answers from previous measurements (dependent interviewing) has a positive effect on cooperation.

In general, prior experiences and especially 'respondent enjoyment' is related to cooperation at

subsequent waves [3]. A short and well-designed questionnaire helps to reduce response burden. Researchers should realize this and not try to get as much as possible out of the research participants at the first waves. In general, make the experience as nice as possible and provide positive feedback at each contact.

Many survey design features that limit *locating* problems, such as sending birthday and holiday cards and newsletters, also serve to nurture a good relationship with respondents and keep them motivated. In addition to these intrinsic incentives, explicit incentives also work well in retaining cooperation, and do not appear to have a negative effect on data quality [1]. Again the better the incentives fit the respondent and the survey, the better the motivational power (e.g., free downloadable software in a student-internet panel, air miles in travel studies, cute t-shirt and toys in infant studies). When research participants have to travel to a special site, a strong incentive is a special transportation service, such as a shuttle bus or car. Of course, all real transportation costs of participants should be reimbursed. In general, everything that can be done to make participation in a study as easy and comfortable as possible should be done. For example, provide for child-care during an on-site health study of teenage mothers.

Finally, a failure to cooperate at a specific time point does not necessarily imply a complete dropout from the study. A respondent may drop out temporarily because of time pressure or lifetime changes (e.g., change of job, birth of child, death of spouse). If a special attempt is made, the respondent may not be lost for the next waves.

In addition to the general measures described above, each longitudinal study can and should use data from earlier time points to design for nonresponse prevention. Analysis of nonrespondents (persons unable to locate again and refusals) provides profiles for groups at risk. Extra effort then may be put into research participants with similar profiles who are still in the study (e.g., offer an extra incentive, try to get additional network information). In addition, these nonresponse analyses provide data for better statistical adjustment.

With special techniques, it is possible to reduce dropout in longitudinal studies considerably, but it can never be prevented completely. Therefore, adjustment procedures will be necessary during analysis.

4 Dropouts in Longitudinal Data

Knowing why dropout occurs makes it possible to choose the correct statistical adjustment procedure. Research participants may drop out of longitudinal studies for various reasons, but of one thing one may be assured: they do not drop out completely at random. If the reasons for dropout are not related to the topic of the study, responses are missing at random and relatively simple weighting or imputation procedures can be adequately employed. But if the reasons for dropout *are* related to the topic, responses are *not* missing at random and a special model for the dropout must be included in the analysis to prevent bias. In longitudinal studies, usually auxiliary data are available from earlier time points, but one can only guess at the reasons why people drop out. It is advisable to ask for these reasons directly in a special short exit-interview. The data from this exit interview, together with auxiliary data collected at earlier time points, can then be used to statistically model the dropout and avoid biased results.

References

- [1] Special issue on survey nonresponse, *Journal of Official Statistics (JOS)* (1999). **15**(2), Accessible free of charge on www.jos.nu.

- [2] De Leeuw, E.D., Hox, J. & Huisman, M. (2003). Prevention and treatment of item nonresponse, *Journal of Official Statistics* **19**(2), 153–176.
- [3] Lepkowski, J.M. & Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little, eds, Wiley, New York.
- [4] Dillman, D.A. (2000). *Mail and Internet Surveys*, Wiley, New York, see also Dillman (1978) *Mail and telephone surveys*.
- [5] Groves, R.M. & Couper, M.P. (1998). *Nonresponse in Household Surveys*, Wiley, New York.

Further Reading

- Kasprzyk, D., Duncan, G.J., Kalton, G. & Singh, M.P. (1989). *Panel Surveys*, Wiley, New York.
- The website of the *Journal of Official Statistics* <http://www.jos.nu> contains many interesting articles on survey methodology, including longitudinal studies and panel surveys.

(See also **Generalized Linear Mixed Models**)

EDITH D. DE LEEUW

Dropouts in Longitudinal Studies: Methods of Analysis

RODERICK J. LITTLE

Volume 1, pp. 518–522

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Dropouts in Longitudinal Studies: Methods of Analysis

Introduction

In longitudinal behavioral studies, it is difficult to obtain outcome measures for all participants throughout the study. When study entry is staggered, participants entering late may not have a complete set of measures at the time of analysis. Some participants may move and lose contact with the study, and others may drop out for reasons related to the study outcomes; for example, in a study of pain, individuals who do not obtain relief may discontinue treatments, or in a study of treatments to stop smoking, people who continue to smoke may be more likely to drop out of the study rather than admit to lack of success. These mechanisms of drop out create problems for the analysis, since the cases that remain are a biased sample and may distort treatment comparisons, particularly if the degree of dropout is differential between treatment arms.

In the **clinical trial** setting, a useful distinction [11] is between *treatment* dropouts, where individuals discontinue an assigned treatment, and *analysis* dropouts, where outcome data are not recorded. A treatment dropout is not necessarily an analysis dropout, in that study outcomes can still be recorded after the lapse in treatment protocol. Since these outcomes do not reflect the full effect of the treatment, the values that would have been recorded if the participant had remained in the study might still be regarded as missing, converting a treatment dropout into an analysis dropout. For discussion of treatment dropouts and, more generally, *treatment compliance*, see [2]. From now on, I focus the discussion on methods for handling analysis dropouts.

In general, any method for handling dropouts requires assumptions, and cannot fully compensate for the loss of information. Hence, the methods discussed here should not substitute for good *study design* to minimize dropout, for example, by keeping track of participants and encouraging them to continue in the study. If participants drop out, efforts should be made to obtain some information (for

example, the reason for drop out) since that can be useful for statistical analysis.

Complete-case Analysis and Imputation

A simple way of dealing with **missing data** is complete-case (CC) analysis, also known as listwise deletion, where incomplete cases are discarded and standard analysis methods are applied to the complete cases (e.g., [10, Chapter 3]). In many statistical packages, this is the default analysis. The exclusion of incomplete cases represents a loss of information, but a more serious problem is that the complete cases are often a biased sample. A useful way of assessing this is to compare the observed characteristics of completers and dropouts, for example, with *t* Tests, comparing means or chi-squared tests, comparing categorical variables. A lack of significant differences indicates that there is no evidence of bias, but this is far from conclusive since the groups may still differ on the outcomes of interest.

A simple approach to incomplete data that retains the information in the incomplete cases is to impute or fill in the missing values (e.g., [10, Chapter 4], and see **Multiple Imputation**). It is helpful to think of imputations as being based on an imputation model that leads to a predictive distribution of the missing values. Missing values are then either imputed using the mean of this predictive distribution, or as a random draw from the predictive distribution. Imputing means leads to consistent estimates of means and totals from the filled-in data; imputing draws is less efficient, but has the advantage that nonlinear quantities, such as variances and percentiles, are also consistently estimated from the imputed data.

Examples of predictive mean imputation methods include unconditional mean imputation, where the sample mean of the observed cases is imputed, and regression imputation, where each missing value is replaced by a prediction from a regression on observed variables (see **Multiple Linear Regression**). In the case of univariate nonresponse, with $Y_1, \dots, Y_{k-1}$ fully observed and Y_k sometimes missing, the regression of Y_k on $Y_1, \dots, Y_{k-1}$ is estimated from the complete cases, including interactions, and the resulting prediction equation is used to impute the estimated conditional mean for each missing value of Y_k . Regression imputation is superior to unconditional mean imputation since it exploits and preserves

relationships between imputed and observed variables that are otherwise distorted. For repeated-measures data with dropouts, missing values can be filled in sequentially, with each missing value for each subject imputed by regression on the observed or previously imputed values for that subject.

Imputation methods that impute draws include *stochastic regression* imputation [10, Example 4.5], where each missing value is replaced by its regression prediction plus a random error with variance equal to the estimated residual variance. A common approach for longitudinal data imputes missing values for a case with the last recorded observation for that case. This method is common, but not recommended since it makes the very strong and often unjustified assumption that the missing values in a case are all identical to the last observed value. Better methods for longitudinal imputation include imputation based on row and column fits [10, Example 4.11].

The imputation methods discussed so far can yield consistent estimates of the parameters under well-specified imputation models, but the analysis of the filled-in data set does not take into account the added uncertainty from the imputations. Thus, statistical inferences are distorted, in the sense that standard errors of parameter estimates computed from the filled-in data will typically be too small, **confidence intervals** will not have their nominal coverage, and *P* values will be too small. An important refinement of imputation, **multiple imputation**, addresses this problem [18]. A predictive distribution of plausible values is generated for each missing value using a statistical model or some other procedure. We then impute, not just one, but a set of *M* (say *M* = 10) draws from the predictive distribution of the missing values, yielding *M* data-sets with different draws plugged in for each of the missing values. For example, the stochastic regression method described above could be repeated *M* times. We then apply the analysis to each of the *M* data-sets and combine the results in a simple way. In particular, for a single parameter, the multiple-imputation estimate is the average of the estimates from the *M* data-sets, and the variance of the estimate is the average of the variances from the *M* data-sets plus $1 + 1/5$ times the sample variance of the estimates over the *M* data-sets (The factor $1 + 1/M$ is a small-*M* correction). The last quantity here estimates the contribution to the variance from imputation uncertainty, missed by single imputation methods. Similar formulae apply

for more than one parameter, with variances replaced by covariance matrices. For other forms of multiple-imputation inference, see [10, 18, 20]. Often, multiple imputation is not much more difficult than doing single imputation; most of the work is in creating good predictive distributions for the missing values. Software for multiple imputation is becoming more accessible, see PROC MI in [15], [19], [20] and [22].

Maximum Likelihood Methods

Complete-case analysis and imputation achieve a rectangular data set by deleting the incomplete cases or filling in the gaps in the data set. There are other methods of analysis that do not require a rectangular data set, and, hence, can include all the data without deletion or imputation. One such approach is to define a summary measure of the treatment effect for each individual based on the available data, such as change in an outcome between baseline and last recorded measurement, and then carry out an analysis of the summary measure across individuals (*see Summary Measure Analysis of Longitudinal Data*). For example, treatments might be compared in terms of differences in means of this summary measure. Since the precision of the estimated summary measure varies according to the number of measurements, a proper statistical analysis gives less weight to measures from subjects with shorter intervals of measurement. The appropriate choice of weight depends on the relative size of intraindividual and interindividual variation, leading to complexities that negate the simplicity of the approach [9].

Methods based on **generalized estimating equations** [7, 12, 17] also do not require rectangular data. The most common form of estimating equation is to generate a likelihood function for the observed data based on a statistical model, and then estimate the parameters to maximize this likelihood [10, chapter 6]. **Maximum likelihood** methods for multilevel or **linear multilevel models** form the basis of a number of recent statistical software packages for repeated-measures data with missing values, which provide very flexible tools for statistical modeling of data with dropouts. Examples include SAS PROC MIXED and PROC NL MIXED [19], methods for longitudinal data in S-PLUS functions *lme* and *nlme* [13], HLM [16], and the Stata programs

gllamm in [14] (see **Software for Statistical Analyses**). Many of these programs are based on linear multilevel models for normal responses [6], but some allow for binary and ordinal outcomes [5, 14, 19] (see **Generalized Linear Mixed Models**).

These maximum likelihood analyses are based on the *ignorable* likelihood, which does not include a term for the missing data mechanism. The key assumption is that the data are *missing at random*, which means that dropout depends only on the observed variables for that case, and not on the missing values or the unobserved random effects (see [10], chapter 6). In other words, missingness is allowed to depend on values of covariates, or on values of repeated measures recorded prior to drop out, but cannot depend on other quantities. Bayesian methods (see **Bayesian Statistics**) [3] under noninformative priors are useful for small sample inferences.

Some new methods allow us to deal with situations where the data are not missing at random, by modeling the joint distribution of the data and the missing data mechanism, formulated by including a variable that indicates the pattern of missing data [Chapter 15 in 10], [1, 4, 8, 14, 23, 24]. However, these nonignorable models are very hard to specify and vulnerable to model misspecification. Rather than attempting simultaneously to estimate the parameters of the dropout mechanism and the parameters of the complete-data model, a more reliable approach is to do a **sensitivity analysis** to see how much the answers change for various assumptions about the dropout mechanism (see [Examples 15.10 and 15.12 in 10], [21]). For example, in a smoking cessation trial, a common practice is to treat dropouts as treatment failures. An analysis based on this assumption might be compared with an analysis that treats the dropouts as missing at random. If substantive results are similar, the analysis provides some degree of confidence in the robustness of the conclusions.

Conclusion

Complete-case analysis is a limited approach, but it might suffice with small amounts of dropout. Otherwise, two powerful general approaches to statistical analysis are maximum likelihood estimation and multiple imputation. When the imputation model and the

analysis model are the same, these methods have similar large-sample properties. One useful feature of multiple imputation is that the imputation model can differ from the analysis model, as when variables not included in the final analysis model are included in the imputation model [10, Section 10.2.4]. Software for both approaches is gradually improving in terms of the range of models accommodated. Deviations from the assumption of missing at random are best handled by a sensitivity analysis, where results are assessed under a variety of plausible alternatives.

Acknowledgment

This research was supported by National Science Foundation Grant DMS 9408837.

References

- [1] Diggle, P. & Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion), *Applied Statistics* **43**, 49–94.
- [2] Frangakis, C.E. & Rubin, D.B. (1999). Addressing complications of intent-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes, *Biometrika* **86**, 365–379.
- [3] Gilks, W.R., Wang, C.C., Yvonnet, B. & Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs' sampling, *Biometrics* **49**, 441–453.
- [4] Hausman, J.A. & Wise, D.A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment, *Econometrica* **47**, 455–473.
- [5] Hedeker, D. (1993). *MIXOR: A Fortran Program for Mixed-effects Ordinal Probit and Logistic Regression*, Prevention Research Center, University of Illinois at Chicago, Chicago, 60637.
- [6] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.
- [7] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
- [8] Little, R.J.A. (1995). Modeling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association* **90**, 1112–1121.
- [9] Little, R.J.A. & Raghunathan, T.E. (1999). On summary-measures analysis of the linear mixed-effects model for repeated measures when data are not missing completely at random, *Statistics in Medicine* **18**, 2465–2478.
- [10] Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition, John Wiley, New York.
- [11] Meinert, C.L. (1980). Terminology - a plea for standardization, *Controlled Clinical Trials* **2**, 97–99.

4 Dropouts in Longitudinal Studies: Methods of Analysis

- [12] Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements, *Statistics in Medicine* **12**, 1723–1732.
- [13] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*, Springer-Verlag, New York.
- [14] Rabe-Hesketh, S., Pickles, A. & Skrondal, A. (2001). GLLAMM Manual, Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London, For associated software see <http://www.gllamm.org/>
- [15] Raghunathan, T., Lepkowski, J. VanHoewyk, J. & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology* **27**(1), 85–95. For associated IVEWARE software see <http://www.isr.umich.edu/src/smp/ive/>
- [16] Raudenbush, S.W., Bryk, A.S. & Congdon, R.T. (2003). *HLM 5*, SSI Software, Lincolnwood.
- [17] Robins, J., Rotnitzky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**, 106–121.
- [18] Rubin, D.B. (1987). *Multiple Imputation in Sample Surveys and Censuses*, John Wiley, New York.
- [19] SAS. (2003). *SAS/STAT Software, Version 9*, SAS Institute, Inc., Cary.
- [20] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, CRC Press, New York, For associated multiple imputation software, see <http://www.stat.psu.edu/~jls/>
- [21] Scharfstein, D., Rotnitzky, A. & Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric models, *Journal of the American Statistical Association* **94**, 1096–1146 (with discussion).
- [22] Van Buuren, S., and Oudshoorn, C.G.M. (1999). Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054. For associated software, see <http://www.multiple-imputation.com>.
- [23] Wu, M.C. & Bailey, K.R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics* **45**, 939–955.
- [24] Wu, M.C. & Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* **44**, 175–188.

(See also **Dropouts in Longitudinal Data; Longitudinal Data Analysis**)

RODERICK J. LITTLE

Dummy Variables

JOSE CORTINA

Volume 1, pp. 522–523

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Dummy Variables

A categorical variable with more than two levels is, in effect, a collection of $k - 1$ variables where k is the number of levels of the categorical variable in question. Consider the categorical variable Religious Denomination. For the sake of simplicity, let us say that it contains three levels: Christian, Muslim, and Jewish. If we were to code these three levels 1, 2, and 3, we might have a data set as in Table 1:

We could use this variable as a predictor of Political Conservatism (PC). Thus, we could regress a measure of PC onto our Religion variable. The regression weight for the Religion variable would be gibberish, however, because the predictor was coded arbitrarily. A regression weight gives the expected change in the dependent variable per single point increase in the predictor. When a predictor is arbitrarily coded, a single point increase has no meaning. The numbers in the RelDen column are merely labels, and they could have been assigned to the groups in any combination. Thus, the regression weight would be very different if we had chosen a different arbitrary coding scheme.

The problem stems from the fact that this categorical variable actually contains $k - 1 = 2$ comparisons among the k groups. In order to capture all of the information contained in the distinctions among these groups, we must have all $k - 1 = 2$ of these comparisons. The generic term for such a comparison variable is *Dummy Variable*.

Strictly speaking, a dummy variable is a dichotomous variable such that if a given subject belongs to a particular group, that subject is given a score of 1 on the dummy variable. Members of other groups are given a zero. One way of handling the Religion

Table 1 Data for a 3-level categorical variable

Subject #	RelDen
1	1
2	2
3	3
4	1
5	2
6	3

Table 2 Data and dummy codes for a 3-level categorical variable

Subject #	RelDen	Dummy1	Dummy2
1	1	1	0
2	2	0	1
3	3	0	0
4	1	1	0
5	2	0	1
6	3	0	0

variable above would be to create two dummy variables to represent its three levels. Thus, we would have data such as those in Table 2.

Dummy1 is coded such that Christians receive a 1 while Muslims and Jews receive a zero. Dummy2 is coded such that Muslims receive a 1 while Jews receive a zero. These two dummy variables as a set contain all of the information contained in the three-level Religion variable. That is, if I know someone's score on both of the dummy variables, then I know exactly which group that person belongs to: someone with a 1 and a 0 is a Christian, someone with a 0 and 1 is a Muslim, and someone with two zeros is a Jew. Because there is no variable for which Jews receive a 1, this is labeled the uncoded group.

Consider once again the prediction of PC from Religion. Whereas the three-level categorical variable cannot be used as a predictor, the two dummy variables can. Regression weights for dummy variables involve comparisons to the uncoded group. Thus, the weight for Dummy1 would be the difference between the PC mean for Christians and the PC mean for Jews (the uncoded group). The weight for Dummy2 would be the difference between the PC mean for Muslims and the PC mean for Jews. The R-squared from the regression of PC onto the set of dummy variables represents the percentage of PC variance accounted for by Religious Denomination.

The more general term for such coded variables is *Design Variable*, of which dummy coding is an example. Other examples are effect coding (in which the uncoded group is coded -1 instead of 0) and contrast coding (in which coded variables can take on any number of values). The appropriateness of a coding scheme depends on the sorts of comparisons that are of most interest.

JOSE CORTINA

Encyclopedia of Statistics in

BEHAVIORAL SCIENCE

Editors in Chief **Brian Everitt** **David Howell**

VOLUME

2

eco-lor

VOLUME 2

Ecological Fallacy.	525-527	Factor Analysis: Multitrait-Multimethod.	623-628
Educational Psychology: Measuring Change Over Time.	527-532	Factor Analysis of Personality Measures.	628-636
Effect Size Measures.	532-542	Factor Score Estimation.	636-644
Eigenvalue/Eigenvector.	542-543	Factorial Designs.	644-645
Empirical Quantile-Quantile Plots.	543-545	Family History Versus Family Study Methods in Genetics.	646-647
Epistasis.	545-546	Family Study and Relative Risk.	647-648
Equivalence Trials.	546-547	Fechner, Gustav T.	649-650
Error Rates.	547-549	Field Experiment.	650-652
Estimation.	549-553	Finite Mixture Distributions.	652-658
ETA and ETA Squared.	553-554	Fisher, Sir Ronald Aylmer.	658-659
Ethics in Research.	554-562	Fisherian Tradition in Behavioral Genetics.	660-664
Evaluation Research.	563-568	Fixed and Random Effects.	664-665
Event History Analysis.	568-575	Fixed Effect Models.	665-666
Exact Methods for Categorical Data.	575-580	Focus Group Techniques.	666-668
Expectancy Effect.	581-582	Free Response Data Scoring.	669-673
Expectation.	582-584	Friedman's Test.	673-674
Experimental Design.	584-586	Functional Data Analysis	675-678
Exploratory Data Analysis.	586-588	Fuzzy Cluster Analysis.	678-686
External Validity.	588-591	Galton, Francis.	687-688
Face-to-Face Surveys.	593-595	Game Theory.	688-694
Facet Theory.	595-599	Gauss, Johann Carl Friedrich.	694-696
Factor Analysis: Confirmatory.	599-606	Gene-Environment Correlation.	696-698
Factor Analysis: Exploratory.	606-617	Gene-Environment Interaction.	698-701
Factor Analysis: Multiple Groups.	617-623		

Generalizability.	702-704	Heteroscedasticity and Complex Variation.	790-795
Generalizability Theory: Basics.	704-711	Heuristics.	795
Generalizability Theory: Estimation.	711-717	Heuristics: Fast and Frugal.	795-799
Generalizability Theory: Overview.	717-719	Hierarchical Clustering.	799-805
Generalized Additive Model.	719-721	Hierarchical Item Response Theory Modeling.	805-810
Generalized Estimating Equations (GEE).	721-729	Hierarchical Models.	810-816
Generalized Linear Mixed Models.	729-738	High-Dimensional Regression.	816-818
Generalized Linear Models (GLM)	739-743	Hill's Criteria of Causation.	818-820
Genotype.	743-744	Histogram.	820-821
Geometric Mean.	744-745	Historical Controls.	821-823
Goodness of Fit.	745-749	History of Analysis of Variance.	823-826
Goodness of Fit for Categorical Variables.	749-753	History of Behavioral Statistics.	826-829
Gosset, William Sealy.	753-754	History of Correlational Measurement.	836-840
Graphical Chain Models.	755-757	History of Discrimination and Clustering.	840-842
Graphical Methods Pre-twentieth Century.	758-762	History of Factor Analysis: A Psychological Perspective	842-851
Graphical Presentation of Longitudinal Data.	762-772	History of Factor Analysis: Statistical Perspective.	851-858
Growth Curve Modeling.	772-779	History of Intelligence Measurement.	858-861
Guttman, Louise (Eliyahu).	780-781	History of Mathematical Learning Theory.	861-864
Harmonic Mean.	783-784	History of Multivariate Analysis of Variance.	864-869
Hawthorne Effect.	784-785	History of Path Analysis.	869-875
Heritability.	786-787	History of Psychometrics.	875-878
Heritability: Overview.	787-790		

History of Surveys of Sexual Behavior.	878-887	Internal Consistency.	934-936
History of the Control Group.	829-836	Internal Validity.	936-937
Hodges-Lehman Estimator.	887-888	Internet Research Methods.	937-940
Horseshoe Pattern.	889	Interquartile Range.	941
Hotelling, Howard.	889-891	Interrupted Time Series Design.	941-945
Hull, Clark L.	891-892	Intervention Analysis.	946-948
Identification.	893-896	Intraclass Correlation.	948-954
Inbred Strain Study.	896-898	Intrinsic Linearity.	954-955
Incidence.	898-899	INUS Conditions.	955-958
Incomplete Contingency Tables.	899-900	Item Analysis.	958-967
Incompleteness of Probability Models.	900-902	Item Bias Detection: Classical Approaches.	967-970
Independence: Chi-square and likelihood Ratio Tests.	902-907	Item Bias Detection: Modern Approaches.	970-974
Independent Components Analysis.	907-912	Item Exposure Detection.	974-978
Independent Pathway Model.	913-914	Item Response Theory: Cognitive Models.	978-982
Index Plots.	914-915	Item Response Theory (IRT) Models for Dichotomous Data.	982-990
Industrial/Organizational Psychology.	915-920	Item Response Theory Models for Rating Scale Data.	995-1003
Influential Observations.	920-923	Item Response Theory Models for Polytomous Response Data.	990-995
Information Matrix.	923-924	Jackknife.	1005-1007
Information Theory.	924-927	Jonckheere-Terpstra Test.	1007-1008
Instrumental Variable.	928	Kendall, Maurice George.	1009-1010
Intention-to-Treat.	928-929	Kendall's Coefficient of Concordance.	1010-1011
Interaction Effects.	929-933	Kendall's Tau - t.	1011-1012
Interaction Plot.	933-934		

Kernel Smoothing.	1012-1017	Longitudinal Data Analysis.	1098-1101
K-Means Analysis.	1017-1022	Longitudinal Designs in Genetic Research.	1102-1104
Kolmogorov, Andrey Nikolaevich.	1022-1023	Lord, Frederic Mather.	1104-1106
Kolmogorov-Smirnov Tests.	1023-1026	Lord's Paradox.	1106-1108
Kruskal-Wallis Test.	1026-1028		
Kurtosis.	1028-1029		
Laplace, Pierre Simon (Marquis de).	1031-1032		
Latent Class Analysis.	1032-1033		
Latent Transition Models.	1033-1036		
Latent Variable.	1036-1037		
Latin Squares Designs.	1037-1040		
Laws of Large Numbers.	1040-1041		
Least Squares Estimation.	1041-1045		
Leverage Plot.	1045-1046		
Liability Threshold Models.	1046-1047		
Linear Model.	1048-1049		
Linear Models: Permutation Methods.	1049-1054		
Linear Multilevel Models.	1054-1061		
Linear Statistical Models for Causation: A Critical Review.	1061-1073		
Linkage Analysis.	1073-1075		
Logistic Regression.	1076-1082		
Log-linear Models.	1082-1093		
Log-linear Rasch Models for Stability and Change.	1093-1097		

Ecological Fallacy

ITA G.G. KREFT

Volume 2, pp. 525–527

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ecological Fallacy

Introduction to Ecological Fallacy

Ecological fallacy is an old term from the fifties, described for the first time by Robinson [4]. The fallacy is defined as the mistake a consumer of research results makes when a statistical effect (e.g., a correlation), measured and calculated over groups, is used as a statement valid for individuals. The opposite, a less common mistake in research, is the same fallacy. In the world outside research, this last fallacy would be called *discrimination*. Discrimination is blaming a group for the behavior of an individual, such as if someone who is Dutch, like myself, behaves badly, and the inference is made that all Dutch people behave that way. To avoid ecological fallacies, all data analysis results should be stated at the level where the analysis was executed. In sum, statements that make cross-level inferences should be avoided, unless good evidence exists that they are safe to make. A better solution is to analyze data at both levels of the hierarchy in multilevel analysis [1] and [3], which is a separate issue (*see Linear Multilevel Models*).

The Robinson Effect

Robinson [4] published the first example of an ecological fallacy, based on data collected in 1930. Reported correlations showed a strong relationship between the percentage immigrants and the level of illiteracy in the United States. The contradiction in the data, according to Robinson, was between the large and negative correlation ($r = -0.53$) using states as the unit of analysis, and the much smaller and positive correlation ($r = 0.12$) when individuals are the unit of analysis. The negative correlation indicates that if a state has a high percentage of foreign-born people, the level of illiteracy of that state is low, while the individual correlation indicates that foreign-born individuals are more often illiterate. This reversal of the correlation sign can be traced to an unmeasured third variable, affluence of a state, as shown by Hanushek, Jackson, and Kain [2]. In the thirties, affluent states attracted more immigrants, and, at the same time, affluence gave children more

opportunities to learn and go to school, thus lowering the illiteracy in the state as a whole.

Ecological Fallacy is Still Around

Kreft and De Leeuw [3] report an example of the relationship between education and income measured over people working in 12 different industries. The individual correlation is low but positive ($r = 0.12$). However, the aggregated correlation over industries is high and has a negative sign ($r = -0.71$). At individual level, a positive relation between education and income is found, but when the data are aggregated to industry level, the opposite appears to be true. The data suggest that individuals can expect some gain in income by going back to school, but, at the industry level, the reverse seems to be true. The opposite conclusion at the industry level is the result of a confounding factor, the type of industry, which is either private or public. In the private sector, high-paying industries exist that do not require high levels of education (e.g., transportation, real estate), while in the public sector, some low-paying industries are present, such as universities and schools, demanding high levels of schooling. Again, it is clear that characteristics of industries are a threat to the validity of cross-level inferences.

Conditions Where Ecological Fallacies May Appear

When interested in individual as well as group effects, the data matrix (and the covariance matrix) will be divided in two parts: a within (group) and a between (group) part, as shown in equation (1), where C = covariance, t = total, b = between, and w = within.

$$C(x_t, y_t) = C(x_b, y_b) + C(x_w, y_w) \quad (1)$$

The variances (V) of variables x and y can be defined in similar ways. A measure for between-group variance for a variable is defined as η^2 . η^2 is the ratio of between-group and total variation (*see Effect Size Measures*). Equally, the within-group variation is $1 - \eta^2$ the ratio of the within-group variation and the total variation. Ecological differences in regression coefficients occur when between and within variations are very different, and/or one of the two is (almost) zero. Given that

2 Ecological Fallacy

the total regression coefficient (b_t) is a weighted sum of the between-group regression b_b and the pooled-within group regression b_w , as in $b_t = \eta^2 b_b + (1 - \eta^2) b_w$, it follows that if $b_b = 0$, $b_t = b_w$, and the total regression are equal to the pooled-within regression. Following the same reasoning, $b_t = b_b$ when $b_w = 0$. In this situation, there is no pooled-within regression, and all the variation is between groups.

Using the definition of the regression coefficient for the total, the between as well as for the within group, as $\{C(xy)/V(x)\}$, it can be shown that the total regression coefficient is different from the within and/or the between-coefficient in predictable ways. The definition of the regression coefficient b_b in (2) is:

$$b(x_b, y_b) = \frac{C(x_b, y_b)}{V(x_b)} \quad (2)$$

Using the definition of η^2 and $1 - \eta^2$, we can replace the between-covariance, $C(x_b, y_b)$, in the numerator by the total covariance minus the within-covariance. The between-regression coefficient is redefined in (3):

$$b(x_b, y_b) = \frac{\{C(x_t, y_t) - C(x_w, y_w)\}}{V(x_b)} \quad (3)$$

Rearranging terms in (3), and replacing the total- as well as the within-covariances in the numerator by $b(x_t, y_t)V(x_t)$ and $b(x_w, y_w)V(x_w)$ respectively results in (4).

$$b(x_b, y_b) = \frac{\{b(x_t, y_t)V(x_t) - b(x_w, y_w)V(x_w)\}}{V(x_b)} \quad (4)$$

Dividing numerator and denominator in (4) by the total variance $V(x_t)$ and replacing the resulting $V(x_b)/V(x_t)$ by $\eta^2(x)$, doing the same with $V(x_w)/V(x_t)$ and replacing it with $1 - \eta^2(x)$ results in (5).

$$b(x_b, y_b) = \frac{\{b(x_t, y_t) - b(x_w, y_w)(1 - \eta^2(x))\}}{\eta^2(x)} \quad (5)$$

Equation (5) shows that the between-regression coefficient is the total regression coefficient minus the within-coefficient, weighted by the within- and between- η^2 . In situations where the regression coefficient between x and y is zero, $\{b(x_t, y_t) = 0\}$, the between-groups coefficients will be a weighted sum of the within-regression as in (6) and have an opposite sign.

$$b(x_b, y_b) = \frac{\{-b(x_w, y_w)(1 - \eta^2(x))\}}{\eta^2(x)} \quad (6)$$

The same equation (6) also shows that, if the group effect, $\eta^2(x)$, is large, $\{1 - \eta^2(x)\}$ will be small, both resulting in a larger between-coefficient than a within-coefficient, explaining the ‘blowing up’ of the aggregated coefficient as compared to the individual one.

References

- [1] Aitkin, M.A. & Longford, N.T. (1986). Statistical Modelling issues in school effectiveness research, *Journal of the Royal Statistical Society, Series A* **149**, 1–43.
- [2] Hanushek, E.A., Jackson, J. & Kain, J. (1974). Model specification, use and aggregate data and the ecological fallacy, *Political Methodology* **1**, 89–107.
- [3] Kreft, I.G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling*, Sage, London.
- [4] Robinson, W.S. (1950). Ecological correlations and the behavior of individuals, *Sociological Review* **15**, 351–357.

ITA G.G. KREFT

Educational Psychology: Measuring Change Over Time

JONNA M. KULIKOWICH

Volume 2, pp. 527–532

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Educational Psychology: Measuring Change Over Time

Educational psychologists examine many teaching and learning phenomena that involve change. For instance, researchers are interested in how students' abilities to read, spell, and write develop in elementary school. Investigators who study social and emotional variables such as personality traits and self-esteem examine how these variables change from adolescence to adulthood. Similarly, educators invested in the promotion of health and fitness are concerned with ways to improve eating, sleeping, and exercise behaviors. All of these examples lead to research questions for which investigators collect data to determine how variable scores change.

To address whether scores change in time, educational researchers often use **repeated measures analyses of variance (ANOVA)**. Studies employing repeated measures designs in educational psychological research are prevalent as any inspection of the best journals in the field shows [8, 17, 23]. Repeated measures ANOVA is simply an extension of the dependent samples t Test. Rather than comparing mean performance for the same group of examinees on two occasions (e.g., days, weeks, years – what statisticians refer to as trials), repeated measures ANOVA allows statisticians to compare means for three or more trials.

GLM for Repeated Measures and Statistical Assumptions

Repeated measures analyses are examined within the framework of the **Generalized Linear Model (GLM)** in statistics. Suppose an educational psychologist is interested in whether spelling scores for a group of elementary students change over a five-week period. Assume the researcher administers a spelling test every Friday of each week. According to Glass and Hopkins [7], any spelling score can be denoted by the symbol, X_{st} , where the subscripts s and t stand for s th student and t th trial, respectively. Thus, any score can be represented by the following equation:

$$X_{st} = \mu + a_s + \beta_t + a\beta_{st} + \varepsilon_{st}, \quad (1)$$

where μ is the grand mean, a_s is the random factor effect for students, β_t is the fixed factor effect for trials, $a\beta_{st}$ is an interaction term, and ε_{st} is the residual of the score X_{st} when predicted from the other terms in the model. Because of the inclusion of one random factor (students) and one fixed factor (trials), this repeated measures ANOVA model is a mixed-effects model (see **Linear Multilevel Models**).

As with all **analysis of variance (ANOVA)** models within the GLM framework, to determine whether there is a significant main effect for trial, the educational psychologist should report the F -ratio, degrees of freedom, the Mean Square Error term, and an index of effect size. Often, the partial eta-squared (η^2) value is reported to indicate small, medium, and large effects (0.01, 0.06, and 0.16 respectively) [3] (see **Effect Size Measures**).

Before investigators conclude, however, that at least one trial is different from the other four, they must ensure that certain statistical assumptions are met. Like other ANOVA models, the typical assumptions of normality and independence of observations are important in repeated measures analyses, too. In addition, statisticians must test the assumption of **sphericity** when analyzing repeated measures data. This assumption requires that the variances for all possible differences between trials be equal. The assumption relates to the dependency that exists between distributions of trials because the data come from the same students. Of course, researchers do not expect the variances of the differences to be exactly identical. Instead, they want to know if they differ to a significant degree.

Fortunately, there are many well-known tests of the sphericity assumption. Greenhouse and Geisser's [9] and Huynh and Feldt's [12] procedures adjust the degrees of freedom of the F test when the sphericity assumption is violated. To determine if it is violated, their procedures result in epsilon values (ε 's, which are reported in printouts such as those of SPSS [22]). Values close to 1.00 are desirable. If the Greenhouse–Geisser or Huynh–Feldt values suggest some departure from sphericity, it might be wise to adjust the degrees of freedom. Adjustment makes for a more *conservative* test, which means that a larger F -ratio is required to reject H_0 .

After researchers determine that there is a significant main effect for trials, two follow-up questions can be addressed: (a) Between which pairs of means is there significant change? (b) Is there

a significant trend across weeks? These two questions pertain to *post hoc* comparisons and trend analysis (see **Multiple Comparison Procedures**), respectively. Statistically, both topics are addressed extensively in the literature [15, 21]. However, with repeated measures designs, educational psychologists must be very careful about the interpretations they make. The next two sections describe some methodological considerations.

Post Hoc Comparisons

Readers can browse any textbook containing chapters on ANOVA, and find ample information about *post hoc* (see **A Priori v Post Hoc Testing**) comparisons. These mean comparisons are made after a so-called omnibus test is significant [13]. Thus, any time there is a main effect for a factor with more than two levels, *post hoc* comparisons indicate where the significant difference or differences are. *Post hoc* comparisons control familywise error rates (see **Error Rates**) (the probability of a Type I error is α for the set of comparisons). For between-subjects designs, educational psychologists can choose from many *post hoc* comparisons (e.g., Duncan, Fisher's least significant difference (LSD), Student–Newman–Kuels, Tukey's honestly significant difference (HSD)). For repeated measures analyses, do well-known *post hoc* comparison procedures work? Unfortunately, the answer is not a simple one, and statisticians vary in their opinions about the best way to approach the question [11, 16]. Reasons for the complexity pertain to the methodological considerations of analyzing repeated measures data. One concern about locating mean differences relates to violations of the assumption of sphericity. Quite often, the degrees of freedom must be adjusted for the omnibus F test because the assumption is violated. While statistical adjustments such as the Greenhouse–Geisser correction assist in accurately rejecting the null hypothesis for the main effect, *post hoc* comparisons are not protected by the same adjustment [18].

Fortunately, Keselman and his colleagues [13, 14] continue to study repeated measures designs extensively. In their papers, they recommend the best methods of comparing means given violations of sphericity or multiple sphericity (for designs with grouping variables). Reviewing these recommendations is beyond the scope of this chapter, but

Bonferroni methods appear particularly useful and result in fewer Type I errors than other tests (e.g., Tukey) when assumptions (such as sphericity) are violated [16].

Trend Analysis

While *post hoc* comparison procedures are one way to examine the means for repeated measures designs, another approach to studying average performance over time is trend analysis. Tests of linear and non-linear trends in studies of growth and development in educational psychology appear periodically [4]. Glass and Hopkins [7] indicate that as long as the repeated trials constitute an ordinal or interval scale of measurement, such as the case of weeks, tests for significant trends are appropriate. However, if the repeated factors are actually related measures, such as different subscale averages of a standardized test battery, then trend analyses are not appropriate. Thus, educational psychologists should not use trend analysis to study within-student differences on the graduate record examination (GRE) quantitative, verbal, and analytic subscales (see **Growth Curve Modeling**).

Figure 1 depicts a trend for hypothetical spelling data collected over a five-week period. Using standard contrasts, software programs such as SPSS [22] readily report whether the data fit linear, quadratic, cubic, or higher-order polynomial models. For the means reported in Table 1, the linear trend is significant, $F(1, 9) = 21.16$, $p = 0.001$, $MSe = 1.00$, partial $\eta^2 = 0.70$. The effect size, 0.70, is large. The data support the idea that students' spelling ability increases in a linear fashion.

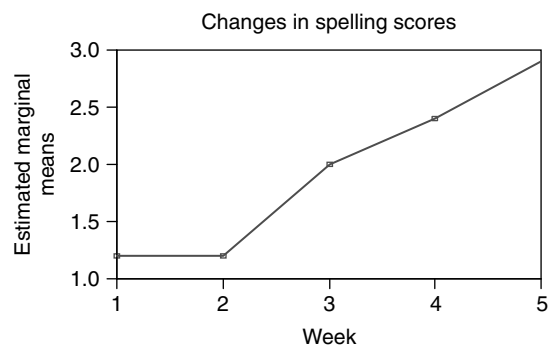


Figure 1 Trend for spelling scores across weeks

Table 1 Means and standard deviations for hypothetical (SD) spelling data

Trial	Phonics instruction M (SD)	Control M (SD)	Total M (SD)
Week 1	1.60 (0.70)	1.20 (0.92)	1.40 (0.82)
Week 2	3.00 (1.15)	1.20 (0.79)	2.10 (1.33)
Week 3	4.90 (1.66)	2.00 (1.15)	3.45 (2.04)
Week 4	6.10 (2.08)	2.40 (1.17)	4.25 (2.51)
Week 5	7.70 (2.26)	2.90 (1.45)	5.30 (3.08)

Table 2 Analysis of Variance for instruction level (I) and Trials (T)

Source of variation	SS	df	MS	F
Between subjects	322.20	19		
Instruction (I)	184.96	1	184.96	24.27 ^a
Students (s : I)	137.24	18	7.62	
Within subjects	302.8	80		
Trials (T)	199.50	4	49.87	79.15*
I × T	57.74	4	14.44	22.92*
T × s : I	45.56	72	0.63	
Total	625	99		

^a $p < 0.0001$.

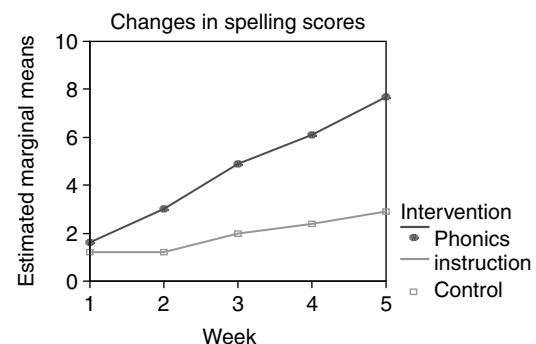
The Mixed-effects Model: Adding a Between-subjects Factor

To this point, the features of a repeated measures analysis of one factor (i.e., trials) have been presented. However, educational psychologists rarely test such simple models. Instead, they often study whether instructional interventions differ and whether the differences remain constant across time [5, 6]. Suppose an experimental variable is added to the repeated measures model for the hypothetical spelling data. Assume students are now randomly assigned to two conditions, a treatment condition that teaches students to spell using phonics-based instruction (Phonics) and a no instruction (Control) condition. Again, spelling scores are collected every Friday for five weeks. Means and standard deviations are reported in Table 1. Treatment (Phonics vs Control) is a between-subjects variable, Trials remains a within-subjects variable, and now there is an interaction between the two variables. In this analysis, students are nested in treatment condition. Table 2 presents the ANOVA source table for this model.

Three F-ratios are of interest. Researchers want to know if there is (a) a main effect for treatment; (b) a

main effect for trials; and, (c) an interaction between treatment and trials. In this analysis, the I (instruction mode) × T (trials) interaction is significant, $F(4, 72) = 22.92$, $p < 0.0001$, $MSe = 0.63$, partial $\eta^2 = 0.55$ (a large effect) as is the main effect for treatment, $F(1, 18) = 24.27$, $p < 0.0001$, $MSe = 7.62$, partial $\eta^2 = 0.55$, and trials, $F(4, 72) = 79.15$, $p < 0.0001$, $MSe = 0.63$, partial $\eta^2 = 0.55$. Statisticians recommend that researchers describe significant interactions before describing main effects because main effects for the first factor do not generalize over levels of the second factor (*see Interaction Effects*). Thus, even though the F-ratios are large for both main effects tests (i.e., treatment and trials), differences between treatment groups are not consistently the same for every weekly trial. Similarly, growth patterns across weekly trials are not similar for both treatment conditions.

Figure 2 illustrates the interaction visually. It displays two trend lines (Phonics and Control). Linear trends for both groups are significant. However, the slope for the Phonics Instruction group is steeper than that observed for the Control Group resulting in a significant I × T linear trend interaction, $F(1, 18) = 34.99$, $p < 0.0001$, $MSe = 1.64$, partial $\eta^2 = 0.56$. Descriptively, the results support the idea that increases in spelling ability over time are larger for the Phonics Instruction group than they are for the Control group. In fact, while hypothetical data are summarized here to show that Phonics Instruction has an effect on spelling performance when compared with a control condition, the results reflect those reported by researchers in educational psychology [24].

**Figure 2** Trends for spelling scores across weeks by treatment group

Assumption Considerations

The same statistical assumptions considered earlier apply to this mixed model. Thus, sphericity is still important. Because of the inclusion of two mutually exclusive groups, the assumption is called *multisample sphericity*. Not only must the variances of the differences of scores for all trial pairs be equal within a group but also the variances of the differences for all trial pairs must be consistent across groups. In the data analyzed, the assumption was not met. When analyzed by SPSS [22], adjustments of the degrees of freedom are recommended. For example, the uncorrected degrees of freedom for the interaction in Table 2 are 4 and 72, but the Greenhouse–Geisser degrees of freedom are 2.083 for the numerator and 37.488 for the denominator.

New Developments in Repeated Measures Analysis

The hypothetical spelling examples assume that students are randomly selected from the population of interest, and that they are randomly assigned to treatment conditions. The design in the second study was balanced as there were 10 students in each instruction level. Data were recorded for each student in a condition for every weekly trial. There were no missing data. Statistically, the design features of the examples are ideal.

Realistically, however, educational psychologists encounter design problems because data are not so ideal. Schools are not places where it is easy to randomly select students or randomly assign students to treatment conditions. Even if a special classroom or academic laboratory is available to the researcher, students interact, and converse regularly about their academic and social experiences. Further, instructors teach spelling independently of what they might learn from a Phonics Instruction intervention. From a statistical perspective, these data distributions likely violate the assumption of independence of observations.

To overcome this problem, researchers use several procedures. First, they can randomly select larger units of analysis such as classrooms, schools, or districts. Students, of course, would be nested in these larger units. Districts or schools would then be randomly assigned to treatment condition. Hierarchical Linear Modeling (HLM) [1], an extension of

the GLM, allows researchers to test multilevel models (see **Linear Multilevel Models**) where nested units of analysis are of interest to the researcher. Bryk and Raudenbush [1] show how these models can incorporate repeated measures factors and result in **growth curve modeling**. Thus, it is possible to analyze trends in Figures 1 and 2 with consideration of nested effects (e.g., how teachers in specific classrooms affect students' development, or how school districts affect achievement – independently of the effects of a treatment).

While HLM models help overcome violations of the assumption of independence, and, while they can incorporate repeated measures designs, they have limitations. Specifically, statistical **power** is an important consideration when selecting adequate sample sizes to test treatment effects. One must recall that these sample sizes now represent the highest level of the nested unit. As such, samples of districts, schools, or classrooms are selected rather than students who may be enrolled in one or two schools. Researchers, therefore, must consider the feasibility of their designs in terms of time and personnel needed to conduct investigations at this level.

Additionally, educational psychologists must consider the number of variables studied. In the hypothetical example, spelling is a simple construct representing one general kind of ability. Recent publications reveal constructs that are complex and multidimensional. Further, because the constructs are psychological traits, they are **latent variables** [2] rather than manifest variables such as hours, height, speed, and weight. Typically, latent variables are judged by how they predict scores on multiple-choice or rating scale items. Data reduction techniques like **exploratory factor analysis** (EFA) or **confirmatory factor analysis** (CFA) are then used to establish construct validity of scales. Educational psychologists then study the stability of these latent variables over time.

Guay, Marsh, and Boivin [10], for example, analyzed the relationships between academic self-concept and academic achievement of elementary students over time using a form of repeated measures testing **Structural Equation Models** (SEM) [19, 20]. These models allow researchers to inspect relationships between multiple dependent variables (either latent or manifest) as they are predicted from a set of independent variables. Further, the relationships between independent and dependent variables

can be examined in waves. That is, construct relationships can be examined for stability over time.

McDonald and Ho [20] provide a good resource of recommended practices in testing SEMs. Results in testing SEMs are best when (a) researchers outline a very good theory for how constructs are related and how they will change; (b) several exploratory studies guide the models' structure; and, (c) sample sizes are sufficient to ensure that the estimates of model parameters are stable.

A Summary of Statistical Techniques Used by Educational Psychologists

Unfortunately, educational psychologists have not relied extensively on HLM and SEM analyses with repeated measures of manifest or latent variables to test their hypotheses. A review of 116 studies in *Contemporary Educational Psychology* and *Journal of Educational Psychology* between 2001 and 2003 showed that 48 articles included tests of models with within-subjects factors. Of these, only four studies tested HLM models with repeated measures with at least two time periods. Of these four studies, only one incorporated growth curve modeling to examine differences in student performance over multiple time periods. As for SEM, eight studies tested multiple waves of repeated and/or related measures. Perhaps the lack of HLM and SEM models may be due to sample size limitations. HLM and SEM model parameters are estimated using **maximum likelihood** procedures. Maximum likelihood procedures require large sample sizes for estimation [10]. Alternatively, HLM and SEM research may not be prevalent, since the investigations of educational psychologists are often exploratory in nature. Thus, researchers may not be ready to confirm relationships or effects using HLM or SEM models [19, 20].

Certainly, the exploratory investigations could help the researchers replicate studies that eventually lead to test of HLM or SEM models where relationships or effects can be confirmed. Additionally, because educational psychologists are interested in individual differences, more studies should examine changes in latent and manifest variables at the student level for multiple time points [2]. **Time series analysis** is one statistical technique that educational psychologists can use to study developmental changes within individuals. Only 3 of the 168 studies reported use of time series analyses.

Table 3 Within-subjects methods in contemporary educational psychology research (2001–2003)

Technique	Frequency
Repeated-measures ANOVA	11
Related-measures ANOVA	11
Reporting effect sizes	9
SEM	8
Regression	5
MANOVA for repeated measures	4
Nonparametric repeated measures	4
HLM	4
Time series	3
Testing of assumptions	2

Note: Frequencies represent the number of articles that used a specific technique. Some articles reported more than one technique.

Table 3 lists the methods used in the 48 within-subjects design studies. As can be seen, only two studies addressed the assumptions of repeated measures designs (e.g., normality of distributions, independence of observations, sphericity). Only 9 of the 48 reported effect sizes. In a few instances, investigators used nonparametric procedures when their scales of measurement were nominal or ordinal, or when their data violated normality assumptions.

As long as researchers in educational psychology are interested in change and development, repeated measures analysis will continue to be needed to answer their empirical questions. Methodologists should ensure that important assumptions such as sphericity and independence of observations are met. Finally, there have been many recent developments in repeated measures techniques. HLM and SEM procedures can be used to study complex variables, their relationships, and how these relationships change in time. Additionally, time series analyses are recommended for examination of within-subject changes, especially for variables that theoretically should not remain stable over time (e.g., anxiety, situational interest, selective attention). As with all quantitative research, sound theory, quality measurement, adequate sampling, and careful consideration of experimental design factors help investigators contribute useful and lasting information to their field.

References

- [1] Bryk, A. & Raudenbush, S.W. (1992). *Hierarchical Linear Models for Social and Behavioral Research:*

- Applications and Data Analysis Methods*, Sage, Newbury Park.
- [2] Boorsboom, D., Mellenbergh, G.J. & van Heerden, J. (2003). The theoretical status of latent variables, *Psychological Bulletin* **110**(2), 203–219.
 - [3] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 3rd Edition, Academic Press, New York.
 - [4] Compton, D.L. (2003). Modeling the relationship between growth in rapid naming speed and growth in decoding skill in first-grade children, *Journal of Educational Psychology* **95**(2), 225–239.
 - [5] Desoete, A., Roeyers, H. & DeClercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology* **91**(1), 188–200.
 - [6] Gaskill, P.J. & Murphy, P.K. (2004). Effects of a memory strategy on second-graders' performance and self-efficacy, *Contemporary Educational Psychology* **29**(1), 27–49.
 - [7] Glass, G.V. & Hopkins, K.D. (1996). *Statistical Methods in Education and Psychology*, 3rd Edition, Allyn & Bacon, Boston.
 - [8] Green, L., McCutchen, D., Schwiebert, C., Quinlan, T., Eva-Wood, A. & Juelis, J. (2003). Morphological development in children's writing, *Journal of Educational Psychology* **95**(4), 752–761.
 - [9] Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data, *Psychometrika* **24**, 95–112.
 - [10] Guay, F., Marsh, H.W. & Boivin, M. (2003). Academic self-concept and academic achievement: developmental perspectives on their causal ordering, *Journal of Educational Psychology* **95**(1), 124–136.
 - [11] Howell, D.C. (2002). *Statistical Methods for Psychology*, Duxbury Press, Belmont.
 - [12] Huynh, H. & Feldt, L. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F distributions, *Journal of the American Statistical Association* **65**, 1582–1589.
 - [13] Keselman, H.J., Algina, J. & Kowalchuk, R.K. (2002). A comparison of data analysis strategies for testing omnibus effects in higher-order repeated measures designs, *Multivariate Behavioral Research* **37**(3), 331–357.
 - [14] Keselman, H.J., Keselman, J.C. & Shaffer, J.P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity, *Psychological Bulletin* **110**(1), 162–170.
 - [15] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole Publishing, Monterey.
 - [16] Kowalchuk, R.K. & Keselman, H.J. (2001). Mixed-model pairwise comparisons of repeated measures means, *Psychological Methods* **6**(3), 282–296.
 - [17] Lumley, M.A. & Provenzano, K.M. (2003). Stress management through written emotional disclosure improves academic performance among college students with physical symptoms, *Journal of Educational Psychology* **95**(3), 641–649.
 - [18] Maxwell, S.E. (1980). Pairwise multiple comparisons in repeated measures designs, *Journal of Educational Statistics* **5**, 269–287.
 - [19] McDonald, R.P. (1999). *Test Theory: A Unified Treatment*, Lawrence Erlbaum Associates, Mahwah.
 - [20] McDonald, R.P. & Ho, M.-H.R. (2002). Principles and practice in reporting structural equation analyses, *Psychological Methods* **7**(1), 64–82.
 - [21] Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th Edition, Irwin, Chicago.
 - [22] SPSS Inc. (2001). *SPSS for Windows 11.0.1*, SPSS Inc., Chicago.
 - [23] Troia, G.A. & Whitney, S.D. (2003). A close look at the efficacy of Fast ForWord Language for children with academic weaknesses, *Contemporary Educational Psychology* **28**, 465–494.
 - [24] Vandervelden, M. & Seigel, L. (1997). Teaching phonological processing skills in early literacy: a developmental approach, *Learning Disability Quarterly* **20**, 63–81.

(See also **Multilevel and SEM Approaches to Growth Curve Modeling**)

JONNA M. KULIKOWICH

Effect Size Measures

ROGER E. KIRK

Volume 2, pp. 532–542

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Effect Size Measures

Measures of effect magnitude help researchers determine if research results are important, valuable, or useful. A variety of statistics are used to measure effect magnitude. Many of the statistics fall into one of two categories: measures of effect size (typically, standardized mean differences) and measures of strength of association. In addition, there is a large group of statistics that do not fit into either category. A partial listing of effect magnitude statistics is given in Table 1.

The statistics play an important role in behavioral research. They are used to (a) estimate the sample size required to achieve an acceptable **power**, (b) integrate the results of empirical research studies in **meta-analyses**, (c) supplement the information provided by null-hypothesis significance tests, and (d) determine whether research results are practically significant [47]. Practical significance is concerned with the usefulness of results. Statistical significance, the focus of null-hypothesis significance tests, is concerned with whether results are due to chance or sampling variability.

Limitations of Null-hypothesis Significance Testing

To appreciate the role that measures of effect magnitude play in the research enterprise, it is necessary to understand the limitations of classical null-hypothesis significance testing. Null-hypothesis significance testing procedures, as practiced today, were developed between 1915 and 1933 by three men: **Ronald A. Fisher** (1890–1962), **Jerzy Neyman** (1894–1981), and **Egon S. Pearson** (1895–1980). Fisher was primarily responsible for the new paradigm and for advocating 0.05 as the standard significance level [51]. **Cohen** [13, p. 1307] observed that, ‘The fact that Fisher’s ideas quickly became the basis for statistical inference in the behavioral sciences is not surprising—they were very attractive. They offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes-no decisions.’ In spite of these apparent advantages, the logic and usefulness of null-hypothesis significance testing has been debated for over 70 years. One of

the earliest negative critiques appeared in a 1938 article by Joseph Berkson [4]. Since then, the number of negative critiques has escalated [2, 7, 8, 13, 15, 23, 39, 55, 68, 71, 73]. One frequent criticism of null-hypothesis significance testing is that it answers the wrong question [3, 7, 15, 19, 22]. For example, in scientific inference, what researchers want to know is the probability that the null hypothesis (H_0) is true, given that they have obtained a set of data (D); that is, $p(H_0|D)$. What null-hypothesis significance testing tells them is the probability of obtaining these data or more extreme data if the null hypothesis is true, $p(D|H_0)$. Unfortunately for researchers, obtaining data for which $p(D|H_0)$ is low does not imply that $p(H_0|D)$ also is low. Falk [22] pointed out that $p(D|H_0)$ and $p(H_0|D)$ can be equal, but only under rare mathematical conditions.

Another criticism is that null-hypothesis significance testing is a trivial exercise because all null hypotheses are false. **John Tukey** [79, p. 100] expressed this idea as follows: ‘the effects of A and B are always different—in some decimal place—for any A and B. Thus asking “Are the effects different?” is foolish.’ More recently, Jones and Tukey [40, p. 412] reiterated this view.

For large, finite, treatment populations, a total census is at least conceivable, and we cannot imagine an outcome for which $\mu_A - \mu_B = 0$ when the dependent variable (or any other variable) is measured to an indefinitely large number of decimal places. . . . The population mean difference may be trivially small but will always be positive or negative.

The view that all null hypotheses are false, except those we construct, for Monte Carlo tests of statistical procedures is shared by many researchers [2, 4, 13, 34, 77]. Hence, because Type I errors cannot occur, statistically significant results are assured if large enough samples are used. Thompson [77, p. 799] captured the essence of this view when he wrote, ‘Statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we’ve been too lazy to drag in enough participants.’

A third criticism of null-hypothesis significance testing is that by adopting a fixed level of significance, a researcher turns a continuum of uncertainty into a dichotomous reject-do not reject decision [25, 30, 67, 81]. A null-hypothesis significance test provides a probability, called a *P* value, of observing a research result given that the null hypothesis is true.

2 Effect Size Measures

Table 1 Measures of effect magnitude

Measures of effect size	Measures of strength of association	Other measures
Cohen's [12] d, f, g, h, q, w	$r, r_{pb}, r_s, r^2, R, R^2, \eta, \eta^2, \eta_{mult}^2, \phi, \phi^2$	Absolute risk reduction (ARR)
Glass's [28] g'	Chambers' [9] r_e	Cliff's [10] p
Hedges's [36] g	Cohen's [12] f^2	Cohen's [12] U_1, U_2, U_3
Mahalanobis's D	Contingency coefficient (C)	Doksum's [20] shift function
Mean ₁ – Mean ₂	Cramér's [17] V	Dunlap's [21] common language effect size for bivariate correlation (CL _R)
Median ₁ – Median ₂	Fisher's [24] Z	Grisson's [31] probability of superiority (PS)
Mode ₁ – Mode ₂	Friedman's [26] r_m	Logit d'
Rosenthal and Rubin's [62] Π	Goodman & Kruskal's [29] λ, γ	McGraw & Wong's [54] common language effect size (CL)
Tang's [75] ϕ	Hays's [35] $\hat{\omega}^2, \hat{\omega}_{Y A,B,AB}^2, \hat{\rho}_1, \hat{\rho}_2$	Odds ratio ($\hat{\omega}$)
Thompson's [78] d^*	$\hat{\rho}_{Y A,B,AB}$	Preece's [59] ratio of success rates
Wilcox's [82] $\hat{\Lambda}_{Mdn, \sigma_b}$	Herzberg's [38] R^2	Probit d'
Wilcox & Muska's [84] $\hat{Q}_{0.632}$	Kelley's [42] ε^2	Relative risk (RR)
	Kendall's [43] W	Risk difference
	Lord's [53] R^2	Sánchez-Meca, Marín-Martínez, & Chacón-Moscó [69] d_{Cox}
	Olejnik & Algina [58] $\hat{\omega}_G^2, \hat{\eta}_G^2$	Rosenthal and Rubin's [61] binomial effect size display (BESD)
	Rosenthal and Rubin's [64] $r_{equivalent}^2$	Rosenthal & Rubin's [63] counternull value of an effect size ($ES_{counternull}$)
	Rosnow, Rosenthal, & Rubin's [66] $r_{alerting}, r_{contrast}, r_{effect\ size}$	Wilcox's [82] probability of superiority (λ)
	Tatsuoka's [76] $\hat{\omega}_{mult. c}^2$	
	Wherry's [80] R^2	

A P value only slightly larger than the level of significance is treated the same as a much larger P value. The adoption of 0.05 as the dividing point between significance and nonsignificance is quite arbitrary. The comment by Rosnow and Rosenthal [65] is pertinent, 'surely, God loves the 0.06 nearly as much as the 0.05'.

A fourth criticism of null-hypothesis significance testing is that it does not address the question of whether results are important, valuable, or useful, that is, their practical significance. The fifth edition of the *Publication Manual of the American Psychological Association* [1, pp. 25–26] explicitly recognizes this limitation of null-hypothesis significance tests.

Neither of the two types of probability value (significance level and P value) directly reflects the magnitude of an effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section.

Researchers want to answer three basic questions from their research [48]: (a) Is an observed effect real or should it be attributed to chance? (b) If the effect is real, how large is it? and (c) Is the effect

large enough to be useful; that is, is it practically significant? As noted earlier, null-hypothesis significance tests only address the first question. Descriptive statistics, measures of effect magnitude, and **Confidence Intervals** address the second question and provide a basis for answering the third question. Answering the third question, is the effect large enough to be useful or practically significant?, calls for a judgment. The judgment is influenced by a variety of considerations including the researcher's value system, societal concerns, assessment of costs and benefits, and so on. One point is evident, statistical significance and practical significance address different questions. Researchers should follow the advice of the *Publication Manual of the American Psychological Association* [1], 'provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship' (pp. 25–26). In the following sections, a variety of measures of effect magnitude are described that can help a researcher assess the practical significance of research results.

Effect Size

In 1969, Cohen introduced the first effect size measure that was explicitly labeled as such. His measure

is given by,

$$\delta = \frac{\psi}{\sigma} = \frac{\mu_E - \mu_C}{\sigma}, \quad (1)$$

where μ_E and μ_C denote the population means of the experimental and control groups and σ denotes the common population standard deviation [11]. Cohen recognized that the size of the contrast $\psi = \mu_E - \mu_C$ is influenced by the scale of measurement of the means. He divided the contrast by σ to rescale the contrast in units of the amount of error variability in the data. Rescaling is useful when the measurement units are arbitrary or have no inherent meaning. Rescaling also can be useful in performing power and sample size computations and in comparing effect sizes across research literatures involving diverse, idiosyncratic measurement scales. However, rescaling serves no purpose when a variable is always measured on a standard scale. Safety experts, for example, always measure contrasts of driver reaction times in seconds and pediatricians always measure contrasts of birth weights in pounds and ounces [5].

For nonstandard scales, Cohen's contribution is significant because he provided guidelines for interpreting the magnitude of δ . He said that $\delta = 0.2$ is a small effect, $\delta = 0.5$ is a medium effect, and $\delta = 0.8$ is a large effect. According to Cohen [14], a medium effect of 0.5 is visible to the naked eye of a careful observer. A small effect of 0.2 is noticeably smaller than medium, but not so small as to be trivial. Only an expert would be able to detect a small effect. A large effect of 0.8 is the same distance above medium as small is below it. A large effect would be obvious to anyone. From another perspective, a small effect is one for which 58% of participants in one group exceed the mean of participants in another group. A medium effect is one for which 69% of participants in one group exceed the mean of another group. And finally, a large effect is one for which 79% of participants exceed the mean of another group.

By assigning numbers to the labels small, medium, and large, Cohen provided researchers with guidelines for interpreting the size of treatment effects. His effect size measure is a valuable supplement to the information provided by a P value. A P value of 0.0001 loses its luster if the effect turns out to be trivial.

In most research settings, the parameters of Cohen's δ are unknown. In such cases, the sample means of the experimental and control groups are used to estimate μ_E and μ_C . An estimator of σ can

be obtained in a number of ways. Under the assumption that σ_E and σ_C are equal, the sample variances of the experimental and control groups are pooled as follows:

$$\hat{\sigma}_{\text{Pooled}} = \sqrt{\frac{(n_E - 1)\hat{\sigma}_E^2 + (n_C - 1)\hat{\sigma}_C^2}{(n_E - 1) + (n_C - 1)}}, \quad (2)$$

where n_E and n_C denote respectively the sample size of the experimental and control groups. An estimator of δ is

$$d = \frac{\bar{Y}_E - \bar{Y}_C}{\hat{\sigma}_{\text{Pooled}}}, \quad (3)$$

where $\bar{Y}_E$ and $\bar{Y}_C$ denote respectively the sample mean of the experimental and control groups, and $\hat{\sigma}_{\text{Pooled}}$ denotes the pooled estimator of σ . Gene Glass's [28] pioneering work on **meta-analysis** led him to recommend a different approach to estimating σ . He reasoned that if there were several experimental groups and a control group, pairwise pooling of the sample standard deviation of each experimental group with that of the control group could result in different values of $\hat{\sigma}_{\text{Pooled}}$ for the various contrasts. Hence, when the standard deviations of the experimental groups differed, the same size difference between experimental and control means would result in different effect sizes. Glass's solution was to use the sample standard deviation of the control group, $\hat{\sigma}_C$, to estimate σ . Glass's estimator of δ is

$$g' = \frac{\bar{Y}_{E_j} - \bar{Y}_C}{\hat{\sigma}_C}, \quad (4)$$

where $\bar{Y}_{E_j}$ and $\bar{Y}_C$ denote respectively the sample mean of the j th experimental group and the sample mean of the control group. Larry Hedges [36] recommended yet another approach to estimating σ . He observed that population variances are often homogeneous, in which case the most precise estimate of the population variance is obtained by pooling the $j = 1, \dots, p$ sample variances. His pooled estimator,

$$\hat{\sigma}_{\text{Pooled}} = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + \dots + (n_p - 1)\hat{\sigma}_p^2}{(n_1 - 1) + \dots + (n_p - 1)}}, \quad (5)$$

is identical to the square root of the within-groups mean square in a completely randomized analysis of variance. Hedges' estimator of δ is

$$g = \frac{\bar{Y}_{E_j} - \bar{Y}_C}{\hat{\sigma}_{\text{Pooled}}}. \quad (6)$$

4 Effect Size Measures

Hedges [36] observed that all three estimators of δ — d , g' , and g —are biased. He recommended correcting g for bias as follows,

$$g_c = J(N - 2)g, \quad (7)$$

where $J(N - 2)$ is the bias correction factor described in Hedges and Olkin [37]. The correction factor is approximately

$$J(N - 2) \cong \left(1 - \frac{3}{4N - 9}\right), \quad (8)$$

where $N = n_E + n_C$. Hedges [36] showed that g_c is the unique, uniformly minimum variance-unbiased estimator of δ . He also described an approximate confidence interval for δ :

$$g_c - z_{\alpha/2}\hat{\sigma}(g_c) \leq \delta \leq g_c + z_{\alpha/2}\hat{\sigma}(g_c),$$

where $z_{\alpha/2}$ denotes the two-tailed critical value that cuts off the upper $\alpha/2$ region of the standard normal distribution and

$$\hat{\sigma}(g_c) = \sqrt{\frac{n_E + n_C}{n_E n_C} + \frac{g_c^2}{2(n_E + n_C)}}. \quad (9)$$

Procedures for obtaining exact confidence intervals for δ using noncentral sampling distributions are described by Cumming and Finch [18].

Cohen's δ has been widely embraced by researchers because (a) it is easy to understand and interpret across different research studies, (b) the sampling distributions of estimators of δ are well understood, and (c) estimators of δ can be easily computed from t statistics and F statistics with one-degree-of-freedom that are reported in published articles. The latter feature is particularly attractive to researchers who do meta-analyses.

The correct way to conceptualize and compute the denominator of δ can be problematic when the treatment is a classification or organismic variable [27, 32, 57]. For experiments with a manipulated treatment and random assignment of the treatment levels to participants, the computation of an effect size such as g_c is relatively straightforward. The denominator of g_c is the square root of the within-groups mean square. This mean square provides an estimate of σ that reflects the variability of observations for the full range of the manipulated treatment. However, when the treatment is an organismic variable, such as gender, boys and girls, the square root of the

within-groups mean square may not reflect the variability for the full range of the treatment because it is a pooled measure of the variation of boys alone and the variation of girls alone. If there is a gender effect, the within-groups mean square reflects the variation for a partial range of the gender variable. The variation for the full range of the gender variable is given by the total mean square and will be larger than the within-groups mean square. Effect sizes should be comparable across different kinds of treatments and experimental designs. In the gender experiment, use of the square root of the total mean square to estimate σ gives an effect size that is comparable to those for treatments that are manipulated. The problem of estimating σ is exacerbated when the experiment has several treatments, repeated measures, and covariates. Gillett [27] and Olejnik and Algina [57] provide guidelines for computing effect sizes for such designs.

There are other problems with estimators of δ . For example, the three estimators, d , g' , and g , assume normality and a common standard deviation. Unfortunately, the value of the estimators is greatly affected by heavy-tailed distributions and heterogeneous standard deviations [82]. Considerable research has focused on ways to deal with these problems [6, 44, 49, 50, 82, 83]. Some solutions attempt to improve the estimation of δ , other solutions call for radically different ways of conceptualizing effect magnitude. In the next section, measures that represent the proportion of variance in the dependent variable that is explained by the variance in the independent variable are described.

Strength of Association

Another way to supplement null-hypothesis significance tests and help researchers assess the practical significance of research results is to provide a measure of the strength of the association between the independent and dependent variables. A variety of measures of strength of association are described by Carroll and Nordholm [6] and Särndal [70]. Two popular measures are omega squared, denoted by ω^2 , for a fixed-effects (see **Fixed and Random Effects**) treatment and the **intraclass correlation** denoted by ρ_1 , for a random-effects (see **Fixed and Random Effects**) treatment. A fixed-effects treatment is one in which all treatment levels about which inferences

are to be drawn are included in the experiment. A random-effects treatment is one in which the p treatment levels in the experiment are a random sample from a much larger population of P levels. For a completely randomized analysis of variance design, omega squared and the intraclass correlation are defined as

$$\frac{\sigma_{\text{Treat}}^2}{\sigma_{\text{Treat}}^2 + \sigma_{\text{Error}}^2},$$

where σ_{Treat}^2 and σ_{Error}^2 denote respectively the treatment and error variance. Both omega squared and the intraclass correlation represent the proportion of the population variance in the dependent variable that is accounted for by specifying the treatment-level classification. The parameters σ_{Treat}^2 and σ_{Error}^2 for a completely randomized design are generally unknown, but they can be estimated from sample data. Estimators of ω^2 and ρ_I are respectively

$$\begin{aligned}\hat{\omega}^2 &= \frac{SS_{\text{Treat}} - (p-1)MS_{\text{Error}}}{SS_{\text{Total}} + MS_{\text{Error}}} \\ \hat{\rho}_I &= \frac{MS_{\text{Treat}} - MS_{\text{Error}}}{MS_{\text{Treat}} + (n-1)MS_{\text{Error}}},\end{aligned}\quad (10)$$

where SS denotes a sum of squares, MS denotes a mean square, p denotes the number of levels of the treatment, and n denotes the number of observations in each treatment level. Omega squared and the intraclass correlation are biased estimators because they are computed as the ratio of unbiased estimators. The ratio of unbiased estimators is, in general, not an unbiased estimator. Carroll and Nordholm (1975) showed that the degree of bias in $\hat{\omega}^2$ is slight.

The usefulness of Cohen's δ was enhanced because he suggested guidelines for its interpretation. On the basis of Cohen's [12] classic work, the following guidelines are suggested for interpreting omega squared:

$$\begin{aligned}\omega^2 = 0.010 &\text{ is a small association} \\ \omega^2 = 0.059 &\text{ is a medium association} \\ \omega^2 = 0.138 \text{ or larger} &\text{ is a large association.}\end{aligned}\quad (11)$$

According to Sedlmeier and Gigerenzer [72] and Cooper and Findley [16], the typical strength of association in the journals that they examined was around 0.06—a medium association.

Omega squared and the intraclass correlation, like the measures of effect size, have their critics. For

example, O'Grady [56] observed that $\hat{\omega}^2$ and $\hat{\rho}_I$ may underestimate the true proportion of explained variance. If, as is generally the case, the dependent variable is not perfectly reliable, measurement error will reduce the proportion of variance that can be explained. Years ago, Gulliksen [33] pointed out that the absolute value of the product-moment correlation coefficient, r_{XY} , cannot exceed $(r_{XX'})^{1/2} (r_{YY'})^{1/2}$, where $r_{XX'}$ and $r_{YY'}$ are the reliabilities of X and Y . O'Grady [56] also criticized omega squared and the intraclass correlation on the grounds that their value is affected by the choice and number of treatment levels. As the diversity and number of treatment levels increases, the value of measures of strength of association also tends to increase. Levin [52] criticized $\hat{\omega}^2$ on the grounds that it is not very informative when an experiment contains more than two treatment levels. A large value of $\hat{\omega}^2$ simply indicates that the dependent variable for at least one treatment level is substantially different from the other levels. As is true for all omnibus measures, $\hat{\omega}^2$ and $\hat{\rho}_I$ do not pinpoint which treatment level(s) is responsible for a large value.

The last criticism can be addressed by computing omega squared and the intraclass correlation for two-mean contrasts as is typically done with Hedges' g_c . This solution is in keeping with the preference of many researchers to ask focused one-degree-of-freedom questions of their data [41, 66] and the recommendation of the *Publication Manual of the American Psychological Association* [1, p. 26], 'As a general rule, multiple degree-of-freedom effect indicators tend to be less useful than effect indicators that decompose multiple degree-of-freedom tests into meaningful one-degree-of-freedom effects – particularly when these are the results that inform the discussion.'

The formulas for omega squared and the intraclass correlation can be modified to give the proportion of variance in the dependent variable that is accounted for by the i th contrast, $\hat{\psi}_i$. The formulas for a completely randomized design are

$$\begin{aligned}\hat{\omega}_{Y|\psi_i}^2 &= \frac{SS \hat{\psi}_i - MS_{\text{Error}}}{SS_{\text{Total}} + MS_{\text{Error}}} \\ \hat{\rho}_{IY|\psi_i} &= \frac{SS \hat{\psi}_i - MS_{\text{Error}}}{SS \hat{\psi}_i + (n-1)MS_{\text{Error}}},\end{aligned}\quad (12)$$

where $SS \hat{\psi}_i = \hat{\psi}_i^2 / \sum_{j=1}^p c_j^2 / n_j$ and the c_j s are coefficients that define the contrast [45]. These measures

6 Effect Size Measures

answer focused one-degree-of-freedom questions as opposed to omnibus questions about one's data.

To determine the strength of association in experiments with more than one treatment or experiments with a blocking variable, partial omega squared can be computed. A comparison of omega squared and partial omega squared for treatment A for a two-treatment, completely randomized factorial design is

$$\omega_{Y|A}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_{\text{Error}}^2}$$

and

$$\omega_{Y|A \cdot B, AB}^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{\text{Error}}^2}, \quad (13)$$

where partial omega squared ignores treatment B and the $A \times B$ interaction. If one or more of the variables in a multitreatment experiment is an organismic or blocking variable, Olejnik and Algina [58] show that partial omega squared is not comparable across different experimental designs. Furthermore, Cohen's guidelines for small, medium, and large effects are not applicable. They propose a measure of strength of association called *generalized omega squared*, denoted by ω_G^2 , that is appropriate, and they provide extensive formulas for its computation.

Meta-analysts often use the familiar product-moment correlation coefficient, r , to assess strength of association. The square of r called the *coefficient of determination* indicates the sample proportion of variance in the dependent variable that is accounted for by the independent variable. The product-moment

correlation and its close relatives can be used with a variety of variables:

Product-moment correlation	X and Y are continuous and linearly related
phi correlation, ϕ	X and Y are dichotomous
Point-biserial correlation, r_{pb}	X is dichotomous, Y is continuous
Spearman rank correlation, r_s	X and Y are in rank form.

The point-biserial correlation coefficient is particularly useful for answering focused questions. The independent variable is coded 0 and 1 to indicate the treatment level to which each observation belongs.

Two categories of measures of effect magnitude, measures of effect size and strength of association, have been described. Researchers are divided in their preferences for the two kinds of measures. As Table 2 shows, it is a simple matter to convert from one measure to another. Table 2 also gives formulas for converting the t statistic found in research reports into each of the measures of effect magnitude.

Other Measures of Effect Magnitude

Researchers continue to search for ways to supplement the null-hypothesis significance test and obtain a better understanding of their data. Their primary focus has been on measures of effect size and strength of association. But, as Table 1 shows, there are many other ways to measure effect magnitude. Some of the statistics in the 'Other measures' column of Table 1 are radically different from anything described thus

Table 2 Conversion formulas for four measures of effect magnitude

	t	g	r_{pb}	$\hat{\omega}_{Y \psi_i}^2$	$\hat{\rho}_{1 \psi_i}$
g	$t\sqrt{\frac{2}{n}}$		$\sqrt{\frac{r_{pb}^2 df 2n}{n^2(1-r_{pb}^2)}}$	$\sqrt{\frac{2(\hat{\omega}^2 - 2n\hat{\omega}^2 - 1)}{n(\hat{\omega}^2 - 1)}}$	$\sqrt{\frac{2(\hat{\rho}_1 - n\hat{\rho}_1 - 1)}{n(\hat{\rho}_1 - 1)}}$
r_{pb}	$\frac{t}{\sqrt{t^2 + df}}$	$\sqrt{\frac{g^2 n^2}{g^2 n^2 + df 2n}}$		$\sqrt{\frac{\hat{\omega}^2(1 - 2n) - 1}{\hat{\omega}^2(1 - 2n) + df(\hat{\omega}^2 - 1) - 1}}$	$\sqrt{\frac{\hat{\rho}_1(1 - n) - 1}{\hat{\rho}_1(1 - n) + df(\hat{\rho}_1 - 1) - 1}}$
$\hat{\omega}_{Y \psi_i}^2$	$\frac{t^2 - 1}{t^2 + 2n - 1}$	$\frac{ng^2 - 2}{ng^2 + 4n - 2}$	$\frac{r_{pb}^2(df + 1) - 1}{r_{pb}^2 df + (1 - r_{pb}^2)(2n - 1)}$		$\frac{\hat{\rho}_1}{2 - \hat{\rho}_1}$
$\hat{\rho}_{1 \psi_i}$	$\frac{t^2 - 1}{t^2 + n - 1}$	$\frac{ng^2 - 2}{ng^2 + 2n - 2}$	$\frac{r_{pb}^2(df + 1) - 1}{r_{pb}^2 df + (1 - r_{pb}^2)(n - 1)}$	$\frac{2\hat{\omega}^2}{\hat{\omega}^2 + 1}$	

far. One such measure for the two-group case is the *probability of superiority*, denoted by PS [31]. PS is the probability that a randomly sampled member of a population given one treatment level will have a score, Y_1 , that is superior to the score, Y_2 , of a randomly sampled member of another population given the other treatment level. The measure is easy to compute: $PS = U/n_1n_2$, where U is the Mann–Whitney statistic (see **Wilcoxon–Mann–Whitney Test**) and n_1 and n_2 are the two sample sizes. The value of U indicates the number of times that the n_1 participants who are given treatment level 1 have scores that outrank those of the n_2 participants who are given treatment level 2, assuming no ties or an equal allocation of ties. An unbiased estimator of the population $Pr(Y_1 > Y_2)$ is obtained by dividing U by n_1n_2 , the number of possible comparisons of the two treatment levels. An advantage of PS according to Grissom [31] is that it does not assume equal variances and is robust to nonnormality.

The **odds ratio** is another example of a different way of assessing effect magnitude. It is applicable to two-group experiments when the dependent variable has only two outcomes, say, success and failure. The term *odds* is frequently used by those who place bets on the outcomes of sporting events. The odds that an event will occur are given by the ratio of the probability that the event will occur to the probability that the event will not occur. If an event can occur with probability p , the odds in favor of the event are $p/(1 - p)$ to 1. For example, suppose an event occurs with probability $3/4$, the odds in favor of the event are $(3/4)/(1 - 3/4) = (3/4)/(1/4) = 3$ to 1.

The computation of the odds ratio is illustrated using the data in Table 3 where the performance of participants in the experimental and control groups is classified as either a success or a failure. For participants in the experimental group, the odds of success are

$$Odds(\text{Success}|\text{Exp. Grp.}) = \frac{n_{11}/(n_{11} + n_{12})}{n_{21}/(n_{21} + n_{22})}$$

$$= \frac{n_{11}}{n_{12}} = \frac{43}{7} = 6.1429. \quad (14)$$

For participants in the control group, the odds of success are

$$\begin{aligned} Odds(\text{Success}|\text{Control Grp.}) &= \frac{n_{21}/(n_{21} + n_{22})}{n_{22}/(n_{21} + n_{22})} \\ &= \frac{n_{21}}{n_{22}} = \frac{27}{23} = 1.1739. \end{aligned} \quad (15)$$

The ratio of the two odds is the odds ratio, $\hat{\omega}$,

$$\begin{aligned} \hat{\omega} &= \frac{Odds(\text{Success}|\text{Exp. Grp.})}{Odds(\text{Success}|\text{Control Grp.})} \\ &= \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 5.233. \end{aligned} \quad (16)$$

In this example, the odds of success for participants in the experiment group are approximately 5 times greater than the odds of success for participants in the control group. When there is no difference between the groups in terms of odds of success, the two rows (or two columns) are proportional to each other and $\hat{\omega} = 1$. The more the groups differ, the more $\hat{\omega}$ departs from 1. A value of $\hat{\omega}$ less than 1 indicates reduced odds of success for the experimental participants; a value greater than 1 indicates increased odds of success for the experimental participants. The lower bound for $\hat{\omega}$ is 0 and occurs when $n_{11} = 0$; the upper bound is arbitrarily large, in effect infinite, and occurs when $n_{21} = 0$.

The probability distribution of the odds ratio is positively skewed. In contrast, the probability distribution of the natural log of $\hat{\omega}$, $\ln \hat{\omega}$, is more symmetrical. Hence, when calculating a confidence interval for $\hat{\omega}$, it is customary to use $\ln \hat{\omega}$ instead of $\hat{\omega}$. A $100(1 - \alpha)$ confidence interval for $\ln \omega$ is given by

$$\ln \hat{\omega} - z_{\alpha/2} \hat{\sigma}_{\ln \hat{\omega}} < \ln \omega < \ln \hat{\omega} + z_{\alpha/2} \hat{\sigma}_{\ln \hat{\omega}},$$

Table 3 Classification of participants

	Success	Failure	Total
Experimental group	$n_{11} = 43$	$n_{12} = 7$	$n_{11} + n_{12} = 50$
Control group	$n_{21} = 27$	$n_{22} = 23$	$n_{21} + n_{22} = 50$
Total	$n_{11} + n_{21} = 70$	$n_{12} + n_{22} = 30$	

8 Effect Size Measures

where $z_{\alpha/2}$ denotes the two-tailed critical value that cuts off the upper $\alpha/2$ region of the standard normal distribution and $\hat{\sigma}_{\ln \hat{\omega}}$ denotes the standard error of $\ln \hat{\omega}$ and is given by

$$\hat{\sigma}_{\ln \hat{\omega}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \quad (17)$$

Once the lower and upper bounds of the confidence interval are found, the values are exponentiated to find the confidence interval for ω . The computation will be illustrated for the data in Table 3 where $\hat{\omega} = 5.233$. For $\hat{\omega} = 5.233$, $\ln \hat{\omega} = 1.6550$. A 100(1 - 0.05)% confidence interval for $\ln \omega$ is

$$\begin{aligned} 1.6550 - 1.96(0.4966) &< \ln \omega \\ &< 1.6550 + 1.96(0.4966) \\ 0.6817 &< \ln \omega < 2.6283. \end{aligned}$$

The confidence interval for ω is

$$\begin{aligned} e^{0.6817} &< \omega < e^{2.6283} \\ 2.0 &< \omega < 13.9 \end{aligned}$$

The researcher can be 95% confident that the odds of success for participants in the experiment group are between 2.0 and 13.9 times greater than the odds of success for participants in the control group. Notice that the interval does not include 1. The odds ratio is widely used in the medical sciences, but less often in the behavioral and social sciences. Table 1 provides references for a variety of other measures of effect magnitude. Space limitations preclude an examination of other potentially useful measures of effect magnitude.

From the foregoing, the reader may have gotten the impression that small effect magnitudes are never or rarely ever important or useful. This is not true. Prentice and Miller [60] and Spencer [74] provide numerous examples in which small effect magnitudes are both theoretically and practically significant. The assessment of practical significance always involves a judgment in which a researcher must calibrate the magnitude of an effect by the benefit possibly accrued from that effect [46].

References

- [1] American Psychological Association. (2001). *Publication Manual of the American Psychological Association*, 5th Edition, American Psychological Association, Washington.
- [2] Bakan, D. (1966). The test of significance in psychological research, *Psychological Bulletin* **66**, 423–37.
- [3] Berger, J.O. & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity, *American Scientist* **76**, 159–65.
- [4] Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test, *Journal of the American Statistical Association* **33**, 526–542.
- [5] Bond, C.F., Wiitala, W.L. & Richard, F.D. (2003). Meta-analysis of raw mean scores, *Psychological Methods* **8**, 406–418.
- [6] Carroll, R.M. & Nordholm, L.A. (1975). Sampling characteristics of Kelley's ϵ^2 and Hays' ω^2 , *Educational and Psychological Measurement* **35**, 541–554.
- [7] Carver, R.P. (1978). The case against statistical significance testing, *Harvard Educational Review* **48**, 378–399.
- [8] Carver, R.P. (1993). The case against statistical significance testing, revisited, *Journal of Experimental Education* **61**, 287–292.
- [9] Chambers, R.C. (1982). Correlation coefficients from 2×2 tables and from biserial data, *British Journal of Mathematical and Statistical Psychology* **35**, 216–227.
- [10] Cliff, N. (1993). Dominance statistics: ordinal analyses to answer ordinal questions, *Psychological Bulletin* **114**, 494–509.
- [11] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York.
- [12] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.
- [13] Cohen, J. (1990). Things I have learned (so far), *American Psychologist* **45**, 1304–1312.
- [14] Cohen, J. (1992). A power primer, *Psychological Bulletin* **112**, 115–159.
- [15] Cohen, J. (1994). The earth is round ($p < 0.05$), *American Psychologist* **49**, 997–1003.
- [16] Cooper, H. & Findley, M. (1982). Expected effect sizes: estimates for statistical power analysis in social psychology, *Personality and Social Psychology Bulletin* **8**, 168–173.
- [17] Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton.
- [18] Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions, *Educational and Psychological Measurement* **61**, 532–574.
- [19] Dawes, R.M., Mirels, H.L., Gold, E. & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments, *Psychological Science* **4**, 396–400.
- [20] Doksum, K.A. (1977). Some graphical methods in statistics. A review and some extensions, *Statistica Neerlandica* **31**, 53–68.
- [21] Dunlap, W.P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations, *Psychological Bulletin* **116**, 509–511.
- [22] Falk, R. (1998). Replication—A step in the right direction, *Theory and Psychology* **8**, 313–321.

- [23] Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology* **5**, 75–98.
- [24] Fisher, R.A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample, *Metron* **1**, 1–32.
- [25] Frick, R.W. (1996). The appropriate use of null hypothesis testing, *Psychological Methods* **1**, 379–390.
- [26] Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation, *Psychological Bulletin* **70**, 245–251.
- [27] Gillett, R. (2003). The comparability of meta-analytic effect-size estimators from factorial designs, *Psychological Methods* **8**, 419–433.
- [28] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.
- [29] Goodman, L.A. & Kruskal, W.H. (1954). Measures of association for cross classification, *Journal of the American Statistical Association* **49**, 732–764.
- [30] Grant, D.A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models, *Psychological Review* **69**, 54–61.
- [31] Grissom, R.J. (1994). Probability of the superior outcome of one treatment over another, *Journal of Applied Psychology* **79**, 314–316.
- [32] Grissom, R.J. & Kim, J.J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size, *Psychological Methods* **6**, 135–146.
- [33] Gulliksen, H. (1950). *Theory of Mental Tests*, Wiley, New York.
- [34] Harris, R.J. (1994). *ANOVA: An Analysis of Variance Primer*, F. E. Peacock Publishers, Itasca.
- [35] Hays, W.L. (1963). *Statistics for Psychologists*, Holt Rinehart Winston, New York.
- [36] Hedges, L.V. (1981). Distributional theory for Glass’s estimator of effect size and related estimators, *Journal of Educational Statistics* **6**, 107–128.
- [37] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, Orlando.
- [38] Herzberg, P.A. (1969). The parameters of cross-validation, *Psychometrika Monograph Supplement* **16**, 1–67.
- [39] Hunter, J.E. (1997). Needed: a ban on the significance test, *Psychological Science* **8**, 3–7.
- [40] Jones, L.V. & Tukey, J.W. (2000). A sensible formulation of the significance test, *Psychological Methods* **5**, 411–414.
- [41] Judd, C.M., McClelland, G.H. & Culhane, S.E. (1995). Data analysis: continuing issues in the everyday analysis of psychological data, *Annual Reviews of Psychology* **46**, 433–465.
- [42] Kelley, T.L. (1935). An unbiased correlation ratio measure, *Proceedings of the National Academy of Sciences* **21**, 554–559.
- [43] Kendall, M.G. (1963). *Rank Correlation Methods*, 3rd Edition, Griffin Publishing, London.
- [44] Kendall, P.C., Marss-Garcia, A., Nath, S.R. & Sheldrick, R.C. (1999). Normative comparisons for the evaluation of clinical significance, *Journal of Consulting and Clinical Psychology* **67**, 285–299.
- [45] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole Publishing, Monterey.
- [46] Kirk, R.E. (1996). Practical significance: a concept whose time has come, *Educational and Psychological Measurement* **56**, 746–759.
- [47] Kirk, R.E. (2001). Promoting good statistical practices: some suggestions, *Educational and Psychological Measurement* **61**, 213–218.
- [48] Kirk, R.E. (2003). The importance of effect magnitude, in *Handbook of Research Methods in Experimental Psychology*, S.F. Davis, ed., Blackwell Science, Oxford, pp. 83–105.
- [49] Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: use in meta-analysis, *Journal of Educational Statistics* **8**, 93–101.
- [50] Lax, D.A. (1985). Robust estimators of scale: finite sample performance in long-tailed symmetric distributions, *Journal of the American Statistical Association* **80**, 736–741.
- [51] Lehmann, E.L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two, *Journal of the American Statistical Association* **88**, 1242–1248.
- [52] Levin, J.R. (1967). Misinterpreting the significance of “explained variation,” *American Psychologist* **22**, 675–676.
- [53] Lord, F.M. (1950). *Efficiency of Prediction When a Regression Equation From One Sample is Used in a New Sample*, Research Bulletin 50–110, Educational Testing Service, Princeton.
- [54] McGraw, K.O. & Wong, S.P. (1992). A common language effect size statistic, *Psychological Bulletin* **111**, 361–365.
- [55] Meehl, P.E. (1967). Theory testing in psychology and physics: a methodological paradox, *Philosophy of Science* **34**, 103–115.
- [56] O’Grady, K.E. (1982). Measures of explained variation: cautions and limitations, *Psychological Bulletin* **92**, 766–777.
- [57] Olejnik, S. & Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations, *Contemporary Educational Psychology* **25**, 241–286.
- [58] Olejnik, S. & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for common research designs, *Psychological Methods* **8**, 434–447.
- [59] Preece, P.F.W. (1983). A measure of experimental effect size based on success rates, *Educational and Psychological Measurement* **43**, 763–766.
- [60] Prentice, D.A. & Miller, D.T. (1992). When small effects are impressive, *Psychological Bulletin* **112**, 160–164.
- [61] Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect, *Journal of Educational Psychology* **74**, 166–169.

- [62] Rosenthal, R. & Rubin, D.B. (1989). Effect size estimation for one-sample multiple-choice-type data: design, analysis, and meta-analysis, *Psychological Bulletin* **106**, 332–337.
- [63] Rosenthal, R. & Rubin, D.B. (1994). The counternull value of an effect size: a new statistic, *Psychological Science* **5**, 329–334.
- [64] Rosenthal, R. & Rubin, D.B. (2003). $r_{\text{equivalent}}$: a simple effect size indicator, *Psychological Methods* **8**, 492–496.
- [65] Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science, *American Psychologist* **44**, 1276–1284.
- [66] Rosnow, R.L., Rosenthal, R. & Rubin, D.B. (2000). Contrasts and correlations in effect-size estimation, *Psychological Science* **11**, 446–453.
- [67] Rossi, J.S. (1997). A case study in the failure of psychology as cumulative science: the spontaneous recovery of verbal learning, in *What if There Were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum, Hillsdale, pp. 175–197.
- [68] Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test, *Psychological Bulletin* **57**, 416–428.
- [69] Sánchez-Meca, J., Marín-Martínez, F. & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis, *Psychological Methods* **8**, 448–467.
- [70] Särndal, C.E. (1974). A comparative study of association measures, *Psychometrika* **39**, 165–187.
- [71] Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for the training of researchers, *Psychological Methods* **1**, 115–129.
- [72] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* **105**, 309–316.
- [73] Shaver, J.P. (1993). What statistical significance testing is, and what it is not, *Journal of Experimental Education* **61**, 293–316.
- [74] Spencer, B. (1995). Correlations, sample size, and practical significance: a comparison of selected psychological and medical investigations, *Journal of Psychology* **129**, 469–475.
- [75] Tang, P.C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use, *Statistics Research Memorandum* **2**, 126–149.
- [76] Tatsuoaka, M.M. (1973). *An Examination of the Statistical Properties of a Multivariate Measure of Strength of Association*, Contract No. OEG-5-72-0027, U.S. Office of Education, Bureau of Research, Urbana-Champaign.
- [77] Thompson, B. (1998). In praise of brilliance: where that praise really belongs, *American Psychologist* **53**, 799–800.
- [78] Thompson, B. (2002). “Statistical,” “practical,” and “clinical”: how many kinds of significance do counselors need to consider? *Journal of Counseling and Development* **80**, 64–71.
- [79] Tukey, J.W. (1991). The philosophy of multiple comparisons, *Statistical Science* **6**, 100–116.
- [80] Wherry, R.J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation, *Annals of Mathematical Statistics* **2**, 440–451.
- [81] Wickens, C.D. (1998). Commonsense statistics, *Ergonomics in Design* **6**(4), 18–22.
- [82] Wilcox, R.R. (1996). *Statistics for the Social Sciences*, Academic Press, San Diego.
- [83] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego.
- [84] Wilcox, R.R. & Muska, J. (1999). Measuring effect size: a non-parametric analogue of $\hat{\omega}^2$, *British Journal of Mathematical and Statistical Psychology* **52**, 93–110.

ROGER E. KIRK

Eigenvalue/Eigenvector

IAN JOLLIFFE

Volume 2, pp. 542–543

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Eigenvalue/Eigenvector

The terms eigenvalue and eigenvector are frequently encountered in **multivariate analysis** in particular in computer software that implements multivariate methods (see **Software for Statistical Analyses**). The reason for the widespread appearance of the terms is that many multivariate techniques reduce to finding the maximum or minimum of some quantity, and optimization of the quantity is achieved by solving what is known as an *eigenproblem*. Any $(p \times p)$ matrix has associated with it a set of p eigenvalues (not necessarily all distinct), which are scalars, and associated with each eigenvalue is its eigenvector, a vector of length p . Note that there are a number of alternative terminologies, including *latent roots/latent vectors*, *characteristic roots/vectors* and *proper values/vectors*. We discuss the general mathematical form of eigenproblems later, but start by explaining eigenvectors and eigenvalues in perhaps their most common statistical context, **principal component analysis**.

In principal component analysis, suppose we have n measurements on a vector $\mathbf{x}$ of p random variables. If p is large, it may be possible to derive a smaller number, q , of linear combinations of the variables in $\mathbf{x}$ that retain most of the variation in the full data set. Principal component analysis formalizes this by finding linear combinations, $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \dots, \mathbf{a}_q'\mathbf{x}$, called *principal components*, that successively have maximum variance for the data, subject to being uncorrelated with previous $\mathbf{a}_k'\mathbf{x}$ s. Solving this maximization problem, we find that the vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_q$ are the eigenvectors, corresponding to the q largest eigenvalues, of the **covariance matrix** $\mathbf{S}$, of the data. Thus, in these circumstances, eigenvectors provide vectors of coefficients, weights or loadings that define the principal components in terms of the p original

variables. The eigenvalues also have a straightforward interpretation in principal component analysis, namely, they are the variances of their corresponding components.

Eigenvalues and eigenvectors also appear in other multivariate techniques. For example, in **discriminant analysis** those linear functions of $\mathbf{x}$ are found that best separate a number of groups, in the sense of maximizing the ratio of ‘between group’ to ‘within group’ variability for the linear functions. The coefficients of the chosen linear functions are derived as eigenvectors of $\mathbf{B}\mathbf{W}^{-1}$, where $\mathbf{B}$, $\mathbf{W}$ are matrices of between- and within-group variability for the p variables. A second example (there are many others) is **canonical correlation analysis**, in which linear functions of two sets of variables $\mathbf{x}_1, \mathbf{x}_2$ are found that have maximum correlation. The coefficients in these linear functions are once again derived as eigenvectors of a product of matrices. In this case, the elements of the matrices are variances and covariances for $\mathbf{x}_1$ and $\mathbf{x}_2$. The corresponding eigenvalues give squared correlations between the pairs of linear functions.

Finally, we turn to the mathematical definition of eigenvalues and eigenvectors. Consider a $(p \times p)$ matrix $\mathbf{S}$, which in the context of the multivariate techniques noted above often (though not always) consists of variances, covariances, or correlations. A vector $\mathbf{a}$ is an eigenvector of $\mathbf{S}$, and l is the corresponding eigenvalue, if $\mathbf{S}\mathbf{a} = l\mathbf{a}$. Geometrically, this means that $\mathbf{a}$ is a direction that is unchanged by the linear operator $\mathbf{S}$. If $\mathbf{S}$ is symmetric, the eigenvalues are real, but otherwise they may be complex. The sum of the eigenvalues equals the trace of the matrix $\mathbf{S}$ (the sum of its diagonal elements) and the product of the eigenvalues is the determinant of the matrix.

IAN JOLLIFFE

Empirical Quantile–Quantile Plots

SANDY LOVIE

Volume 2, pp. 543–545

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Empirical Quantile–Quantile Plots

An empirical quantile–quantile (EQQ) plot provides a graphical comparison between measures of *location* (means and medians, for instance) and *spread* (standard deviations, variances, ranges, etc.) for two ordered sets of observations, hence the name of the plot, where **quantiles** are ordered values and *empirical* refers to the source of the data. What one has with an EQQ plot, therefore, is the graphical equivalent of significance tests of differences in location and spread. The display itself is a form of added value **scatterplot**, in which the x and y axes represent the ordered values of the two sets of data.

Interpreting the resulting graph is easiest if the axes are identical, since an essential part of the plot is a 45-degree *comparison line* running from the bottom left-hand corner (the origin) to the upper right-hand corner of the display. Decisions about the data are made with respect to this comparison line; for instance, are the data parallel to it, or coincident with it, or is the bulk of the data above or below it? Indeed, so much information can be extracted from an EQQ plot that it is helpful to provide a summary table of data/comparison line outcomes and their statistical interpretation (see Table 1).

Three example EQQ plots are shown below. The data are taken from Minitab's *Fish* and *Crowd*

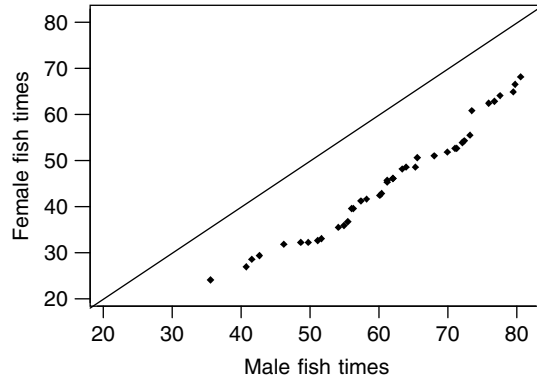


Figure 1 EQQ plot of average nest-guarding times for male and female fish

datasets (the latter omitting one problematic pair of observations).

In Figure 1, the location measure (for example, the average) for the male fish guarding time is higher than the equivalent measure for the female fish, since all the data lie *below* the 45% comparison line. However, the spreads seem the same in each sample as the data are clearly parallel with the comparison line.

In Figure 2, the data are almost exactly coincident with the 45% comparison line, thus showing graphically that there are no differences in either location or spread of the estimates of the crowdedness of the room by male and female students.

Table 1 Interpreting an EQQ plot

Relationship of data to comparison line	Interpretation with respect to location and spread
Bulk of data lies ABOVE	Locations different; location of y variable higher than x variable
Bulk of data lies BELOW	Locations different; location of x variable higher than y variable
Data PARALLEL	No difference in spreads
Data NOT PARALLEL	Difference in spreads; if the data starts ABOVE the comparison line and crosses it, then y variable spread smaller than x 's; <i>vice versa</i> for data starting BELOW the line
Data BOTH not parallel AND lies above or below	Differences in BOTH locations AND spreads
Data COINCIDENT	No differences in locations or spreads; samples are basically identical

2 Empirical Quantile–Quantile Plots

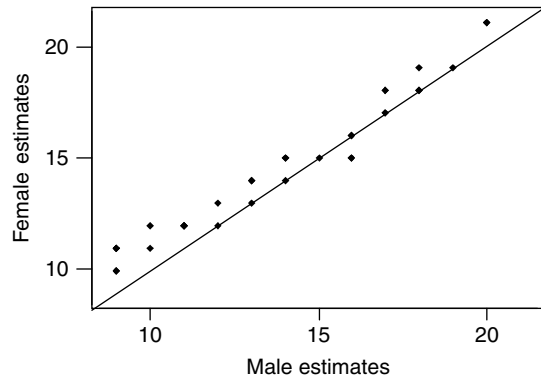


Figure 2 EQQ plot of estimates by male and female students of the crowdedness of a room

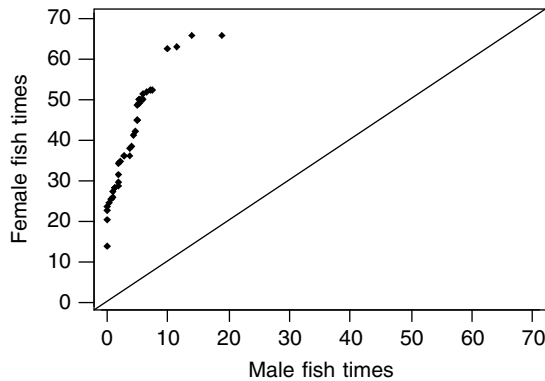


Figure 3 EQQ plot of nest-fanning times by male and female fish

In Figure 3 we have the time taken by both male and female fishes fanning their nests in order to

increase the flow of oxygenated water over the eggs. Clearly, the average female-fanning times are longer than those of the males since all the data lie above the comparison line, but it is also the case that the plotted data are not parallel to the comparison line, thereby strongly suggesting a difference in spreads as well. Here the female times seem to show a considerably larger spread than those of the male fish. One way of seeing this is to mentally project the largest and smallest values from the plot onto the x - and y -axes, and note which cuts off the larger or smaller length.

Finally, EQQ plots are most useful when the samples are independent because the necessity to order the data discards any correlational information in dependent samples. It is also useful to have equal numbers in the samples, although an interpolation routine is offered by Chambers et al. to substitute for any missing values ([1], pp. 54–55; see also pp. 48–57 for more on EQQ plots).

Reference

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.

(See also **Probability Plots**)

SANDY LOVIE

Epistasis

DAVID M. EVANS

Volume 2, pp. 545–546

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Epistasis

Several biologically motivated definitions of epistasis exist [1, 2], however, most modern textbooks of genetics define epistasis to include any interaction between genes at different loci to produce effects which are different from that expected from the individual loci themselves. For example, seed capsules produced by the shepard's purse plant (*Bursa bursa-pastoris*) are normally triangular as a result of two separate dominant pathways. It is only when both pathways are blocked through the action of a double recessive that oval capsules are produced. Thus, crosses from plants that are doubly heterozygous produce triangular and oval shaped purses in the ratio 15:1 rather than the usual Mendelian ratio of 3:1 [10, 13].

Epistasis may also be defined in quantitative genetic terms [5]. In this case, epistasis refers to a deviation from additivity in the effect of alleles and genotypes at different loci with respect to their contribution to a quantitative phenotype [2, 5]. Although the biological and quantitative genetic definitions of epistasis are related, it is important to realize that they are not formally equivalent [2]. In terms of the standard biometrical model, epistasis may involve interactions between the additive and/or dominance effects at two or more loci. For example, in the case of two loci, there may be additive $\times$ additive interactions, additive $\times$ dominance interactions, and dominance $\times$ dominance interactions. As the number of loci contributing to the trait increases, the number of possible epistatic interactions increases rapidly also (i.e., three-way interactions, four-way interactions etc.). It is important to realize that choice of scale is important, since a trait that demonstrates epistasis on one scale may not show evidence of epistasis on another transformed scale and vice versa.

Whilst in the case of a single locus, the total genetic variance of a quantitative trait is simply the sum of the additive and dominance components of variance, when multiple loci are considered, the genetic variance may also contain additional components of variation due to epistasis. The proportion of genetic variance due to epistatic interactions is termed the *epistatic variance*. It is the residual genetic variance, which is unexplained by the additive and dominance components. Similar to other **variance components**, the epistatic variance is a property of

the particular population of interest being dependent on population parameters such as allele and multi-locus genotype frequencies. The interested reader is referred to any of the classic texts in quantitative genetics for a formal mathematical derivation of epistatic variance [3, 6, 7, 9].

It is impossible to estimate epistatic variance components using the classical twin study (see **Twin Designs**). Several large twin studies have yielded low correlations between dizygotic twins, which cannot be explained through the effect of genetic dominance alone [4]. While it is possible to resolve higher order epistatic effects through careful breeding studies in experimental organisms (e.g., *Nicotiana Rustica*, *Drosophila*), this is unrealistic in human populations for a number of reasons. First, the analysis of higher order epistatic interactions would yield more parameters than could ever be estimated from any set of relatives [4]. Second, the correlation between the different components would be so high as to make it impossible to resolve them reliably [4, 8, 14]. Finally, when gene frequencies are unequal, multiple sources of gene action contribute to the variance components making interpretation of these components in terms of the underlying gene action problematic [4, 9].

Finally, it is possible to fit a variety of linkage and association models which incorporate epistatic effects at measured loci [2]. For example, it is possible to fit a two-locus variance components linkage model which includes a component of variance due to additive $\times$ additive epistatic interactions (for an example of how to do this see [11]). Although the power to detect epistasis will in general be low, Purcell and Sham [12] make the valuable point that it is still possible to detect a quantitative trait locus (QTL), which has no main effect but interacts epistatically with another unmeasured locus using a single locus model. This is because allele-sharing variables, which index epistatic and non-epistatic effects, are correlated. In other words, a single locus model will soak up most of the variance due to epistatic effects even though the power to detect epistasis formally through a multi-locus model is low [12].

References

- [1] Bateson, W. (1909). *Mendel's Principles of Heredity*, Cambridge University Press, Cambridge.

2 Epistasis

- [2] Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans, *Human Molecular Genetics* **11**, 2463–2468.
- [3] Crow, J.F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*, Harper & Row, New York.
- [4] Eaves, L.J. (1988). Dominance alone is not enough, *Behavior Genetics* **18**, 27–33.
- [5] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transaction of the Royal Society. Edinburgh* **52**, 399–433.
- [6] Kempthorne, O. (1957). *An Introduction to Genetic Statistics*, John Wiley & Sons, New York.
- [7] Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*, Sinauer Associates, Sunderland.
- [8] Martin, N.G., Eaves, L.J., Kearsley, M.J. & Davies, P. (1978). The power of the classical twin study, *Heredity* **40**, 97–116.
- [9] Mather, K. & Jinks, J.L. (1982). *Biometrical Genetics*, Chapman & Hall, New York.
- [10] Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases, *Human Heredity* **56**, 73–82.
- [11] Neale, M.C. (2002). QTL mapping in sib-pairs: the flexibility of Mx, in *Advances in Twin and Sib-Pair Analysis*, T.D. Spector, H. Sneider & A.J. MacGregor, eds, Greenwich Medical Media, London, pp. 219–244.
- [12] Purcell, S. & Sham, P.C. (2004). Epistasis in quantitative trait locus linkage analysis: interaction or main effect? *Behavior Genetics* **34**, 143–152.
- [13] Shull, G.H. (1914). Duplicate genes for capsule form in BURSA bursa Bastoris, *Zeitschrift für Induktive Abstammungs-und Vererbungslehre* **12**, 97–149.
- [14] Williams, C.J. (1993). On the covariance between parameter estimates in models of twin data, *Biometrics* **49**, 557–568.

DAVID M. EVANS

Equivalence Trials

JOHN P. HATCH

Volume 2, pp. 546–547

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Equivalence Trials

Methods for statistical equivalence testing grew out of the need by the pharmaceutical industry to demonstrate the bioequivalence of drugs [1–3]. When the patent expires on a brand-name drug, other companies may file abbreviated applications with the Food and Drug Administration (FDA) for approval of a generic equivalent without full efficacy and safety testing. What is needed is evidence that the generic and brand-name products differ only to a clinically unimportant degree. This is known as *equivalence testing*. Furthermore, as the number of drugs proven to be effective increases, it becomes increasingly questionable whether new drugs can ethically be compared to a placebo (inactive control). In such cases, the researcher may wish to compare the new drug against one already proven to be safe and effective (positive control). The aim of such a study may not be to demonstrate the superiority of the test drug but rather simply to demonstrate that it is not clinically inferior to the proven one. This is known as *noninferiority testing*.

Classical approaches to hypothesis testing (*see Classical Statistical Inference: Practice versus Presentation*), which test the null hypothesis of exact equality, are inappropriate for the equivalence problem. A conclusion of no real difference based upon the lack of a statistically significant difference is based on insufficient evidence. The lack of statistical significance might merely be due to insufficient statistical **power** or to excessive measurement error. What is needed is a method that permits us to decide whether the difference between treatments is small enough to be safely ignored in practice.

Equivalence testing begins with the *a priori* statement of a definition of equivalence. This should correspond to the largest difference that can be considered unimportant for the substantive problem at hand. Equivalence margins then are defined, bounded by lower and upper end points Δ_1 and Δ_2 , respectively. These margins define the range of mean differences ($M_1 - M_2$) that will be considered equivalent. It is not necessary that $\Delta_2 = -\Delta_1$; a larger difference in one direction than the other is allowed. For equivalence testing, both lower and upper margins are needed, but for noninferiority testing only the lower margin is needed. Next, data are collected and

the $(1 - 2\alpha)$ **confidence interval** for the difference between treatment means is calculated. Equivalence is inferred if this confidence interval falls entirely within the equivalence margins.

Equivalence testing also may be accomplished by simultaneously performing two one-sided hypothesis tests (*see Directed Alternatives in Testing*). One test seeks to reject the null hypothesis that the mean difference is less than or equal to Δ_1 :

$$\begin{aligned} H_0: M_1 - M_2 &\leq \Delta_1 \\ H_A: M_1 - M_2 &> \Delta_1 \end{aligned} \quad (1)$$

The second test seeks to reject the null hypothesis that the mean difference is greater than or equal to Δ_2 :

$$\begin{aligned} H_0: M_1 - M_2 &\geq \Delta_2 \\ H_A: M_1 - M_2 &< \Delta_2 \end{aligned} \quad (2)$$

In other words, we test the composite null hypothesis that the mean difference falls outside the equivalence margins versus the composite alternative hypothesis that the mean difference falls within the margins. The type I error rate is not affected by performing two tests because they are disjoint. The method just described is referred to as *average bioequivalence testing* because only population means are considered. Other methods are available for population bioequivalence testing (comparability of available drugs that a physician could prescribe for an individual new patient) and individual bioequivalence testing (switchability of available drugs within an individual patient) [1].

References

- [1] Chow, S.-C. & Shao, J. (2002). *Statistics in Drug Research: Methodologies and Recent Developments*, Marcel Dekker, New York, pp. 107–146, Chapter 5.
- [2] FDA. (2001). *Guidance for Industry on Statistical Approaches to Establishing Bioequivalence*, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville.
- [3] Hauck, W.W. & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials, *Journal of Pharmacokinetics and Biopharmaceutics* **12**, 83–91, 657.

JOHN P. HATCH

Error Rates

BRUCE THOMPSON

Volume 2, pp. 547–549

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Error Rates

As some recent histories [1, 2] of null hypothesis statistical significance testing (NHSST) confirm, contemporary NHSST practices are an amalgam of the contradictory philosophies espoused by **Sir Ronald Fisher** as against **Jerzy Neyman** and **Egon Pearson**. In the appendix to their chapter, Mulaik, Raju, and Harshman [4] provide a thoughtful summary of those arguments.

Within the contemporary amalgam of NHSST practices, today, most researchers acknowledge the possibility with sample data of rejecting a null hypothesis that in reality is true in the population. This mistake has come to be called a *Type I error*. Conversely, the failure to reject a null hypothesis when the null hypothesis is untrue in the population is called a *Type II error*. Of course, a given decision regarding a single null hypothesis cannot be both a Type I and a Type II error, and we can possibly make a Type I error only if we reject the null hypothesis (i.e., the result is ‘statistically significant’).

Unless we later collect data from the full population, we will never know for sure whether the decisions we take with sample data are correct, or the decisions are errors. However, we can mathematically estimate the probability of errors, and these can range between 0.0 and 1.0. Unless we are perverse scientists with an evil fascination with untruth, we prefer the probabilities of decision errors to be small (e.g., 0.01, 0.05, 0.10).

The ceiling we select as that maximum probability of a Type I error, called α or α_{TW} (e.g., $\alpha = 0.05$), on a given, single hypothesis test is called the *testwise error rate*. In other words, ‘error rate’ always refers only to Type I errors. (When researchers say only α , they are implicitly always referring to α_{TW} .)

However, when we test multiple hypotheses (see **Multiple Comparison Procedures**), the probability of making one or more Type I errors in the set of decisions, called the *experimentwise error rate* (α_{EW}), is not necessarily equal to the probability we select as the ceiling for testwise error rate (α_{TW}). The α_{EW} is always equal to or greater than the α_{TW} .

There are two situations in which in a given study $\alpha_{EW} = \alpha_{TW}$. First, if we conduct a study in which we test only a single hypothesis, then α_{EW} must equal α_{TW} . Second, if we test multiple hypotheses in which the outcome variables or the hypotheses are

perfectly correlated with each other, then α_{EW} still must equal α_{TW} .

For example, if we are investigating the effects of an intervention versus a control condition on self-concept, we might employ two different self-concept tests, because we are not totally confident that either test is perfect at measuring the construct. If it turned out that the two outcome variables were perfectly correlated, even if we performed two *t* Tests or two ANOVAs (analyses of variance) to analyze the data, the α_{EW} would still equal α_{TW} .

The previous example (correctly) suggests that the correlations of the outcome variables or the hypotheses impact the inflation of the testwise error rate (i.e., the experimentwise error rate). Indeed, α_{EW} is most inflated when the outcome variables or the hypotheses are perfectly uncorrelated. At this extreme, the formula due to Bonferroni can be used to compute the experimentwise error rate:

$$\alpha_{EW} = 1 - (1 - \alpha_{TW})^k, \quad (1)$$

where k equals the number of outcome variables or hypotheses tested.

Love [3] provides a mathematical proof that the Bonferroni formula is correct. If the outcome variables or hypotheses are neither perfectly correlated nor perfectly uncorrelated, α_{EW} is computationally harder to determine, but would fall within the range of α_{TW} to $[1 - (1 - \alpha_{TW})^k]$.

For example, at one extreme if we tested five perfectly correlated hypotheses, each at $\alpha_{TW} = 0.05$, then $\alpha_{EW} = \alpha_{TW} = 0.05$. At the other extreme, if the hypotheses or outcomes were perfectly uncorrelated, the Bonferroni formula applies, and α_{EW} would be

$$\begin{aligned} &1 - (1 - 0.05)^5 \\ &1 - (0.95)^5 \\ &1 - 0.774 = 0.226. \end{aligned} \quad (2)$$

In other words, if five uncorrelated hypotheses are tested, each at $\alpha_{TW} = 0.05$, then the probability of one or more Type I errors being made is 22.6%. Two very big problems with this disturbing result are that the probability does not tell us (a) exactly how many Type I errors (e.g., 1, 2, 3...) are being made or (b) where these errors are.

This is one reason why multivariate statistics are often necessary. If we test one multivariate

2 Error Rates

hypothesis, rather than conducting separate tests of the five outcome variables, $\alpha_{EW} = \alpha_{TW}$.

Years ago, researchers noticed that α_{EW} approximately equals $k(\alpha_{TW})$ (e.g., 22.6% approximately equals $5(0.05) = 25.0\%$). Thus was born the ‘Bonferroni correction’, which adjusts the original α_{TW} downward, so that given the new testwise alpha level (α_{TW}^*), the α_{EW} would be roughly equal to α_{TW} . With the present example, α_{TW}^* would be set equal to 0.01, because $\alpha_{TW} = 0.05/5$ is 0.01. However, one problem with using the Bonferroni correction in this manner is that although the procedure controls the experiment-wise Type I error rate, the probability of making Type II error gets correspondingly larger with this method.

One common application of the Bonferroni correction that is more appropriate involves *post hoc* tests in ANOVA. When we test whether the means of more than two groups are equal, and determine that some differences exist, the question arises as to exactly which groups differ. We address this question by invoking one of the myriad *post hoc* test procedures (e.g., Tukey, Scheffè, Duncan).

Post hoc tests always compare one mean versus a second mean. Because the differences in two means are being tested, *post hoc* tests invoke a variation on

the *t* Test invented by a Guinness brewery worker roughly a century ago. Conceptually, ANOVA *post hoc* tests (e.g., Tukey, Scheffè, Duncan) are *t* Tests with build in variations on the Bonferroni correction being invoked so as to keep α_{EW} from becoming too inflated.

References

- [1] Hubbard, R. & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects, *Educational and Psychological Measurement* **60**, 661–681.
- [2] Huberty, C.J. (1999). On some history regarding statistical testing, in *Advances in Social Science Methodology*, Vol. 5, B. Thompson, ed., JAI Press, Stamford, pp. 1–23.
- [3] Love, G. (November 1988). Understanding experiment-wise error probability, in *Paper Presented at the Annual Meeting of the Mid-South Educational Research Association*, Louisville, (ERIC Document Reproduction Service No. ED 304 451).
- [4] Mulaik, S.A., Raju, N.S. & Harshman, R.A. (1997). There is a time and place for significance testing, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Erlbaum, Mahwah, pp. 65–115.

BRUCE THOMPSON

Estimation

SARA A. VAN DE GEER

Volume 2, pp. 549–553

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Estimation

In the simplest case, a data set consists of observations on a single variable, say real-valued observations. Suppose there are n such observations, denoted by $X_1, \dots, X_n$. For example, X_i could be the reaction time of individual i to a given stimulus, or the number of car accidents on day i , and so on. Suppose now that each observation follows the same probability law P . This means that the observations are relevant if one wants to predict the value of a new observation X (say, the reaction time of a hypothetical new subject, or the number of car accidents on a future day, etc.). Thus, a common underlying distribution P allows one to generalize the outcomes.

The emphasis in this paper is on the data and estimators derived from the data, and less on the estimation of population parameters describing a model for P . This is because the data exist, whereas population parameters are a theoretical construct (see **Model Evaluation**). An estimator is any given function $T_n(X_1, \dots, X_n)$ of the data. Let us start with reviewing some common estimators.

The Empirical Distribution. The unknown P can be estimated from the data in the following way. Suppose first that we are interested in the probability that an observation falls in A , where A is a certain set chosen by the researcher. We denote this probability by $P(A)$. Now, from the frequentist point of view, a probability of an event is nothing else than the limit of relative frequencies of occurrences of that event as the number of occasions of possible occurrences n grows without limit. So, it is natural to estimate $P(A)$ with the frequency of A , that is, with

$$\begin{aligned} \hat{P}_n(A) &= \frac{\text{number of times an observation } X_i \text{ falls in } A}{\text{total number of observations}} \\ &= \frac{\text{number of } X_i \in A}{n}. \end{aligned} \quad (1)$$

We now define the empirical distribution $\hat{P}_n$ as the probability law that assigns to a set A the probability $\hat{P}_n(A)$. We regard $\hat{P}_n$ as an estimator of the unknown P .

The Empirical Distribution Function. The distribution function of X is defined as

$$F(x) = P(X \leq x), \quad (2)$$

and the empirical distribution function is

$$\hat{F}_n(x) = \frac{\text{number of } X_i \leq x}{n}. \quad (3)$$

Figure 1 plots the distribution function $F(x) = 1 - 1/x^2$, $x \geq 1$ (smooth curve) and the empirical distribution function $\hat{F}_n$ (stair function) of a sample from F with sample size $n = 200$.

Sample Moments and Sample Variance. The theoretical mean

$$\mu = E(X), \quad (4)$$

(E stands for *Expectation*), can be estimated by the sample average

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}. \quad (5)$$

More generally, for $j = 1, 2, \dots$ the j th sample moment

$$\hat{\mu}_{j,n} = \frac{X_1^j + \dots + X_n^j}{n}, \quad (6)$$

is an estimator of the j th moment $E(X^j)$ of P (see **Moments**).

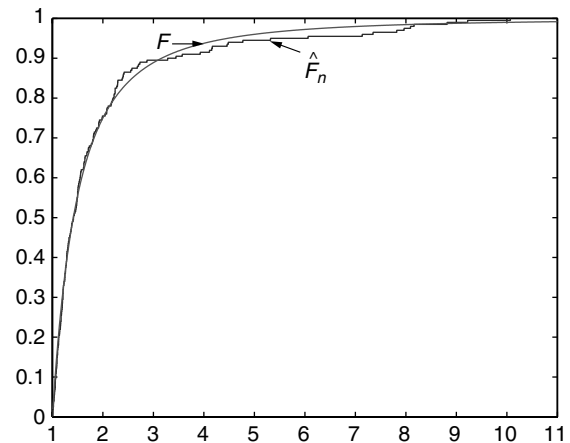


Figure 1 The empirical and theoretical distribution function

2 Estimation

The sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (7)$$

is an estimator of the variance $\sigma^2 = E(X - \mu)^2$.

Sample Median. The median of X is the value m that satisfies $F(m) = 1/2$ (assuming there is a unique solution). Its empirical version is any value $\hat{m}_n$ such that $\hat{F}_n(\hat{m}_n)$ is equal or as close as possible to $1/2$. In the above example $F(x) = 1 - 1/x^2$, so that the theoretical median is $m = \sqrt{2} = 1.4142$. In the ordered sample, the 100th observation is equal to 1.4166, and the 101th observation is equal to 1.4191. A common choice for the sample median is taking the average of these two values. This gives $\hat{m}_n = 1.4179$.

Parametric Models. The distribution P may be partly known beforehand. The unknown parts of P are called *parameters* of the model. For example, if the X_i are yes/no answers to a certain question (the binary case), we know that P allows only two possibilities, say 1 and 0 (yes = 1, no = 0). There is only one parameter, say the probability of a yes answer $\theta = P(X = 1)$. More generally, in a parametric model, it is assumed that P is known up to a finite number of parameters $\theta = (\theta_1, \dots, \theta_d)$. We then often write $P = P_\theta$. When there are infinitely many parameters (which is, for example, the case when P is completely unknown), the model is called nonparametric.

If $P = P_\theta$ is a parametric model, one can often apply the maximum likelihood procedure to estimate θ (see **Maximum Likelihood Estimation**).

Example 1 The time one stands in line for a certain service is often modeled as exponentially distributed. The random variable X representing the waiting time then has a density of the form

$$f_\theta(x) = \theta e^{-\theta x}, \quad x > 0, \quad (8)$$

where the parameter θ is the so-called *intensity* (a large value of θ means that - on average - the waiting time is short), and the maximum likelihood estimator of θ is

$$\hat{\theta}_n = \frac{1}{\bar{X}_n}. \quad (9)$$

Example 2 In many cases, one assumes that X is normally distributed. In that case there are two parameters: the mean $\theta_1 = \mu$ and the variance $\theta_2 = \sigma^2$. The maximum likelihood estimators of (μ, σ^2) are $(\hat{\mu}_n, \hat{\sigma}_n^2)$, where $\hat{\mu}_n = \bar{X}_n$ is the sample mean and $\hat{\sigma}_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$.

The Method of Moments. Suppose that the parameter θ can be written as a given function of the moments $\mu_1, \mu_2, \dots$. The *methods of moments estimator* replaces these moments by their empirical counterparts $\hat{\mu}_{n,1}, \hat{\mu}_{n,2}, \dots$

Example 3 Vilfredo Pareto [2] noticed that the number of people whose income exceeds level x is often approximately proportional to $x^{-\theta}$, where θ is a parameter that differs from country to country. Therefore, as a model for the distribution of incomes, one may propose the Pareto density

$$f_\theta(x) = \frac{\theta}{x^{\theta+1}}, \quad x > 1. \quad (10)$$

When $\theta > 1$, one has $\theta = \mu/(\mu - 1)$. Hence, the method of moments estimator of θ is in this case $t_1(\hat{P}_n) = \bar{X}_n/(\bar{X}_n - 1)$. After some calculations, one finds that the maximum likelihood estimator of θ is $t_2(\hat{P}_n) = (n/\sum_{i=1}^n \log X_i)$. Let us compare these on the simulated data in Figure 1. We generated in this simulation a sample from the Pareto distribution with $\theta = 2$. The sample average turns out to be $\bar{X}_n = 1.9669$, so that the methods of moments estimate is 2.0342. The maximum likelihood estimate is 1.9790. Thus, the maximum likelihood estimate is a little closer to the true θ than the methods of moments estimate.

Properties of Estimators. Let $T_n = T_n(X_1, \dots, X_n)$ be an estimator of the real-valued parameter θ . Then it is desirable that T_n is in some sense close to θ . A minimum requirement is that the estimator approaches θ as the sample size increases. This is called *consistency*. To be more precise, suppose the sample $X_1, \dots, X_n$ are the first n of an infinite sequence $X_1, X_2, \dots$ of independent copies of X . Then T_n is called consistent if (with probability one)

$$T_n \rightarrow \theta \text{ as } n \rightarrow \infty. \quad (11)$$

Note that consistency of frequencies as estimators of probabilities, or means as estimators of expectations, follows from the (strong) **law of large numbers**.

The *bias* of an estimator T_n of θ is defined as its mean deviation from θ :

$$\text{bias}(T_n) = E(T_n) - \theta. \quad (12)$$

We remark here that the distribution of $T_n = T_n(X_1, \dots, X_n)$ depends on P , and, hence, on θ . Therefore, the expectation $E(T_n)$ depends on θ as well. We indicate this by writing $E(T_n) = E_\theta(T_n)$. The estimator T_n is called *unbiased* if

$$E_\theta(T_n) = \theta \text{ for all possible values of } \theta. \quad (13)$$

Example 4 Consider the estimators $S_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ and $\hat{\sigma}_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n$. Note that S_n^2 is larger than $\hat{\sigma}_n^2$, but that the difference is small when n is large. It can be shown that S_n^2 is an unbiased estimator of the variance $\sigma^2 = \text{var}(X)$. The estimator $\hat{\sigma}_n^2$ is biased: it underestimates the variance.

In many models, unbiased estimators do not exist. Moreover, it often heavily depends on the model under consideration, whether or not an estimator is unbiased. A weaker concept is *asymptotic unbiasedness* (see [1]).

The mean square error of T_n as estimator of θ is

$$\text{MSE}(T_n) = E(T_n - \theta)^2. \quad (14)$$

One may decompose the MSE as

$$\text{MSE}(T_n) = \text{bias}^2(T_n) + \text{var}(T_n), \quad (15)$$

where $\text{var}(T_n)$ is the variance of T_n .

Bias, variance, and mean square error are often quite hard to compute, because they depend on the distribution of all n observations $X_1, \dots, X_n$. However, one may use certain approximations for large sample sizes n . Under regularity conditions, the maximum likelihood estimator $\hat{\theta}_n$ of θ is asymptotically unbiased, with asymptotic variance $1/(nI(\theta))$, where $I(\theta)$ is the Fisher information in a single observation (see **Information Matrix**). Thus, maximum likelihood estimators reach the minimum variance bound asymptotically.

Histograms. Our next aim is estimating the density $f(x)$ at a given point x . The density is defined as the derivative of the distribution function F at x :

$$f(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \frac{P(x, x+h]}{h}. \quad (16)$$

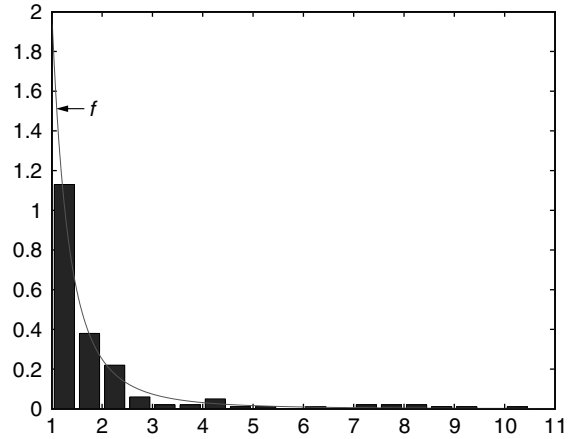


Figure 2 Histogram with bandwidth $h = 0.5$ and true density

Here, $(x, x+h]$ is the interval with left endpoint x (not included) and right endpoint $x+h$ (included). Unfortunately, replacing P by $\hat{P}_n$ here does not work, as for h small enough, $\hat{P}_n(x, x+h]$ will be equal to zero. Therefore, instead of taking the limit as $h \rightarrow 0$, we fix h at a (small) positive value, called the *bandwidth*. The estimator of $f(x)$, thus, becomes

$$\hat{f}_n(x) = \frac{\hat{P}_n(x, x+h]}{h} = \frac{\text{number of } X_i \in (x, x+h]}{nh}. \quad (17)$$

A plot of this estimator at points $x \in \{x_0, x_0 + h, x_0 + 2h, \dots\}$ is called a **histogram**.

Example 3 continued Figure 2 shows the histogram, with bandwidth $h = 0.5$, for the sample of size $n = 200$ from the Pareto distribution with parameter $\theta = 2$. The solid line is the density of this distribution.

Minimum Chi-square. Of course, for real (not simulated) data, the underlying distribution/density is not known. Let us explain in an example a procedure for checking whether certain model assumptions are reasonable. Suppose that one wants to test whether data come from the exponential distribution with parameter θ equal to 1. We draw a histogram of the sample (sample size $n = 200$), with bandwidth $h = 1$ and 10 cells (see Figure 3). The cell counts are (151, 28, 4, 6, 1, 1, 4, 3, 1, 1). Thus, for example, the number of observations that falls in

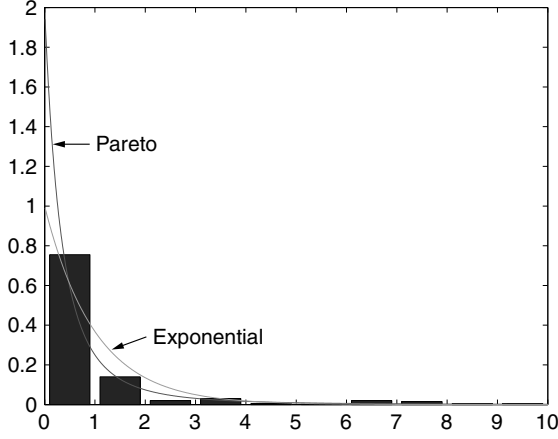


Figure 3 Histogram with bandwidth $h = 1$, exponential and Pareto density

the first cell, that is, that have values between 0 and 1, is equal to 151. The cell probabilities are, therefore, (0.755, 0.140, 0.020, 0.030, 0.005, 0.005, 0.020, 0.015, 0.005, 0.005). Now, according to the exponential distribution, the probability that an observation falls in cell k is equal to $e^{-(k-1)} - e^{-k}$, for $k = 1, \dots, 10$. These cell probabilities are (.6621, .2325, .0855, .0315, .0116, .0043, .0016, .0006, .0002, .0001). Because the probabilities of the last four cells are very small, we merge them together. This gives cell counts $(N_1, \dots, N_7) = (151, 28, 4, 6, 1, 1, 4, 9)$ and cell probabilities $(\pi_1, \dots, \pi_7) = (.6621, .2325, .0855, .0315, .0116, .0043, .0025)$. To check whether the cell frequencies differ significantly from the hypothesized cell probabilities, we calculate Pearson's χ^2 . It is defined as

$$\chi^2 = \frac{(N_1 - n\pi_1)^2}{n\pi_1} + \dots + \frac{(N_7 - n\pi_7)^2}{n\pi_7}. \quad (18)$$

We write this as $\chi^2 = \chi^2(\text{exponential})$ to stress that the cell probabilities were calculated assuming the exponential distribution. Now, if the data are exponentially distributed, the χ^2 statistic is generally not too large. But what is large? Consulting a table of Pearson's χ^2 at the 5% significance level gives the critical value $c = 12.59$. Here we use 6 degrees of freedom. This is because there are $m = 7$ cell probabilities, and there is the restriction $\pi_1 + \dots + \pi_m = 1$, so we estimated $m - 1 = 6$ parameters. After some calculations, one obtains $\chi^2(\text{exponential}) = 168.86$. This exceeds the critical

value c , that is, $\chi^2(\text{exponential})$ is too large to support the assumption of the exponential distribution. In fact, the data considered here are the simulated sample from the Pareto distribution with parameter $\theta = 2$. We shifted this sample one unit to the left. The value of χ^2 for this (shifted) Pareto distribution is

$$\chi^2(\text{Pareto}) = 10.81. \quad (19)$$

This is below the critical value c , so that the test, indeed, does not reject the Pareto distribution. However, this comparison is not completely fair, as our decision to merge the last four cells was based on the exponential distribution, which has much lighter tails than the Pareto distribution.

In Figure 3, the histogram is shown, together with the densities of the exponential and Pareto distribution. Indeed, the Pareto distribution fits the data better in the sense that it puts more mass at small values.

Continuing with the test for the exponential distribution, we note that, in many situations, the intensity θ is not required to be fixed beforehand. One may use an estimator for θ and proceed as before, calculating χ^2 with the estimated value for θ . However, the critical values of the test then become smaller. This is because, clearly, estimating parameters using the sample means that the hypothesized distribution is pulled towards the sample. Moreover, when using, for example, maximum likelihood estimators of the parameters, critical values will in fact be hard to compute. The minimum χ^2 estimator overcomes this problem. Let $\pi_k(\vartheta)$ denote the cell probabilities when the parameter value is ϑ , that is, in the exponential case $\pi_k(\vartheta) = e^{-\vartheta(k-1)} - e^{-\vartheta k}$, $k = 1, \dots, m - 1$, and $\pi_m(\vartheta) = 1 - \sum_{k=1}^{m-1} \pi_k(\vartheta)$. The minimum χ^2 estimator $\hat{\theta}_n$ is now the minimizer over ϑ of

$$\left\{ \frac{(N_1 - n\pi_1(\vartheta))^2}{n\pi_1(\vartheta)} + \dots + \frac{(N_m - n\pi_m(\vartheta))^2}{n\pi_m(\vartheta)} \right\}. \quad (20)$$

The χ^2 test with this estimator for θ now has $m - 2$ degrees of freedom. More generally, the number of degrees of freedom is $m - 1 - d$, where d is the number of estimated free parameters when calculating cell probabilities. The critical values of the test can be found in a χ^2 table.

Sufficiency. A goal of statistical analysis is generally to summarize the (large) data set into a small

number of characteristics. The sample mean and sample variance are such summarizing statistics, but so is, for example, the sample median, and so on. The question arises, to what extent one can summarize data without throwing away information. For example, suppose you are given the empirical distribution function $\hat{F}_n$, and you are asked to reconstruct the original data $X_1, \dots, X_n$. This is not possible since the ordering of the data is lost. However, the index i of X_i is just a label: it contains no information about the distribution P of X_i (assuming that each observation X_i comes from the same distribution, and the observations are independent). We say that the empirical distribution $\hat{F}_n$ is *sufficient*. More generally, a statistic $T_n = T_n(X_1, \dots, X_n)$ is called *sufficient* for P if the distribution of the data given the value of T_n does not depend on P . For example, it can be shown that when P is the exponential distribution with unknown intensity, then the sample mean is sufficient. When P is the normal distribution with unknown mean and variance, then the sample mean and sample variance are sufficient. Cell counts are not sufficient when, for example, P is a continuous distribution. This is

because, if one only considers the cell counts, one throws away information on the distribution within a cell. Indeed, when one compares Figures 2 and 3 (recall that in Figure 3 we shifted the sample one unit to the left), one sees that, by using just 10 cells instead of 20, the strong decrease in the second half of the first cell is no longer visible.

Sufficiency depends very heavily on the model for P . Clearly, when one decides to ignore information because of a sufficiency argument, one may be ignoring evidence that the model's assumptions may be wrong. Sufficiency arguments should be treated with caution.

References

- [1] Bickel, P.J. & Doksum, K.A. (2001). *Mathematical Statistics*, 2nd Edition, Holden-Day, San Francisco.
- [2] Pareto, V. (1897). *Course d'Économie Politique*, Rouge, Lausanne et Paris.

SARA A. VAN DE GEER

Eta and Eta Squared

ANDY P. FIELD

Volume 2, pp. 553–554

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Eta and Eta Squared

Eta-squared (η^2) is an **effect size** measure (typically the effect of manipulations across several groups). When statistical models are fitted to observations, the fit of the model is crucial. There are three main sources of variance that can be calculated: the total variability in the observed data, the model variability (the variability that can be explained by the model fitted), and the residual or error variability (the variability unexplained by the model). If sums of squared errors are used as estimates of variability, then the total variability is the total sum of squared errors (SS_T), that is, calculate the deviation of each score from the mean of all scores (the grand mean), square it, and then add these squared deviations together:

$$SS_T = \sum_{i=1}^n (x_i - \bar{x}_{\text{grand}})^2. \quad (1)$$

This can also be expressed in terms of the variance of all scores: $SS_T = s_{\text{grand}}^2(N - 1)$.

Once a model has been fitted, this total variability can be partitioned into the variability explained by the model, and the error. The variability explained by the model (SS_M) is the sum of squared deviations of the values predicted by the model and the mean of all observations:

$$SS_M = \sum_{i=1}^n (\hat{x}_i - \bar{x}_{\text{grand}})^2. \quad (2)$$

Finally, the residual variability (SS_R) can be obtained through subtraction ($SS_R = SS_T - SS_M$), or for a more formal explanation, see [3].

In **regression models**, these values can be used to calculate the proportion of variance that the model explains (SS_M/SS_T), which is known as the coefficient of determination (R^2). Eta squared is the same but calculated for models on the basis of group means. The distinction is blurred because using group means to predict observed values is a special case of a regression model (see [1] and [3], and **generalized linear models (GLM)**).

As an example, we consider data from Davey et al. [2] who looked at the processes underlying Obsessive Compulsive Disorder by inducing negative, positive, or no mood in people and then asking

them to imagine they were going on holiday and to generate as many things as they could that they should check before they went away (see Table 1).

The total variability can be calculated from the overall variance and the total number of scores (30):

$$SS_T = s_{\text{grand}}^2(N - 1) = 21.43(30 - 1) = 621.47. \quad (3)$$

The model fitted to the data (the predicted values) is the group means. Therefore, the model sum of squares can be rewritten as:

$$SS_M = \sum_{i=1}^n (\bar{x}_i - \bar{x}_{\text{grand}})^2, \quad (4)$$

in which $\bar{x}_i$ is the mean of the group to which observation i belongs. Because there are multiple observations in each group, this can be simplified still further:

$$SS_M = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x}_{\text{grand}})^2, \quad (5)$$

where k is the number of groups. We would get:

$$\begin{aligned} SS_M &= 10(12.60 - 9.43)^2 + 10(7.00 - 9.43)^2 \\ &\quad + 10(8.70 - 9.43)^2 \\ &= 164.87. \end{aligned} \quad (6)$$

Eta squared is simply:

$$\eta^2 = \frac{SS_M}{SS_T} = \frac{164.87}{621.47} = 0.27. \quad (7)$$

Table 1 Number of ‘items to check’ generated under different moods

	Negative	Positive	None
	7	9	8
	5	12	5
	16	7	11
	13	3	9
	13	10	11
	24	4	10
	20	5	11
	10	4	10
	11	7	7
	7	9	5
$\bar{X}$	12.60	7.00	8.70
s^2	36.27	8.89	5.57
Grand Mean = 9.43 Grand Variance = 21.43			

2 Eta and Eta Squared

The literal interpretation is that by fitting the group means to the data, 27% of the variability in the number of items generated can be explained. This is the percentage reduction in error (PRE). Eta squared is accurate and unbiased when describing the sample; however, it is biased as a measure of the effect size in the population because there is sampling error associated with each of the group means that is not reflected in η^2 . Finally, the unsquared Eta (η) can be thought of as the correlation coefficient associated with a curvilinear line connecting the group means. It should be apparent that when groups are unordered, this statistic is not particularly useful.

References

- [1] Cohen, J. (1968). Multiple regression as a general data-analytic system, *Psychological Bulletin* **70**(6), 426–443.
- [2] Davey, G.C.L., Startup, H.M., Zara, A., MacDonald, C.B. & Field, A.P. (2003). The perseveration of checking thoughts and mood-as-input hypothesis, *Journal of Behavior Therapy and Experimental Psychiatry* **34**(2), 141–160.
- [3] Field, A.P. (2005). *Discovering Statistics Using SPSS*, 2nd Edition, Sage, London.

ANDY P. FIELD

Ethics in Research

RICHARD MILLER

Volume 2, pp. 554–562

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ethics in Research

Integrity in conducting research is important to individual researchers, their institutional sponsors, and the public at large. The increasing importance of the pursuit of new knowledge that relies on systematic, empirical investigation has led to greater demands for accountability. Research helps people make sense of the world in which they live and the events they experience. The knowledge gained through psychological research has provided many practical benefits as well as invaluable insights into the causes of human behavior. Despite the obvious advantages of the knowledge provided by research, the process of conducting scientific research can pose serious ethical dilemmas. Because research is a complex process, well-intentioned investigators can overlook the interests of research participants, thereby causing harm to the participants, scientists, science, and society.

A Historical Review of Research Ethics

Regulations affecting the research process can be seen as early as the fourteenth century when Pope Boniface VIII prohibited the cutting up of dead bodies, which was necessary to prevent knights from boiling the bodies of their comrades killed in the Holy Land in order to send them home for burial. While the practice was unrelated to medical research, it nevertheless, had an affect on scientific inquiry for centuries. More systematic regulation of research came about partly because of the atrocities committed by Nazi investigators conducting concentration camp experiments. At the end of World War II, 23 Nazi researchers, mostly physicians, were tried before the Nuremberg Military Tribunal. At the trial, it was important for the prosecutors to distinguish between the procedures used in Nazi experiments and those used by US wartime investigators. To do this, the judges agreed on 10 basic principles for research using human participants. Many of the principles set forth in the Nuremberg Code continue to form the foundation for ethical practices used today, including voluntary consent of the human participant, the avoidance of unnecessary suffering or injury, limitations on the degree of risk allowed, and the opportunity for the research participant to withdraw. While these principles were considered

laudatory, many American investigators viewed them as relevant only to Nazi war crimes and the impact on American scientists was minimal.

In the United States, oversight has come about as a result of a history of ethical abuses and exploitation, including the infamous study at the Willowbrook State School for the Retarded where a mild strain of virus was injected into children in order to study the course of the disease under controlled conditions, and the well-publicized Tuskegee Syphilis Study during which African-American men infected with syphilis were denied actual treatment and told not to seek alternative treatment from outside physicians. Such studies have created reasonable doubt as to the benevolence and value of research, especially among members of groups who received unethical treatment.

Federal Regulation of Research

To address these ethical concerns, the National Commission for the Protection of Human Subjects of Biomedical and Behavior Research was created and is best known for the Belmont Report [6], which identifies three basic ethical principles and their application to research: respect for persons, beneficence, and justice. These principles form the basis for provisions related to procedures insuring informed consent, assessment of risk and potential benefits, and selection of participants. In response to the Belmont Report, federal regulation of research became more systematic. While the primary responsibility for the ethical treatment of participants remains with the individual investigator, research in the United States conducted by individuals affiliated with universities, schools, hospitals, and many other institutions is now reviewed by a committee of individuals with diverse backgrounds who examine the proposed research project for any breach of ethical procedures. These review committees, commonly called *Institutional Review Boards* (IRBs), were mandated by the National Research Act, Public Law 93-348, and require researchers to prepare an application or protocol describing various aspects of the research and to submit this protocol along with informed consent forms for approval prior to the implementation of a research project. The review of the proposed research by the IRB includes an examination of the procedure, the nature of the participants, and other relevant factors in the research design. The IRB also

identifies the relevant ethical issues that may be of concern and decides what is at stake for the participant, the researcher, and the institution with which the researcher is affiliated. If there are ethical concerns, the IRB may suggest alternatives to the proposed procedures. Finally, the IRB will provide the researcher with a formal statement of what must be changed in order to receive IRB approval of the research project.

The attempt by IRBs to ensure ethical practices has caused some dissatisfaction among scientists. Since IRBs are not federal agencies but are instead created by local institutions, they have come under criticism for (a) lack of standard procedures and requirements; (b) delays in completing the review process; (c) creating the fear that IRBs will impose institutional sanctions on individual researchers; and (d) applying rules originally designed for medical studies to behavioral science research projects without acknowledging the important differences between the two. To address these concerns, IRBs should require both board members and principal investigators to undergo training in research ethics, adopt more consistent guidelines for evaluating research protocols, place limits on the power given to the IRB, include an evaluation of the technical merit of a proposal as a means of determining risk/benefit ratios, develop a series of case studies to help sensitize members of an IRB to ethical dilemmas within the social sciences and ways they may be resolved, encourage the recruitment of women, minorities, and children as research participants, adopt provisions that ensure students be given alternatives to participation in research when the research is a class requirement, and carefully review cases where a financial conflict of interest may occur [7].

Ethical Concerns in Recruiting Participants

One of the first ethical issues a researcher must address involves the recruitment of research participants. In the recruitment process, researchers must be guided by the principles of autonomy, respect for persons, and the principle of beneficence that requires that researchers minimize the possible harm to participants while maximizing the benefits from the research. The first stage in the recruitment of participants is often an advertisement for the research project. The advertisement generally describes the

basic nature of the research project and the qualifications that are needed to participate. At this stage, ethical concerns include the use of inducements and coercion, consent and alternatives to consent, institutional approval of access to participants, and rules related to using student subject pools [1]. It is important that researchers avoid ‘hyperclaiming,’ in which the goals the research is likely to achieve are exaggerated. It is also important that researchers not exploit potential participants, especially vulnerable participants, by offering inducements that are difficult to refuse. At the same time, researchers must weigh the costs to the participant and provide adequate compensation for the time they spend in the research process.

Most psychological research is conducted with students recruited from university subject pools, which raises an ethical concern since the students’ grades may be linked with participation. Ethical practice requires that students be given a reasonable alternative to participation in order to obtain the same credit as those who choose to participate in research. The alternatives offered must not be seen by students as either punitive or more stringent than research participation.

In the recruitment process, researchers should attempt to eliminate any potential participants who may be harmed by the research. Research protocols submitted to an IRB typically have a section in which the researcher describes this screening process and the criteria that will be used to include or exclude persons from the study. The screening process is of particular importance when using proxy decisions for incompetent persons and when conducting clinical research. On the other hand, it is important that the sample be representative of the population to which the research findings can be generalized.

Informed Consent and Debriefing

Informed consent is the cornerstone of ethical research. Consent can be thought of as a contract in which the participant agrees to tolerate experimental procedures that may include boredom, deception, and discomfort for the good of science, while the researcher guarantees the safety and well-being of the participant. In all but minimal-risk research, informed consent is a formal process whereby the relevant aspects of the research are described along with the obligations and responsibilities of both the participant

and the researcher. An important distinction is made between 'at risk' and 'minimal risk.' Minimal risk refers to a level of harm or discomfort no greater than that which the participant might expect to experience in daily life. Research that poses minimal risk to the participant is allowed greater flexibility with regard to informed consent, the use of deception, and other ethically questionable procedures. Although, it should still meet methodological standards to ensure that the participants' time is not wasted.

Informed consent presents difficulties when the potential participants are children, the participants speak a different language than the experimenter, or the research is therapeutic but the participants are unable to provide informed consent. Certain research methodologies make it difficult to obtain informed consent, as when the methodology includes disguised observation or other covert methods. The omission of informed consent in covert studies can be appropriate when there is a need to protect participants from nervousness, apprehension, and in some cases criminal prosecution. Studies that blur the distinction between consent for treatment or therapy and consent for research also pose ethical problems as can the use of a consent form that does not provide the participant with a true understanding of the research. While most psychological research includes an informed consent process, it should be noted that federal guidelines permit informed consent to be waived if (a) the research involves no more than minimal risk to the participants; (b) the waiver will not adversely affect the rights and welfare of the participants; and (c) the research could not be feasibly conducted if informed consent were required [4].

The Use of Deception in Psychological Research

At one time, deception was routinely practiced in behavioral science research, and by the 1960s research participants, usually college students, expected deception and as a result sometimes produced different results than those obtained from unsuspecting participants. In general, psychologists use deception in order to prevent participants from learning the true purpose of the study, which might in turn affect their behavior. Many forms of deception exist, including the use of an experimental confederate posing as another participant, providing false feedback to participants, presenting two related studies as unrelated,

and giving incorrect information regarding stimuli. The acceptability of deception remains controversial although the practice is common. Both participants and researchers tend to conduct a kind of cost-benefit analysis when assessing the ethics of deception. Researchers tend to be more concerned about the dangers of deception than do research participants. Participants' evaluation of studies that use deception are related to the studies' scientific merit, value, methodological alternatives, discomfort experienced by the participants, and the efficacy of the debriefing procedures.

Several alternatives to using deception are available. Role-playing and simulation can be used in lieu of deception. In field research, many researchers have sought to develop reciprocal relationships with their participants in order to promote acceptance of occasional deception. Such reciprocal relationships can provide direct benefits to the participants as a result of the research process. In cases where deception is unavoidable, the method of assumed consent can be used [3]. In this approach, a sample taken from the same pool as the potential participants is given a complete description of the proposed study, including all aspects of the deception, and asked whether they would be willing to participate in the study. A benchmark of 95% agreement allows the researcher to proceed with the deception manipulation.

Avoiding Harm: Pain and Suffering

Participants' consent is typically somewhat uninformed in order to obtain valid information untainted by knowledge of the researcher's hypothesis and expectations. Because of this lack of full disclosure, it is important that the researcher ensures that no harm will come to the participant in the research process. Protection from harm is a foundational issue in research ethics. Types of harm that must be considered by the researcher include physical harm, psychological stress, feelings of having one's dignity, self-esteem, or self-efficacy compromised, or becoming the subject of legal action. Other types of potential harm include economic harm, including the imposition of financial costs to the participants, and social harms that involve negative affects on a person's interactions or relationships with others. In addition to considering the potential harm that may accrue to the research participant, the possibility of harm to the

participants' family, friends, social group, and society must be considered.

While conducting research, it is the researcher's responsibility to monitor actual or potential harm to the participant in case the level of harm changes during the course of the research. One way that the level of potential harm can change is as a result of a mistake made by the researcher. In the case of increased likelihood of harm, the researcher should inform the participant and remind him or her that voluntary withdrawal without penalty is available.

A particular kind of harm addressed in the 1992 American Psychological Association (APA) Code of Ethics [2] is the harm caused by culturally incompetent researchers whose perceptions of gender and race are misinformed by the dominant group's view of social reality. Research designs constructed by researchers with uninformed views can reinforce negative stereotypes about the group studied. One way to avoid this ethical bias is to view research participants as partners as opposed to subjects in the research process. The perception of partnership can be fostered by taking the participants into the researchers' confidence, providing a thorough debriefing and the opportunity for further involvement in a role other than that of a 'subject.' Another type of harm, of special concern to those engaged in field research, is the harm that can result from disclosure of uncensored information.

While psychological research into certain processes, for example, anxiety, depends on the arousal of some discomfort in the participant, it is the responsibility of the researcher to look for ways to minimize this discomfort. In many situations, discomfort is inherent in what is being studied. When nothing can be done to eliminate this type of discomfort, some ways that may minimize the psychological consequences of the discomfort include full and candid disclosure of the experimental procedures, providing opportunities for the participant to withdraw, and ensuring that there are no lingering ill effects of the discomfort. One particular type of lingering ill effect relates to the possibility of embarrassment that participants can experience as a result of their behavior during the research process. To protect participants from this type of harm, it is essential that researchers employ procedures to maintain confidentiality.

Maintaining Confidentiality

Respecting the privacy of the research participant involves much more than just obtaining informed consent. Confidentiality is a complex, multifaceted issue. It involves an agreement, implicit as well as explicit, between the researcher and the participant regarding disclosure of information about the participant and how the participant's data will be handled and transmitted. The participant has the right to decide what information will be disclosed, to whom it will be disclosed, under what circumstances it will be disclosed, and when it will be disclosed.

Participants must be informed about mandatory reporting requirements, for example, illegal activity, plans for sharing information about the participant with others, and the extent to which confidentiality can be legally protected. It is the responsibility of review committees to ensure that the proposed research procedures will not unintentionally compromise confidentiality, especially with participants who are vulnerable because of age, gender, status, or disability.

There are exceptions to the rule regarding confidentiality. The 1992 APA Code of Ethics allows for a breach of confidentiality to protect third parties, and several states have embraced the Supreme Court ruling in *Tarasoff versus Board of Regents of the University of California* [9] that requires the psychologist to take reasonable steps to protect potential victims. Researchers not trained in clinical diagnosis can find themselves in a difficult position interpreting the likelihood of harm from statements made by research participants.

New technologies, along with government statutes and access by third parties to data, can threaten confidentiality agreements, although both state and federal courts have been willing to uphold promises of confidentiality made to research participants. Techniques to maintain confidentiality of data include data encryption and electronic security. While most quantified data are presented in aggregate form, some types of data such as video recordings, photographs, and audio recordings require special care in order to protect participants' privacy. Distortion of the images and sounds can be done, but the most important safeguard is to obtain permission from the participant to use the material, including the dissemination of the findings.

Similarly, qualitative research poses special difficulties for maintaining privacy and confidentiality. Techniques for maintaining confidentiality include the use of pseudonyms or fictitious biographies and the coding of tapes and other data recording methods in which participant identification cannot be disguised. Also, it is the researchers' responsibility to take reasonable precautions to ensure that participants respect the privacy of other participants, particularly in research settings where others are able to observe the behavior of the participant.

Assessing Risks and Benefits

One of the responsibilities of an IRB is to ask the question: will the knowledge gained from this research be worth the inconvenience and potential cost to the participant? Both the magnitude of the benefits to the participant and the potential scientific and social value of the research must be considered [5]. Some of the potential types of benefits of psychological research are (a) an increase in basic knowledge of psychological processes; (b) improved methodological and assessment procedures; (c) practical outcomes and benefits to others; (d) benefits for the researchers, including the educational functions of research in preparing students to think critically and creatively about their field; and (e) direct, sometimes therapeutic, benefits to the participants, for example, in clinical research.

Some of the potential costs to the participant are social and physical discomfort, boredom, anxiety, stress, loss of self-esteem, legal risks, economic risks, social risks, and other aversive consequences. In general, the risks associated with the research should be considered from the perspective of the participant, the researcher, and society as a whole, and should include an awareness that the risks to the participant may come not only from the research process, but also from particular vulnerabilities of the participant or from the failure of the researcher to use appropriate strategies to reduce risk.

The IRB's job of balancing these costs and benefits is difficult since the types of costs and benefits are so varied. The deliberations of the IRB in arriving at a 'favorable ratio' should be formed with respect to the guidelines provided in the Belmont Report, which encourages ethical review committees to examine all aspects of the research carefully and to

consider, on behalf of the researcher, alternative procedures to reduce risks to the participants. The careful deliberation of the cost/benefit ratio is of particular importance in research with those unable to provide informed consent, such as the cognitively impaired; research where there is risk without direct benefit to the participant, research with such vulnerable populations as children and adolescents; and therapeutic research in which the participant in need of treatment is likely to overestimate the benefit and underestimate the risk, even when the researcher has provided a full and candid description of the likelihood of success and possible deleterious effects.

Ethical Issues in Conducting Research with Vulnerable Populations

An important ethical concern considered by IRBs is the protection of those who are not able fully to protect themselves. While determining vulnerability can be difficult, several types of people can be considered vulnerable for research purposes, including people who (a) either lack autonomy and resources or have an abundance of resources, (b) are stigmatized, (c) are institutionalized, (d) cannot speak for themselves, (e) engage in illegal activities, and (f) may be damaged by the information revealed about them as a result of the research. One of the principle groups of research participants considered to be vulnerable is children and adolescents. In addition to legal constraints on research with minors adopted by the United States Department of Health and Human Services (DHHS), ethical practices must address issues of risk and maturity, privacy and autonomy, parental permission and the circumstances in which permission can be waived, and the assent of the institution (school, treatment facility) where the research is to be conducted.

Other vulnerable groups addressed in the literature include minorities, prisoners, trauma victims, the homeless, Alzheimer's patients, gays and lesbians, individuals with AIDS and STDs, juvenile offenders, and the elderly, particularly those confined to nursing homes where participants are often submissive to authority.

Research with psychiatric patients poses a challenge to the researcher. A major ethical concern with clinical research is how to form a control group without unethically denying treatment to some participants, for example, those assigned to a placebo

control group. One alternative to placebo-controlled trials is active-controlled trials.

A number of ethical issues arise when studying families at risk and spousal abuse. It is the responsibility of the investigator to report abuse and neglect, and participants must understand that prior to giving consent. Other ethical issues include conflict between research ethics and the investigator's personal ethics, identifying problems that cannot be solved, and balancing the demands made by family members and the benefits available to them.

Alcohol and substance abusers and forensic patients present particular problems for obtaining adequate informed consent. The researcher must take into account the participants' vulnerability to coercion and competence to give consent. The experience of the investigator in dealing with alcoholics and drug abusers can be an important element in maintaining ethical standards related to coercion and competence to give consent.

One final vulnerable population addressed in the literature is the cognitively impaired. Research with these individuals raises issues involving adult guardianship laws and the rules governing proxy decisions. The question is: who speaks for the participant? Research with vulnerable participants requires the researcher to take particular care to avoid several ethical dilemmas including coercive recruiting practices, the lack of confidentiality often experienced by vulnerable participants, and the possibility of a conflict of interest between research ethics and personal ethics.

Ethical Considerations Related to Research Methodology

Ethical Issues in Conducting Field Research

Research conducted in the field confronts an additional ethical dilemma not usually encountered in laboratory studies. Often the participants are unaware that they are being studied, and therefore no contractual understanding can exist. In many field studies, especially those that involve observational techniques, informed consent may be impossible to obtain. This dilemma also exists when the distinction between participant and observer is blurred. Similarly, some laboratory experiments involving deception use procedures similar to field research in introducing the independent variable as

unrelated to the experiment. Covert research that involves the observation of people in public places is not generally considered to constitute an invasion of privacy; however, it is sometimes difficult to determine when a reasonable expectation of privacy exists, for example, behavior in a public toilet.

Because it is not usually possible to assess whether participants have been harmed in covert studies, opinions regarding the ethicality and legality of such methods varies markedly. Four principles that must be considered in deciding on the ethicality of covert field research are (a) the availability of alternative means for studying the same question, (b) the merit of the research question, (c) the extent to which confidentiality or anonymity can be maintained, and (d) the level of risk to the uninformed participant.

One specific type of field research warrants special ethical consideration: socially sensitive research, which is defined as research where the findings can have practical consequences for the participants. The research question, the research process, and the potential application of the research findings are particularly important in socially sensitive research. IRBs have been found to be very wary of socially sensitive research, more often finding fault with the research and overestimating the extent of risk involved as compared to their reviews of less sensitive research. Despite these difficulties, socially sensitive research has considerable potential for addressing many of society's social issues and should be encouraged.

Ethical Issues in Conducting Archival Research

Archival research can provide methodological advantages to the researcher in that unobtrusive measures are less likely to affect how participants behave. However, research involving archival data poses a problem for obtaining informed consent, since the research question may be very different from the one for which the data was originally collected. In most cases, issues of privacy do not exist since an archive can be altered to remove identifying information. A second ethical concern with archival research has to do with the possibility that those who create the archive may introduce systematic bias into the data set. This is of particular concern when the archive is written primarily from an official point of view that may not accurately represent the participants' attitudes, beliefs, or behavior.

Ethical Issues in Conducting Internet Research

The Internet provides an international forum in which open and candid discussions of a variety of issues of interest to behavioral scientists take place (see **Internet Research Methods**). These discussions provide an opportunity for the behavioral scientist to 'lurk' among Usenet discussion groups, Internet Relay Chat, and Multiuser dungeons. Cyberspace is typically thought of as public domain where privacy is not guaranteed and traditional ethical guidelines may be difficult to apply. A second ethical concern in Internet research is the possibility for on-line misrepresentation. For example, children or other vulnerable populations could be inadvertently included in research. To address these concerns, a set of informal guidelines for acceptable behavior in the form of netiquettes has been developed. Among other things, the guidelines suggest that researchers should identify themselves, ensure confidential treatment of personal information, be sensitive to possible unanticipated consequences to participants as a result of the research process, particularly in terms of potential harm to the participant in the form of stress, legal liabilities, and loss of self-esteem, obtain consent from those providing data whenever possible, and provide participants with information about the study.

Debriefing

Debriefing provides the participant an opportunity to discuss the findings of the study. The need to adequately debrief participants in a research study is a clear ethical responsibility of the investigator although it is still the exception rather than the rule. Debriefing can serve four purposes. It can (a) remove fraudulent information about the participant given during the research process, (b) desensitize subjects who have been given potentially disturbing information about themselves, (c) remove the participants' negative arousal resulting from the research procedure, and (d) provide therapeutic or educational value to the participant. Even participants who are screened out of a study or voluntarily withdraw from a study should be debriefed and told why they might have been eliminated from the study. It has also been suggested that a description of the debriefing procedure be included in any scientific publication of the research.

The Use of Animals in Research

Animal research by psychologists can be dated back to rat maze studies at Clark University in 1901. Many medical breakthroughs including such procedures as chemotherapy, vaccines, bypass surgery, and antibiotics are based on animal studies. While animal research will remain controversial among many people, the first federal attempt to provide standards began with the Animal Welfare Act of 1966. Building upon this foundation, the American Psychological Association has also established ethical standards for the 'humane care and use of animals in research.' It is the responsibility of the research scientist to observe all appropriate laws and regulations and professional standards in acquiring, caring for, and disposing of animals. The research scientist must also ensure that those working with animals are trained and experienced in both research methods and animal care in order to provide for the comfort, health, and humane treatment of the animals. A third responsibility is to minimize the discomfort, infection, illness, and pain of the animals involved in research and to only subject animals to pain, stress, or privation when an alternative procedure is not available, and even then only when the potential value of the research makes the negative treatment justifiable. Scientists involved in surgical procedures with animals have a special responsibility to use appropriate anesthesia and procedures both during and after surgery in order to minimize pain and possible infection. Finally, when the life of a research animal is to be terminated, it must be done in a manner designed to minimize pain and observe accepted procedures. In order to promote and ensure the ethical treatment of animals in research, most research facilities and universities have animal review committees (IACUCs) that perform a function similar to the IRB. These committees can judge the adequacy of the procedures being proposed, the training and experience of the investigators, and whether nonanimal models could be used to answer the questions being posed.

Ethical Issues When the Research is Completed

Plagiarism occurs when an investigator uses the work of someone else without giving credit to the original author. There are several steps that the researcher can

take to avoid this ethical breach, including (a) careful acknowledgement of all sources, including secondary sources of information, (b) use of quotation marks to set off direct quotes and taking care that paraphrasing another author is not simply a minor variation of the author's own words, and (c) maintaining complete records of rough notes, drafts, and other materials used in preparing a report.

Several notorious cases, including that of Cyril Burt, have clearly demonstrated the ethical ban on the falsification and fabrication of data, as well as the misuse of statistics to mislead the reader. In addition to fabrication, it is unethical to publish, as original data, material that has been published before. It is also the ethical responsibility of the investigator to share research data for verification. While these are fairly straightforward ethical considerations, it is important to distinguish between honest errors and misconduct in statistical reporting. Currently, there are no federal guidelines that inform our understanding of the differences between common practices and actual misuse. Therefore, it is important that individual investigators consult with statisticians in order to apply the most appropriate tests to their data.

Authorship credit at the time of publication should only be taken for work actually performed and for substantial contributions to the published report. Simply holding an institutional position is not an ethical reason for being included as an author of a report. Students should be listed as the principal author of any article that is primarily based on that student's work, for example, a dissertation.

Summary and Conclusion

Ethical dilemmas often arise from a conflict of interest between the needs of the researcher and the needs of the participant and/or the public at large. A conflict of interest can occur when the researcher occupies multiple roles, for example, clinician/researcher, or within a single role such as a program evaluation researcher who experiences sponsor pressures for results that may compromise scientific rigor. In resolving ethical dilemmas, psychologists are guided in their research practices by APA guidelines as well as Federal regulations that mandate that research be approved by an Institutional Review Board or Institutional Animal Care and Use Committee. A set of

heuristics that can be employed in resolving ethical conflicts include: (a) using the ethical standards of the profession, (b) applying ethical and moral principles, (c) understanding the legal responsibilities placed upon the researcher, and (d) consulting with professional colleagues [8]. In the final analysis, the researcher's conscience determines whether the research is conducted in an ethical manner. To enhance the researcher's awareness of ethical issues, education and training programs have become increasingly available in university courses, workshops, and on governmental websites. The use of role-playing and context-based exercises, and the supervision of student research have been shown to effectively increase ethical sensitivity.

References

- [1] American Psychological Association. (1982). *Ethical Principles in the Conduct of Research with Human Participants*, American Psychological Association, Washington.
- [2] American Psychological Association. (1992). Ethical principles of psychologists & code of conduct, *American Psychologist* **42**, 1597–1611.
- [3] Cozby, P.C. (1981). *Methods in Behavioral Research*, Mayfield, Palo Alto.
- [4] Fischman, M.W. (2000). Informed consent, in *Ethics in Research with Human Participants*, B.D. Sales & S. Folkman, eds, American Psychological Association, Washington, pp. 35–48.
- [5] Fisher, C.B. & Fryberg, D. (1994). Participant partners: college students weigh the costs and benefits of deceptive research, *American Psychologist* **49**, 417–427.
- [6] Office for Protection From Research Risks, Protection of Human Subjects, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical Principles and Guidelines in the Protection of Human Subjects*. (GPO 887-809) U.S. Government Printing Office, Washington.
- [7] Rosnow, R., Rotheram-Borus, M.J., Ccci, S.J., Blanck, P.D. & Koocher, G.P. (1993). The institutional review board as a mirror of scientific and ethical standards, *American Psychologist* **48**, 821–826.
- [8] Sales, B. & Lavin, M. (2000). Identifying conflicts of interest and resolving ethical dilemmas, in *Ethics in Research with Human Participants*, B.D. Sales & S. Folkman, eds, American Psychological Association, Washington, pp. 109–128.
- [9] Tarasoff V. Board of Regents of the University of California (1976). 17 Cal. 3d 425, 551 P.2d 334.

RICHARD MILLER

Evaluation Research

MARK W. LIPSEY AND SIMON T. TIDD

Volume 2, pp. 563–568

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Evaluation Research

From a methodological standpoint, the most challenging task an evaluation researcher faces is determining a program's effects on the social conditions it is expected to ameliorate. This is also one of the most important tasks for the evaluator because a program that does not have the anticipated beneficial effects is unlikely to be viewed favorably, irrespective of how well it functions in other regards. Evaluation researchers refer to this form of evaluation as an impact evaluation or outcome evaluation. Because of the centrality of impact evaluation in evaluation research and its dependence on quantitative methods, it is our focus in this essay.

Impact Evaluation

The basic question asked by an impact evaluation is whether the program produces its intended effects. This is a causal question and the primary method for answering it is an experiment that compares the outcomes for an intervention condition to those from a control condition without the intervention. Evaluation research is conducted largely in field settings where 'true' experiments with random assignment and strict control of extraneous influences are difficult to achieve. Evaluation researchers, therefore, often fall back on quasi-experiments, (see **Quasi-experimental Designs**) most commonly nonequivalent comparison designs lacking random assignment [18]. The application of these designs to social programs raises a variety of methodological challenges. We turn now to some of the most salient of these challenges.

Outcome Measurement

No systematic assessment of a program's effects can be made unless the intended outcomes can be measured. Program representatives often describe the expected outcomes in broad terms (e.g., 'improve the quality of life for children') that must be unpacked to identify their specific observable features (e.g., improved health, cognitive development, and social relationships). A necessary first step, therefore, is usually a negotiation between the evaluator and the

program stakeholders to carefully specify the changes the program is expected to bring about and the indicators that signal whether they occurred. Once adequately specified, these outcomes can often be measured using established procedures; for example, standardized achievement tests and grade point averages are conventional measures of academic performance. However, there may be no established valid and reliable measures for some outcomes and the evaluator must then attempt to develop them.

In addition to validity and reliability, however, evaluators have to be concerned with another measurement characteristic – sensitivity, or the extent to which scores on a measure change when a change actually occurs on an outcome the program is attempting to affect. There are two main ways in which an outcome measure can be insensitive to change. First, the measure may include elements that relate to something other than what the program targets. Consider, for example, a math tutoring program concentrating on fractions and long division problems for elementary school children. The evaluator might choose an off-the-shelf math achievement test as an outcome measure even though it covers a wider selection of math problems than fractions and long division. Large gains in fractions and long division might be obscured by the response to other topics that are averaged into the final score. A measure that covered only the math topics that the program actually taught would be more sensitive to these gains.

Second, outcome measures may be insensitive to program effects if they have been developed to differentiate individuals for diagnostic purposes. Most standardized psychological tests are of this sort, including, for example, measures of personality traits, clinical symptoms, cognitive abilities, and attitudes. These measures are generally good for determining who is high or low on the characteristic measured. However, when applied to program participants who differ on the measured characteristic, they may yield such wide variation in scores that improvement due to the program is lost amid the differences among individuals.

Unit of Analysis

Social programs deliver their services to any of a wide variety of entities, such as individuals, families, schools, neighborhoods, or cities. Correspondingly,

the units in the research sample may be any of these entities. It is not unusual for the program to deliver its services to one level with the intent of producing effects on units nested within this level. This situation occurs frequently in educational programs. A mathematics curriculum, for instance, may be implemented school wide and delivered mainly at the classroom level. The desired outcome, however, is improved math achievement for the students in those classes. Students can be sampled only by virtue of being in a classroom that is, or is not, using the curriculum of interest. Thus, the classroom is the primary sampling unit but the students clustered within the classrooms are of focal interest for the evaluation and are the primary analysis unit.

A common error is to analyze the outcome data at the student level, ignoring the clustering of students within classrooms. This error exaggerates the sample size used in the statistical analysis by counting the number of students rather than the number of classrooms that are the actual sampling unit. It also treats the student scores within each classroom as if they were independent data points when, because of the students' common classroom environment and typically nonrandom assignment to classrooms, their scores are likely to be more similar within classrooms than they would be otherwise. This situation requires the use of specialized multilevel statistical analysis models (*see* **Linear Multilevel Models**) to properly estimate the standard errors and determine the statistical significance of any effects (for further details, see [13, 19].

Selection Bias

When an impact evaluation involves an intervention and control group that show preintervention differences on one or more variables related to an outcome of interest, the result is a postintervention difference that mimics a true intervention effect. Initial nonequivalence of this sort biases the estimate of the intervention effects and undermines the validity of the design for determining the actual program effects. This serious and unfortunately common problem is called *selection bias* because it occurs in situations in which units have been differentially selected into the intervention and control groups.

The best way to achieve equivalence between intervention and control groups is to randomly allocate members of a research sample to the groups (see [2] for a discussion of how to implement randomization) (*see* **Randomization**). However, when intervention and control groups cannot be formed through random assignment, evaluators may attempt to construct a matched control group by selecting either individuals or an aggregate group that is similar on a designated set of variables to those receiving the intervention. In individual **matching**, a partner is selected from a pool of individuals not exposed to the program who matches each individual who does receive the program. For children in a school drug prevention program, for example, the evaluator might deem the relevant matching variables to be age, sex, and family income. In this case, the evaluator might scrutinize the roster of unexposed children at a nearby school for the closest equivalent child to pair with each child participating in the program.

With aggregate matching, individuals are not matched case by case; rather, the overall distributions of values on the matching variables are made comparable for the intervention and control groups. For instance, a control group might be selected that has the same proportion of children by sex and age as the intervention group, but this may involve a 12-year-old girl and an 8-year-old boy in the control group to balance a 9-year-old girl and an 11-year-old boy in the intervention group. For both matching methods, the overall goal is to equally distribute characteristics that may impact the outcome variable. As a further safeguard, additional descriptive variables that have not been used for matching may be measured prior to intervention and incorporated in the analysis as statistical controls (discussed below).

The most common impact evaluation design is one in which the outcomes for an intervention group are compared with those of a control group selected on the basis of relevance and convenience. For a community-wide program for senior citizens, for instance, an evaluator might draw a control group from a similar community that does not have the program and is convenient to access. Because any estimate of program effects based on a simple comparison of outcomes for such groups must be presumed to include selection bias, this is a nonequivalent comparison group design.

Nonequivalent control (comparison) group designs are analyzed using statistical techniques that

attempt to control for the preexisting differences between groups. To apply statistical controls, the control variables must be measured on both the intervention and comparison groups before the intervention is administered. A significant limitation of both matched and nonequivalent comparison designs is that the evaluator generally does not know what differences there are between the groups nor which of those are related to the outcomes of interest.

With relevant control variables in hand, the evaluator must conduct a statistical analysis that accounts for their influence in a way that effectively and completely removes selection bias from the estimates of program effects. Typical approaches include **analysis of covariance** and **multiple linear regression** analysis. If all the relevant control variables are included in these analyses, the result should be an unbiased estimate of the intervention effect.

An alternate approach to dealing with nonequivalence that is becoming more commonplace is selection modeling. Selection modeling is a two-stage procedure in which the first step uses relevant control variables to construct a statistical model that predicts selection into the intervention or control group. This is typically done with a specialized form of regression analysis for binary dependent variables, for example, **probit** or **logistic regression**. The results of this first stage are then used to combine all the control variables into a single composite selection variable, or **propensity score** (propensity to be selected into one group or the other). The propensity score is optimized to account for the initial differences between the intervention and control groups and can be used as a kind of super control variable in an analysis of covariance or multiple regression analysis. Effective selection modeling depends on the evaluator's diligence in identifying and measuring variables related to the process by which individuals select themselves (e.g., by volunteering) or are selected (e.g., administratively) into the intervention or comparison group. Several variants of selection modeling and two-stage estimation of program effects are available. These include Heckman's econometric approach [6, 7], Rosenbaum and Rubin's propensity scores [14, 15], and **instrumental variables** [5].

The Magnitude of Program Effects

The ability of an impact evaluation to detect and describe program effects depends in large part on

the magnitude of those effects. Small effects are more difficult to detect than large ones and their practical significance may also be more difficult to describe. Evaluators often use an **effect size** statistic to express the magnitude of a program effect in a standardized form that makes it comparable across measures that use different units or different scales. The most common effect size statistic is the standardized mean difference (sometimes symbolized d), which represents a mean outcome difference between an intervention group and a control group in standard deviation units. Describing the size of a program effect in this manner indicates how large it is relative to the range of scores recorded in the study. If the mean reading readiness score for participants in a preschool intervention program is half a standard deviation larger than that of the control group, the standardized mean difference effect size is 0.50. The utility of this value is that it can be easily compared to, say, the standardized mean difference of 0.35 for a test of vocabulary. The comparison indicates that the preschool program was more effective in advancing reading readiness than in enhancing vocabulary.

Some outcomes are binary rather than a matter of degree; that is, for each participant, the outcome occurs or it does not. Examples of binary outcomes include committing a delinquent act, becoming pregnant, or graduating from high school. For binary outcomes, an **odds ratio** effect size is often used to characterize the magnitude of the program effect. An odds ratio indicates how much smaller or larger the odds of an outcome event are for the intervention group compared to the control group. For example, an odds ratio of 1.0 for high school graduation indicates even odds; that is, participants in the intervention group are no more and no less likely than controls to graduate. Odds ratios greater than 1.0 indicate that intervention group members are more likely to experience the outcome event; for instance, an odds ratio of 2.0 means that the odds of members of the intervention group graduating are twice as great as for members of the control group. Odds ratios smaller than 1.0 mean that they are less likely to graduate.

Effect size statistics are widely used in the **meta-analysis** of evaluation studies. Additional information can be found in basic meta-analysis texts such as those found in [4, 10, 16].

The Practical Significance of Program Effects

Effect size statistics are useful for summarizing and comparing research findings but they are not necessarily good guides to the practical magnitude of those effects. A small statistical effect may represent a program effect of considerable practical significance; conversely, a large statistical effect may be of little practical significance. For example, a very small reduction in the rate at which people with a particular illness are hospitalized may have important cost implications for health insurers. Statistically larger improvements in the patients' satisfaction with their care, on the other hand, may have negligible practical implications.

To appraise the practical magnitude of program effects, the statistical effect sizes must be translated into terms relevant to the social conditions the program aims to improve. For example, a common outcome measure for juvenile delinquency programs is the rate of rearrest within a given time period. If a program reduces rearrest rates by 24%, this amount can readily be interpreted in terms of the number of juveniles affected and the number of delinquent offenses prevented.

For other program effects, interpretation may not be so simple. Suppose that a math curriculum for low-performing sixth-grade students raised the mean score from 42 to 45 on the mathematics subtest of the Omnibus Test of Basic Skills, a statistical effect size of 0.30 standard deviation units. How much improvement in math skills does this represent in practical terms? Interpretation of statistical effects on outcome measures with values that are not inherently meaningful requires comparison with some external referent that puts the effect size in a practical context. With achievement tests, for instance, we might compare program effects against test norms. If the national norm on the math test is 50, the math curriculum reduced the gap between the students in the program and the norm by about 38% (from 8 points to 5), but still left them short of the average skill level.

Another referent for interpreting the practical magnitude of a program effect is a 'success' threshold on the outcome measure. A comparison of the proportions of individuals in the intervention and control groups who exceed the threshold reveals the practical magnitude of the program effect. For example, a mental health program that treats depression might

use the Beck Depression Inventory as an outcome measure. On this instrument, scores in the 17 to 20 range indicate borderline clinical depression, so one informative index of practical significance is the percent of patients with posttest scores less than 17. If 37% of the control group is below the clinical threshold at the end of the treatment period compared to 65% of the treatment group, the practical magnitude of this treatment effect can be more easily appraised than if the same difference is presented in arbitrary scale units.

Another basis of comparison for interpreting the practical significance of program effects is the distribution of effect sizes in evaluations of similar programs. For instance, a review of evaluation research on the effects of marriage counseling, or a meta-analysis of the effects of such programs, might show that the mean effect size for marital satisfaction was around 0.46, with most of the effect sizes ranging between 0.12 and 0.80. With this information, an evaluator who finds an effect size of 0.34 for a particular marriage-counseling program can recognize it as rather middling performance for a program of this type.

Statistical Power

Suppose that an evaluator has some idea of the magnitude of the effect that a program must produce to have a meaningful impact and can express it as an effect size statistic. An impact evaluation of that program should be designed so it can detect that effect size. The minimal standard for identifying an effect in a quantitative analysis is that it attains statistical significance. The probability that an estimate of the program effect based on sample data will be statistically significant when, in fact, it represents a real (population) effect of a given magnitude is called statistical **power**. Statistical power is a function of the effect size to be detected, the sample size, the type of statistical significance test used, and the alpha level.

Deciding the proper level of statistical power for an impact assessment is a substantive issue. If an evaluator expects that the program's statistical effects will be small and that such small effects are worthwhile, then a design powerful enough to detect them is needed. For example, the effect of an intervention that lowers automobile accident deaths by as little as 1% might be judged worth detecting

because saving lives is so important. In contrast, when an evaluator judges that an intervention is worthwhile only if its effects are large, a design that lacks power to detect small effects may be quite acceptable. Proficiency in statistical power estimation and its implications for sample size and statistical control variables is critical for competent impact evaluation. More detailed information can be found in [3, 9].

Moderator and Mediator Relationships

The experimental and quasi-experimental designs used for impact evaluation are oriented toward determining whether the program produces effects on specific outcome variables. They reveal little about how and when the program brings about its effects. For instance, program effects are rarely identical for all recipient subgroups and all circumstances of service delivery. Differences in outcomes related to the moderator variables (*see* **Moderation**) that describe these variations must be examined to identify the conditions under which the program is most and least effective. In addition, programs usually produce their effects through a causal chain in which they first affect proximal outcomes that, in turn, change other more distal outcomes. A mass media antidrug campaign, for instance, might attempt to change attitudes and knowledge about drug use with the expectation that such changes will lead to changes in drug-use behavior. Analysis of such intervening variables, or mediator relationships (*see* **Mediation**), helps explain the change mechanisms through which a program produces its effects.

To explore moderator relationships, evaluators examine statistical interactions between the potential moderator variables and the outcomes they may moderate. A simple case would be to divide the research sample into male and female subgroups, determine the mean program effect for each gender, and then compare those effects. If the effects are larger, say, for females than for males, it indicates that gender moderates the program effect. Demographic variables such as gender, age, ethnicity, and socioeconomic status often characterize groups that respond differently to a social program. Moreover, it is not unusual for different program sites, personnel configurations, and procedures to be associated with variations in program effects.

In addition to uncovering differential program effects, evaluators can use moderator analysis to test their expectations about what differential effects should occur. In this use of moderator analysis, the evaluator reasons that, if the program is operating as expected and having effects, these effects should be larger here and smaller there – for example, larger where the behavior targeted for change is most prevalent, where more or better service is delivered, for groups that should be naturally more responsive, and so forth. A moderator analysis that confirms these expectations provides evidence that helps confirm the existence of program effects. A moderator analysis that fails to confirm expectations serves as a caution that there may be influence on the effect estimates other than the program itself.

Testing for mediator relationships hypothesized in the program logic is another way of probing evaluation findings to determine if they are consistent with expectations of a successful program. For example, suppose that the intended outcome of a program, in which adult volunteers mentor at-risk youths, is reductions in the youths' delinquent behavior. The hypothesized causal pathway is that contact with mentors influences the youths to emulate the values of their mentors and use leisure time more constructively. This, in turn, is expected to lead to reduced contact with antisocial peers and, finally, to decreased delinquent behavior. In this hypothesized pathway, constructive use of leisure time is the major mediating variable between program exposure and contact with peers. Contact with peers, similarly, is presumed to mediate the relationship between leisure time use and decreased delinquency.

Statistical procedures for examining mediator relationships assess the relationship between the independent variable and the mediator, the independent variable and the dependent variable, and the mediator and the dependent variable. The critical test is whether the relationship between the independent and dependent variables shrinks toward zero when the mediator is controlled statistically. Mediator relationships are usually tested with **multiple linear regression** analyses; discussions of the statistical procedures for conducting these tests can be found in [1, 11].

More sophisticated analysis procedures available for moderator and mediator analysis include multi-level modeling (*see* **Hierarchical Models**) [13, 19] and **structural equation modeling** [8, 17]. Impact evaluation can be designed to include the variables

needed for such analysis, and these analysis techniques can be combined with those for analyzing experimental designs [12]. By providing the tools to examine how, when, and where program effects are produced, evaluators avoid 'black box' evaluations that determine only whether effects were produced.

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator distinction in social psychological research: conceptual, strategic and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] Boruch, R.F. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Sage Publications, Thousand Oaks.
- [3] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.
- [4] Cooper, H. & Hedges, L.V. (1994). *The Handbook of Research Synthesis*, Russell Sage Foundation, New York.
- [5] Greene, W.H. (1993). Selection-incidental truncation, in *Econometric Analysis*, W.H. Greene, ed., Macmillan Publishing, New York, pp. 706–715.
- [6] Heckman, J.J. & Holtz, V.J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training, *Journal of the American Statistical Association* **84**, 862–880 (with discussion).
- [7] Heckman, J.J. & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: an overview, *Journal of Econometrics* **30**, 239–267.
- [8] Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, Guilford Press, New York.
- [9] Kraemer, H.C. & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*, Sage Publications, Newbury Park.
- [10] Lipsey, M.W. & Wilson, D.B. (2001). *Practical Meta-Analysis*, Sage Publications, Thousand Oaks.
- [11] MacKinnon, D.P. & Dwyer, J.H. (1993). Estimating mediated effects in prevention studies, *Evaluation Review* **17**, 144–158.
- [12] Muthén, B.O. & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation, *Psychological Methods* **2**(4), 371–402.
- [13] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage Publications, Newbury Park.
- [14] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**(1), 41–55.
- [15] Rosenbaum, P.R. & Rubin, D.B. (1983). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**, 516–524.
- [16] Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*, (revised ed.), Sage Publications, Thousand Oaks.
- [17] Schumacker, R.E. & Lomax, R.G. (1996). *A Beginner's Guide to Structural Equation Modeling*, Lawrence Erlbaum, Mahwah.
- [18] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, New York.
- [19] Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, Newbury.

MARK W. LIPSEY AND SIMON T. TIDD

Event History Analysis

JEROEN K. VERMUNT AND GUY MOORS

Volume 2, pp. 568–575

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Event History Analysis

Introduction

The purpose of event history analysis is to explain why certain individuals are at a higher risk of experiencing the event(s) of interest than others. This can be accomplished by using special types of methods which, depending on the field in which they are applied, are called *failure-time models*, lifetime models, survival models (see **Survival Analysis**), transition-rate models, response-time models, event history models, duration models, or hazard models. Examples of textbooks discussing this class of techniques are [1, 2, 5, 7, 8, 10], and [12]. Here, we will use the terms event history, survival, and hazard models interchangeably.

A hazard model is a regression model in which the ‘risk’ of experiencing an event at a certain time point is predicted with a set of covariates. Two special features distinguish hazard models from other types of regression models. The first is that they make it possible to deal with **censored observations** which contain only partial information on the timing of the event of interest. Another special feature is that covariates may change their value during the observation period. The possibility of including such time-varying covariates makes it possible to perform a truly dynamic analysis. Before discussing in more detail the most important types of hazard models, we will first introduce some basic concepts.

State, Event, Duration, and Risk Period

In order to understand the nature of event history data and the purpose of event history analysis, it is important to understand the following four elementary concepts: state, event, duration, and risk period. These concepts are illustrated below using an example from the analyses of marital histories.

The first step in the analysis of event histories is to define the *states* that one wishes to distinguish. States are the categories of the ‘dependent’ variable, the dynamics of which we want to explain. At every particular point in time, each person occupies exactly one state. In the analysis of marital histories, four states are generally distinguished: never married, married, divorced, and widow(er). The set of possible states is sometimes also called the *state space*.

An *event* is a transition from one state to another, that is, from an origin state to a destination state. In this context, a possible event is ‘first marriage’, which can be defined as the transition from the origin state, never married, to the destination state, married. Other possible events are: a divorce, becoming a widow(er), and a non-first marriage. It is important to note that the states that are distinguished determine the definition of possible events. If only the states married and not married were distinguished, none of the above-mentioned events could have been defined. In that case, the only events that could be defined would be marriage and marriage dissolution.

Another important concept is the *risk period*. Clearly, not all persons can experience each of the events under study at every point in time. To be able to experience a particular event, one must occupy the origin state defining the event, that is, one must be at risk of the event concerned. The period that someone is at risk of a particular event, or exposed to a particular risk, is called the *risk period*. For example, someone can only experience a divorce when he or she is married. Thus, only married persons are at risk of a divorce. Furthermore, the risk period(s) for a divorce are the period(s) that a subject is married. A strongly related concept is the *risk set*. The risk set at a particular point in time is formed by all subjects who are at risk of experiencing the event concerned at that point in time.

Using these concepts, event history analysis can be defined as the analysis of the *duration of the nonoccurrence of an event* during the risk period. When the event of interest is ‘first marriage’, the analysis concerns the duration of nonoccurrence of a first marriage, in other words, the time that individuals remained in the state of never being married. In practice, as will be demonstrated below, the dependent variable in event history models is not duration or time itself but a transition rate. Therefore, event history analysis can also be defined as the analysis of rates of occurrence of the event during the risk period. In the first marriage example, an event history model concerns a person’s marriage rate during the period that he/she is in the state of never having been married.

Basic Statistical Concepts

Suppose that we are interested in explaining individual differences in women’s timing of the first birth.

2 Event History Analysis

In that case, the event is having a first child, which can be defined as the transition from the origin state no children to the destination state one child. This is an example of what is called a *single nonrepeatable event*, where the term single reflects that the origin state no children can only be left by one type of event, and the term nonrepeatable indicates that the event can occur only once. For the moment, we concentrate on such single nonrepeatable events, but later on we show how to deal with multiple type and repeatable events.

The manner in which the basic statistical concepts of event history models are defined depends on whether the time variable T – indicating the duration of nonoccurrence of an event – is assumed to be continuous or discrete. Even though it seems in most applications it is most natural to treat T as a continuous variable, sometimes this assumption is not realistic. Often, T is not measured accurately enough to be treated as strictly continuous, for example, when the duration variable in a study on the timing of the first birth is measured in completed years instead of months or days. In other applications, the events of interest can only occur at particular points in time, such as in studies on voting behavior.

Here, we will assume that the T is a continuous random variable, for example, indicating the duration of nonoccurrence of the first birth. Let $f(t)$ be the probability density function of T , and $F(t)$ the distribution function of T . As always, the following relationships exist between these two quantities,

$$\begin{aligned} f(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{\partial F(t)}{\partial t}, \\ F(t) &= P(T \leq t) = \int_0^t f(u)du. \end{aligned} \quad (1)$$

The *survival probability* or survival function, indicating the probability of nonoccurrence of an event until time t , is defined as

$$S(t) = 1 - F(t) = P(T \geq t) = \int_t^\infty f(u)du. \quad (2)$$

Another important concept is the *hazard rate* or hazard function, $h(t)$, expressing the instantaneous risk of experiencing an event at $T = t$, given that the event did not occur before t . The hazard rate is

defined as

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \quad (3)$$

in which $P(t \leq T < t + \Delta t | T \geq t)$ indicates the probability that the event will occur during $[t \leq T < t + \Delta t]$, given that the event did not occur before t . The hazard rate is equal to the unconditional instantaneous probability of having an event at $T = t$, $f(t)$, divided by the probability of not having an event before $T = t$, $S(t)$. It should be noted that the hazard rate itself cannot be interpreted as a conditional probability. Although its value is always nonnegative, it can take on values larger than one. However, for small Δt , the quantity $h(t)\Delta t$ can be interpreted as the approximate conditional probability that the event will occur between t and $t + \Delta t$.

Above $h(t)$ was defined as a function of $f(t)$ and $S(t)$. It is also possible to express $S(t)$ and $f(t)$ in terms of $h(t)$; that is,

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t h(u)du\right), \\ f(t) &= h(t)S(t) = h(t)\exp\left(-\int_0^t h(u)du\right). \end{aligned} \quad (4)$$

This shows that the functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ give mathematically equivalent specifications of the distribution of T .

Log-linear Models for the Hazard Rate

When working within a continuous-time framework, the most appropriate method for regressing the time variable T on a set of covariates is through the hazard rate. This makes it straightforward to assess the effects of time-varying covariates – including the time dependence itself and time-covariate interactions – and to deal with censored observations. Censoring is a form of missing data that is explained in more detail below.

Let $h(t|\mathbf{x}_i)$ be the hazard rate at $T = t$ for an individual with covariate vector $\mathbf{x}_i$. Since the hazard rate can take on values between 0 and infinity, most hazard models are based on a log transformation of the hazard rate, which yields a regression model of the form

$$\log h(t|\mathbf{x}_i) = \log h(t) + \sum_j \beta_j x_{ij}. \quad (5)$$

This hazard model is not only log-linear but also proportional. In proportional hazard models, the time dependence is multiplicative (additive after taking logs) and independent of an individual's covariate values. The following section shows how to specify nonproportional log-linear hazard models by including time-covariate interactions.

The various types of continuous-time log-linear hazard models are defined by the functional form that is chosen for the time dependence, that is, for the term $\log h(t)$. In Cox's semiparametric model [3], the time dependence is left unspecified. Exponential models assume the hazard rate to be constant over time, while piecewise exponential model assume the hazard rate to be a step function of T , that is, constant within time periods. Other examples of parametric log-linear hazard models are Weibull, Gompertz, and polynomial models.

As was demonstrated by several authors (for example, see [6] or [10]), log-linear hazard models can also be defined as log-linear Poisson models, which are also known as log-rate models. Assume that we have – besides the event history information – two categorical covariates denoted by A and B . In addition, assume that the time axis is divided into a limited number of time intervals in which the hazard rate is postulated to be constant. In the first-birth example, this could be one-year intervals. The discretized time variable is denoted by T . Let h_{abt} denote the constant hazard rate in the t th time interval for an individual with $A = a$ and $B = b$. To see the similarity with standard **log-linear models**, it should be noted that the hazard rate, sometimes referred to as occurrence-exposure rate, can also be defined as $h_{abt} = m_{abt}/E_{abt}$. Here, m_{abt} denoted the expected number of occurrences of the event of interest and E_{abt} the total exposure time in cell (a, b, t) .

Using the notation of hierarchical log-linear models the saturated model for the hazard rate h_{abt} can now be written as

$$\log h_{abt} = u + u_a^A + u_b^B + u_t^T + u_{ab}^{AB} + u_{at}^{AT} + u_{bt}^{BT} + u_{abt}^{ABT}, \quad (6)$$

in which the u terms are log-linear parameters which are constrained in the usual way, for instance, by means of **analysis of variance**-like restrictions. Note that this is a nonproportional model because of the presence of time-covariate interactions. Restricted variants of model described in (6) can be obtained by

omitting some of the higher-order interaction terms. For example,

$$\log h_{abt} = u + u_a^A + u_b^B + u_t^T \quad (7)$$

yields a model that is similar to the proportional log-linear hazard model described in (5). In addition, different types of hazard models can be obtained by the specification of the time dependence. Setting the u_t^T terms equal to zero yields an exponential model. Unrestricted u_t^T parameters yield a piecewise exponential model. Other parametric models can be approximated by defining the u_t^T terms to be some function of T . And finally, if there are as many time intervals as observed survival times and if the time dependence of the hazard rate is not restricted, one obtains a Cox regression model. Log-rate models can be estimated using standard programs for log-linear analysis or Poisson regression using E_{abt} as a weight or exposure vector (see [10] and generalized linear models).

Censoring

An issue that always receives a great amount of attention in discussions on event history analysis is censoring. An observation is called *censored* if it is known that it did not experience the event of interest during some time, but it is not known when it experienced the event. In fact, censoring is a specific type of missing data. In the first-birth example, a censored case could be a woman who is 30 years of age at the time of interview (and has no follow-up interview) and does not have children. For such a woman, it is known that she did not have a child until age 30, but it is not known whether or when she will have her first child. This is, actually, an example of what is called *right censoring*. Another type of censoring that is more difficult to deal with is left censoring. Left censoring means that we do not have information on the duration of nonoccurrence of the event before the start of the observation period.

As long as it can be assumed that the censoring mechanism is not related to the process under study, dealing with right censored observations in **maximum likelihood estimation** of the parameters of hazard models is straightforward. Let δ_i be a censoring indicator taking the value 0 if observation i is censored and 1 if it is not censored. The contribution of case i to the likelihood function that must be

maximized when there are censored observations is

$$\begin{aligned}\mathcal{L}_i &= h(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i) \\ &= h(t_i|\mathbf{x}_i)^{\delta_i} \exp\left(-\int_0^{t_i} h(u|\mathbf{x}_i) du\right).\end{aligned}\quad (8)$$

As can be seen, the likelihood contribution of a censored case equals its survival probability $S(t_i|\mathbf{x}_i)$, and of a noncensored case the density $f(t_i|\mathbf{x}_i)$, which equals $h(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i)$.

Time-varying Covariates

A strong point of hazard models is that one can use time-varying covariates. These are covariates that may change their value over time. Examples of interesting time-varying covariates in the first-birth example are a woman's marital and work status. It should be noted that, in fact, the time variable and interactions between time and time-constant covariates are time-varying covariates as well.

The saturated log-rate model described in (6), contains both time effects and time-covariate interaction terms. Inclusion of ordinary time-varying covariates does not change the structure of this hazard model. The only implication of, for instance, covariate B being time varying rather than time constant is that in the computation of the matrix with exposure times E_{abt} it has to taken into account that individuals can switch from one level of B to another.

Multiple Risks

Thus far, only hazard rate models for situations in which there is only one destination state were considered. In many applications it may, however, prove necessary to distinguish between different types of events or risks. In the analysis of the first-union formation, for instance, it may be relevant to make a distinction between marriage and cohabitation. In the analysis of death rates, one may want to distinguish different causes of death. And in the analysis of the length of employment spells, it may be of interest to make a distinction between the events voluntary job change, involuntary job change, redundancy, and leaving the labor force.

The standard method for dealing with situations where – as a result of the fact that there is more than one possible destination state – individuals may

experience different types of events is the use of a multiple-risk or competing-risk model. A multiple-risk variant of the hazard rate model described in (5) is

$$\log h_d(t|\mathbf{x}_i) = \log h_d(t) + \sum_j \beta_{jd} x_{ij}. \quad (9)$$

Here, the index d indicates the destination state or the type of event. As can be seen, the only thing that changes compared to the single type of event situation is that we have a separate set of time and covariate effects for each type of event.

Repeatable Events and Other Types of Multivariate Event Histories

Most events studied in social sciences are repeatable, and most event history data contains information on repeatable events for each individual. This is in contrast to biomedical research, where the event of greatest interest is death. Examples of repeatable events are job changes, having children, arrests, accidents, promotions, and residential moves.

Often events are not only repeatable but also of different types, that is, we have a multiple-state situation. When people can move through a sequence of states, events cannot only be characterized by their destination state, as in competing risks models, but they may also differ with respect to their origin state. An example is an individual's employment history: an individual can move through the states of employment, unemployment, and out of the labor force. In that case, six different kinds of transitions can be distinguished, which differ with regard to their origin and destination states. Of course, all types of transitions can occur more than once. Other examples are people's union histories with the states living with parents, living alone, unmarried cohabitation, and married cohabitation, or people's residential histories with different regions as states.

Hazard models for analyzing data on repeatable events and multiple-state data are special cases of the general family of multivariate hazard rate models. Another application of these multivariate hazard models is the simultaneous analysis of different life-course events. For instance, it can be of interest to investigate the relationships between women's reproductive, relational, and employment careers, not only by means of the inclusion of time-varying

covariates in the hazard model, but also by explicitly modeling their mutual interdependence.

Another application of multivariate hazard models is the analysis of dependent or clustered observations. Observations are clustered, or dependent, when there are observations from individuals belonging to the same group or when there are several similar observations per individual. Examples are the occupational careers of spouses, educational careers of brothers, child mortality of children in the same family, or in medical experiments, measures of the sense of sight of both eyes or measures of the presence of cancer cells in different parts of the body. In fact, data on repeatable events can also be classified under this type of multivariate event history data, since in that case there is more than one observation of the same type for each observational unit as well.

The hazard rate model can easily be generalized to situations in which there are several origin and destination states and in which there may be more than one event per observational unit. The only thing that changes is that we need indices for the origin state (o), the destination state (d), and the rank number of the event (m). A log-linear hazard rate model for such a situation is

$$\log h_{od}^m(t|\mathbf{x}_i) = \log h_{od}^m(t) + \sum_j \beta_{jod}^m x_{ij}. \quad (10)$$

The different types of multivariate event history data have in common that there are dependencies among the observed survival times. These dependencies may take several forms: the occurrence of one event may influence the occurrence of another event; events may be dependent as a result of common antecedents; and survival times may be correlated because they are the result of the same causal process, with the same antecedents and the same parameters determining the occurrence or nonoccurrence of an event. If these common risk factors are not

observed, the assumption of statistical independence of observation is violated. Hence, unobserved heterogeneity should be taken into account.

Unobserved Heterogeneity

In the context of the analysis of survival and event history data, the problem of unobserved heterogeneity, or the bias caused by not being able to include particular important explanatory variables in the regression model, has received a great deal of attention. This is not surprising because this phenomenon, which is also referred to as selectivity or frailty, may have a much larger impact in hazard models than in other types of regression models:

We will illustrate the effects of unobserved heterogeneity with a small example. Suppose that the population under study consists of two subgroups formed by the two levels of an observed covariate A , where for an average individual with $A = 2$ the hazard rate is twice as high as for someone with $A = 1$. In addition, assume that within each of the levels of A there is (unobserved) heterogeneity in the sense that there are two subgroups within levels of A denoted by $W = 1$ and $W = 2$, where $W = 2$ has a 5 times higher hazard rate than $W = 1$. Table 1 shows the assumed hazard rates for each of the possible combinations of A and W at four time points. As can be seen, the true hazard rates are constant over time within levels of A and W . The reported hazard rates in the columns labeled 'observed' show what happens if we cannot observe W . First, it can be seen that despite that the true rates are time constant, both for $A = 1$ and $A = 2$ the observed hazard rates decline over time. This is an illustration of the fact that unobserved heterogeneity biases the estimated time dependence in a negative direction. Second, while the ratio between the hazard rates for $A = 2$ and $A = 1$ equals the true value 2.00 at $t = 0$, the

Table 1 Hazard rates illustrating the effect of unobserved heterogeneity

Time point	$A = 1$			$A = 2$			Ratio between $A = 2$ and $A = 1$
	$W = 1$	$W = 2$	Observed	$W = 1$	$W = 2$	Observed	
0	.010	.050	.030	.020	.100	.060	2.00
10	.010	.050	.026	.020	.100	.045	1.73
20	.010	.050	.023	.020	.100	.034	1.50
30	.010	.050	.019	.020	.100	.027	1.39

observed ratio declines over time (see last column). Thus, when estimating a hazard model with these observed hazard rates, we will find a smaller effect of A than the true value of $(\log) 2.00$. Third, in order to fully describe the pattern of observed rates, we need to include a time-covariate interaction in the hazard model: the covariate effect changes (declines) over time or, equivalently, the (negative) time effect is smaller for $A = 1$ than for $A = 2$.

Unobserved heterogeneity may have different types of consequences in hazard modeling. The best-known phenomenon is the downwards bias of the duration dependence. In addition, it may bias covariate effects, time-covariate interactions, and effects of time-varying covariates. Other possible consequences are dependent censoring, dependent competing risks, and dependent observations. The common way to deal with unobserved heterogeneity is included random effects in the model of interest (for example, see [4] and [9]).

The random-effects approach is based on the introduction of a time-constant latent covariate in the hazard model. The latent variable is assumed to have a multiplicative and proportional effect on the hazard rate, that is,

$$\log h(t|\mathbf{x}_i, \theta_i) = \log h(t) + \sum_j \beta_j x_{ij} + \log \theta_i \quad (11)$$

Here, θ_i denotes the value of the **latent variable** for subject i . In the parametric random-effects approach, the latent variable is postulated to have a particular distributional form. The amount of unobserved heterogeneity is determined by the size of the standard deviation of this distribution: The larger the standard deviation of θ , the more unobserved heterogeneity there is.

Heckman and Singer [4] showed that the results obtained from a random-effects continuous-time hazard model can be sensitive to the choice of the functional form of the mixture distribution. They, therefore, proposed using a nonparametric characterization of the mixing distribution by means of a finite set of so-called mass points, or latent classes, whose number, locations, and weights are empirically determined (also, see [10]). This approach is implemented in the Latent GOLD software [11] for **latent class analysis**.

Example: First Interfirm Job Change

To illustrate the use of hazard models, we use a data set from the 1975 Social Stratification and Mobility Survey in Japan reported in Yamaguchi's [12] textbook on event history analysis. The event of interest is the first interfirm job separation experienced by the sample subjects. The time variable is measured in years. In the analysis, the last one-year time intervals are grouped together in the same way as Yamaguchi did, which results in 19 time intervals. It should be noted that contrary to Yamaguchi, we do not apply a special formula for the computation of the exposure times for the first time interval.

Besides the time variable denoted by T , there is information on the firm size (F). The first five categories range from small firm (1) to large firm (5). Level 6 indicates government employees. The most general log-rate model that will be used is of the form

$$\log h_{ft} = u + u_f^F + u_t^T. \quad (12)$$

The log-likelihood values, the number of parameters, as well as the BIC¹ values for the estimated models are reported in Table 2. Model 1 postulates that the hazard rate does neither depend on time or firm size and Model 2 is an exponential survival model with firm size as a nominal predictor. The large difference in the log-likelihood values of these two models shows that the effect of firm size on the rate of job change is significant. A Cox proportional hazard model is obtained by adding an unrestricted time effect (Model 3). This model performs much better than Model 2, which indicates that there is a strong time dependence. Inspection of the estimated time dependence of Model 3 shows that the hazard rate rises in the first time periods and subsequently starts decreasing slowly (see Figure 1). Models 4 and 5 were estimated to test whether it is possible to simplify the time dependence of the hazard rate on the basis of this information. Model 4 contains only time parameters for the first and second time point, which

Table 2 Test results for the job change example

Model	Log-likelihood	# parameters	BIC
1. $\{\}$	-3284	1	6576
2. $\{F\}$	-3205	6	6456
3. $\{Z, F\}$	-3024	24	6249
4. $\{Z_1, Z_2, F\}$	-3205	8	6471
5. $\{Z_1, Z_2, Z_{lin}, F\}$	-3053	9	6174

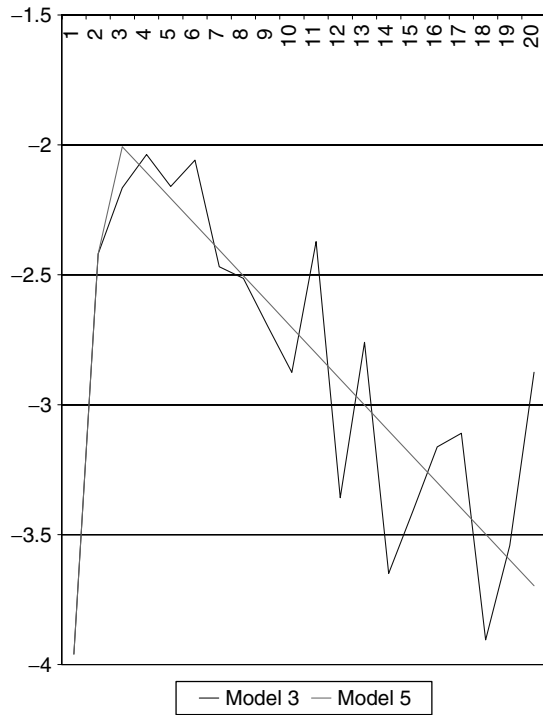


Figure 1 Time dependence according to model 3 and model 5

means that the hazard rate is assumed to be constant from time point 3 to 19. Model 5 is the same as Model 4 except for that it contains a linear term to describe the negative time dependence after the second time point. The comparison between Models 4 and 5 shows that this linear time dependence of the log hazard rate is extremely important: The log-likelihood increases 97 points using only one additional parameter. Comparison of Model 5 with the less restricted Model 3 and the more restricted Model 2 shows that Model 5 captures the most important part of the time dependence. Though according to the likelihood-ratio statistic the difference between Models 3 and 5 is significant, Model 5 is the preferred model according to the BIC criterion. Figure 1 shows how Model 5 smooths the time dependence compared to Model 3.

The log-linear hazard parameter estimates for firm size obtained with Model 5 are 0.51, 0.28, 0.03, -0.01, -0.48, and -0.34, respectively.² These show that there is a strong effect of firm size on the rate of a

first job change: The larger the firm the less likely an employee is to leave the firm or, in other words, the longer he will stay. Government employees (category 6) have a slightly higher (less low) hazard rate than employees of large firm (category 5).

Notes

1. BIC is defined as minus twice the log-likelihood plus $\ln(N)$ times the number of parameters, where N is the sample size (here 1782).
2. Very similar estimates are obtained with Model 3.

References

- [1] Allison, P.D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*, Sage Publications, Beverly Hills.
- [2] Blossfeld, H.P. & Rohwer, G. (1995). *Techniques of Event History Modeling*, Lawrence Erlbaum Associates, Mahwah.
- [3] Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society, Series B* **34**, 187–203.
- [4] Heckman, J.J. & Singer, B. (1982). Population heterogeneity in demographic models, in *Multidimensional Mathematical Demography*, K. Land & A. Rogers, eds, Academic Press, New York.
- [5] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.
- [6] Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.
- [7] Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.
- [8] Tuma, N.B. & Hannan, M.T. (1984). *Social Dynamics: Models and Methods*, Academic Press, New York.
- [9] Vaupel, J.W., Manton, K.G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**, 439–454.
- [10] Vermunt, J.K. (1997). Log-linear models for event history histories, *Advanced Quantitative Techniques in the Social Sciences Series*, Vol. 8, Sage Publications, Thousand Oakes.
- [11] Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD 2.0 User's Guide*, Statistical Innovations Inc., Belmont.
- [12] Yamaguchi, K. (1991). Event history analysis, *Applied Social Research Methods*, Vol. 28, Sage Publications, Newbury Park.

JEROEN K. VERMUNT AND GUY MOORS

Exact Methods for Categorical Data

SCOTT L. HERSHBERGER

Volume 2, pp. 575–580

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Exact Methods for Categorical Data

Introduction

The validity of methods based on *asymptotic theory* is doubtful when sample size is small, as well as when data are sparse, skewed, or heavily tied. One way to form valid statistical inferences under these adverse data conditions is to compute exact P values, based on the distribution of *all* possible values of a test statistic that could be obtained from a given set of data. *Exact tests* are procedures for determining statistical significance using the **sampling distribution** of the test statistic obtained from the observed data: instead of evaluating the data in reference to some underlying theoretical population distribution, the data itself are employed to construct the relevant sampling distribution. Probability values are considered ‘exact’ in exact tests, since the P values are obtained from a sampling distribution composed of all possible values of test statistic computable from the data.

Exact P values can be obtained for a large number of nonparametric (*see* **Distribution-free Inference, an Overview**) statistical tests, including simple linear rank statistics (*see* **Rank Based Inference**) based on Wilcoxon scores, median scores, Van der Waerden scores, and Savage scores. Exact P values can also be obtained for univariate (e.g., the independent means t Test) and multivariate (e.g., Hotelling’s T^2) normal statistical tests.

Creating the ‘Exact’ Sampling Distribution of a Test Statistic

The most important questions to be addressed are (a) how the sampling distribution of a test statistic is obtained from an observed data set, that is, how is each sample constructed, and (b) how many samples are required. Whatever the answer to (a), the number of samples drawn is exhaustive in exact tests. By ‘exhaustive’, we mean all samples that could possibly be constructed in a specified way are selected.

As an example, consider how an exact test of a single median is performed. In this situation, we are required to assign positive and negative signs in all

possible combinations to the data. Let us say our data are composed of four values: 2, 4, 6, and 8. Since there are four numbers, each of which could be positive or negative, there are $2^N = 2^4 = 16$ possible combinations:

- | | |
|-------------------|-------------------|
| 1. 2, 4, 6, 8 | 2. -2, 4, 6, 8 |
| 3. -2, -4, 6, 8 | 4. -2, -4, -6, 8 |
| 5. -2, -4, -6, -8 | 6. 2, -4, 6, 8 |
| 7. 2, -4, -6, 8 | 8. 2, -4, -6, -8 |
| 9. 2, -4, -6, 8 | 10. 2, 4, -6, -8 |
| 11. -2, 4, -6, 8 | 12. -2, 4, -6, -8 |
| 13. 2, 4, 6, -8 | 14. 2, 4, -6, -8 |
| 15. -2, 4, 6, -8 | 16. 2, -4, 6, -8 |

In this situation, the answer to (a) is, for each of the samples, to draw all the numbers from the data and assign each number a positive or negative sign. The answer to (a) also determines the answer to (b): according to the rule for the number of possible sequences of N observations, each of which may have K outcomes, K^N , there must be 16 unique sequences (samples) that could be drawn. From each of these 16 samples, a test statistic would be calculated, creating a sampling distribution of the statistic with 16 observations. (As explained below, the test statistic is the sum of the observations in a sample.)

In our next example, the number of unique ways that N observations can be distributed among K groups determines how many samples are to be drawn and is equal to

$$\frac{N!}{n_1!n_2!\dots n_K!}.$$

As an illustration, we wish to conduct a one-way **analysis of variance** (ANOVA) with three groups. Assume the following data were available:

- Group 1: 6, 8
 Group 2: 9, 11, 9
 Group 3: 17, 15, 16, 16

Therefore, group 1 has two observations ($n_1 = 2$), group 2 has three ($n_2 = 3$), and group 3 has four ($n_3 = 4$), for a total $N = 9$. The number of ways two, three, and four observations can be distributed across three groups is $9!/(2)!(3)!(4)! = 1260$.

Resampling Procedures

Differences among Permutation Tests, Randomization Tests, Exact Tests, and the Bootstrap

One should also note that a great deal of confusion exists for the terms *exact test*, permutation test (see **Permutation Based Inference**), and randomization test (see **Randomization Based Tests**) [7, 9, 13]. How do they differ? In general, no meaningful statistical differences exist among these tests; the few differences that exist are minor, and in most situations produce the same outcome. ‘Permutation test’ is the broadest and most commonly used term of three, generally referring to procedures that repeatedly sample, using various criteria depending on the method, by permutation, using some criterion. Exact tests are permutation tests. Randomization tests are specific types of permutation tests and are applied to the permutation of data from experiments in which subjects have been randomly assigned to treatment groups. Although the term ‘exact test’ is occasionally used instead of the more generic ‘permutation test’, to our knowledge, the only statistical procedure labeled as ‘exact’ is Fisher’s exact test. In the interest of semantic clarity, Fisher’s exact test should be labeled ‘Fisher’s permutation test’.

Permutation, randomization, and exact tests are types of *resampling procedures*, procedures that select samples of scores from the original data set. Good [10] provides a particularly comprehensive review of resampling procedures. A critical difference *does* exist between the three permutation-type tests and the **bootstrap**, a method of resampling to produce random samples of size n from an original data set of size N . The critical difference is that permutation tests are based on sampling without replacement while the bootstrap is based on sampling with replacement. Although both permutation-type tests and the bootstrap often produce similar results, the former appears to be more accurate for smaller samples [18].

Exact Test Examples

We illustrate exact tests with two examples.

The Pitman Test

The basic idea behind the **Pitman test** [15, 17] is that if a random sample is from a symmetric distribution

with an unknown median θ , then symmetry implies it is equally likely any sample value will differ from θ by some positive amount d , or by the same negative amount, $-d$, for all values of d . In the Pitman test, the null hypothesis could be that the population median θ is equal to a specific value θ_0 ($H_0: \theta = \theta_0$), whereas the two-sided alternative hypothesis is $H_1: \theta \neq \theta_0$. It is assumed that under H_0 , the sign of each of the N differences

$$d_i = x_i - \theta_0, \quad i = 1, 2, \dots, N \quad (1)$$

is equally likely to be positive or negative. Clearly, when H_0 is not correct, there is more likely to be preponderance either of positive or negative signs associated with the d_i .

Example

The final exam scores for nine students were 50, 80, 77, 95, 88, 62, 90, 91, and 74, with a median (θ) of 80. In the past, the median final exam score (θ_0) was 68. Thus, the teacher suspects that this class performed better than previous classes. In order to confirm this, the teacher tests the hypothesis that $H_0: \theta = 68$ against the alternative $H_1: \theta > 68$ using the Pitman test.

The only assumption required for the Pitman test is that under H_0 , deviations (d_i) from 68 should be symmetric. Consequently, if H_0 is correct, it follows that the sums of the positive and negative deviations from 68 should not differ. On the other hand, if H_1 is correct, the sum of the positive deviations should exceed the sum of the negative deviations; that is, the class as a whole performed better than a median score of 68 would suggest implying that this class did perform better than classes of previous years.

The first step in the Pitman test is to sum (S) the deviations actually observed in the data: the deviations are $d_1 = 50 - 68 = -18$, $d_2 = 80 - 68 = 12$, $d_3 = 77 - 68 = 9$, $\dots$, $d_9 = 74 - 68 = 6$, their sum $S = 95$. The next step results in a sampling distribution of the sum of deviations. Recall that under H_0 , there is an equal probability that any one of the deviations be positive or negative. The critical assumption in the Pitman test follows from the equal probability of positive and negative deviations: *Any combination of plus and minus signs attached to the nine deviations in the sample is equally likely.* Since there are nine deviations, and each of the

nine deviations can take on either one of two signs (+ or -), there are $2^9 = 512$ possible allocations of signs (two to each of nine differences). The idea behind the Pitman test is allocate signs to the deviations for all possible 512 allocations, and obtain the sum of the deviations for each allocation. The sampling distribution of the sum of the deviations under the null hypothesis is created from the resulting 512 sums. Using this distribution, we can determine what is the probability of obtaining a sum as great as or greater than the sum observed (i.e., $S = 95$) in the sample if the null hypothesis is correct. Our interest in the probability of a sum greater than 95 follows from the teacher's specification of a one-tailed H_1 : The teacher expects the current class's performance to be significantly better than in previous years; or to put it another way, the new median of 80 should be significantly greater than 65. If this probability is low enough, we reject the null hypothesis that the current class's median is 68.

For example, one of the 512 allocations of signs to the deviations gives each deviation a positive sign; that is, 35, 5, 8, 10, 3, 23, 5, 6, and 11. The sum of these deviations is 101. Conversely, another one of the 512 allocations gives each of the deviations a negative sign, resulting in a sum of -101. All 512 sums contribute to the sampling distribution of S . On the basis of this sampling distribution, the sum of 95 has a probability of .026 or less of occurring if the null hypothesis is correct. Given the low probability that H_0 is correct ($p < .026$, we reject H_0 and decide that the current class median of $80(\theta)$ is significantly greater the past median of $65(\theta_0)$. Maritz [13] and Sprent [17] provide detailed discussions of the Pitman test.

Fisher's Exact Test

Fisher's exact test provides an exact method for testing the null hypothesis of independence for categorical data in a 2×2 contingency table with both sets of marginal frequencies fixed in advance. The exact probability is calculated for a sample showing as much or more evidence for independence than that obtained. As with many exact tests, a number of samples are obtained from the data; in this case, all possible contingency tables having the

specified marginal frequencies. An empirical probability distribution is constructed that reflects the probability of observing each of the contingency tables. This test was first proposed in [8], [11], and [20], and is also known as the Fisher-Irwin test and as the Fisher-Yates test. It is discussed in many sources, including [2], [5], [6], [16], and [19].

Consider the following 2×2 contingency table.

$$\begin{array}{cc|c} & B_1 & B_2 & Totals \\ A_1 & a & b & a+b \\ A_2 & c & d & c+d \\ \hline & a+c & b+d & N \end{array}$$

The null hypothesis is that the categories of A and B are independent. Under this null hypothesis, the probability p of observing any particular table with all marginal frequencies fixed follows the hypergeometric distribution (see **Catalogue of Probability Density Functions**):

$$p = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{n!a!b!c!d!}. \quad (2)$$

This equation expresses the distribution of the four cell counts in terms of only one cell (it does not matter which). Since the marginal totals (i.e., $a+c$, $b+d$, $a+b$, $c+d$) are fixed, once the number of observations in one cell has been specified, the number of observations in the other three cells are not free to vary: the count for one cell determines the other three cell counts.

In order to test the null hypothesis of independence, a one-tailed P value can be evaluated as the probability of obtaining a result (a 2×2 contingency table with a particular distribution of observations) as extreme as the observed value in the one cell that is free to vary in one direction. That is, the probabilities obtained from the hypergeometric distribution in one tail are summed. Although in the following example the alternative hypothesis is directional, one can also perform a nondirectional Fisher exact test.

To illustrate Fisher's exact test, we analyze the results of an experiment examining the ability of a subject to discriminate correctly between two objects. The subject is told in advance exactly how many times each object will be presented and is required to make that number of identifications. This

4 Exact Methods for Categorical Data

is to ensure that the marginal frequencies remain fixed.

The results are given in the following contingency table.

	B_1	B_2	$Totals$
A_1	1	6	7
A_2	6	1	7
	7	7	14

Factor A represents the presentation of the two objects and factor B is the subject's identification of these. The null hypothesis is that the presentation of an object is independent of its identification; that is, the subject cannot correctly discriminate between the two objects.

Using the equation for the probability from a hypergeometric distribution, we obtain the probability of observing this table with its specific distribution of observations:

$$p = \frac{7!7!7!7!}{14!1!6!1!6!} = 0.014. \quad (3)$$

However, in order to evaluate the null hypothesis, in addition to the probability $p = 0.014$, we must also compute the probabilities for any sets of observed frequencies that are even more extreme than the observed frequencies. The only result that is more extreme is

	B_1	B_2	$Totals$
A_1	0	7	7
A_2	7	0	7
	7	7	14

The probability of observing this contingency table is

$$p = \frac{7!7!7!7!}{14!0!7!0!7!} = 0.0003. \quad (4)$$

When $p = 0.014$ and $p = 0.0003$ are added, the resulting probability of 0.0143 is the likelihood of obtaining a set of observed frequencies that is equal to or is more extreme than the set of observed frequencies by chance alone. If we use a one-tailed alpha of 0.05 as the criterion for rejecting the null hypothesis, the probability of 0.0143 suggests that the likelihood that the experimental results would occur by chance alone is too small. We therefore reject the null hypothesis: there is a relation between the presentation of the objects and the subject's correct identification of them –

the subject is able to discriminate between the two objects.

Prior to the widespread availability of computers, Fisher's exact was rarely performed for sample sizes larger than the one in our example. The reason for this neglect is attributable to the intimidating number of contingency tables that could possibly be observed with the marginal totals fixed to specific values and the necessity of computing a hypergeometric probability for each. If we arbitrarily choose cell a as the one cell of the four free to vary, the number of possible tables is $m_{\text{Low}} \leq a \leq m_{\text{High}}$, where $m_{\text{Low}} = \max(0, a + c + b + d - N)$ and $m_{\text{High}} = \min(a + c + 1, b + d + 1)$. Applied to our example, there $m_{\text{Low}} = \max(0, 0)$ and $m_{\text{High}} = \min(8, 8)$. Given $a = 1$, there is a range of $0 \leq 1 \leq 8$ possible contingency tables, each with a different distribution of observations. When the most 'extreme' distribution of observations results from an experiment, only one hypergeometric probability must be considered; as the results depart from 'extreme', additional hypergeometric probabilities must be calculated. All eight tables are given in Tables 1–8:

Table 1

	B_1	B_2	$Totals$
A_1	0	7	7
A_2	7	0	7
	7	7	14

$p = 0.0003$

Table 2

	B_1	B_2	$Totals$
A_1	1	6	7
A_2	6	1	7
	7	7	14

$p = 0.014$

Table 3

	B_1	B_2	$Totals$
A_1	2	5	7
A_2	5	2	7
	7	7	14

$p = 0.129$

Table 4

	B_1	B_2	$Totals$
A_1	3	4	7
A_2	4	3	7
	7	7	14

$p = 0.357$

Table 5

	B_1	B_2	$Totals$
A_1	4	4	7
A_2	3	3	7
	7	7	14

$p = 0.357$

Table 6

	B_1	B_2	$Totals$
A_1	5	2	7
A_2	2	5	7
	7	7	14

$p = 0.129$

Table 7

	B_1	B_2	$Totals$
A_1	6	1	7
A_2	1	6	7
	7	7	14

$p = 0.014$

Table 8

	B_1	B_2	$Totals$
A_1	7	0	7
A_2	0	7	7
	7	7	14

$p = 0.0003$

Note that the sum of the probabilities of the eight tables is ≈ 1.00 .

Several algorithms are available for computation of exact P values. A good review of these algorithms is given by Agresti [1]. Algorithms based on direct enumeration are very time consuming and feasible for only smaller data sets. The network algorithm,

authored by Mehta and Patel [14], is a popular and efficient alternative direct enumeration.

Fisher's exact test may also be applied to any $K \times J \times L$ or higher contingency table; in the specific case of $2 \times 2 \times K$ contingency tables, the test is more commonly referred to as the Cochran–Mantel–Haenszel test (see **Mantel–Haenszel Methods**) [3, 12]. Exact tests have also been developed for **contingency tables** having factors with ordered categories [4].

References

- [1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**, 131–177.
- [2] Boik, R.J. (1987). The Fisher–Pitman permutation test: a nonrobust alternative to the normal theory F test when variances are heterogeneous, *British Journal of Mathematical and Statistical Psychology* **40**, 26–42.
- [3] Cochran, W.G. (1954). Some methods of strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.
- [4] Cohen, A., Madigan, D. & Sacrowitz, H.B. (2003). Effective directed tests for models with ordered categorical data, *Australian and New Zealand Journal of Statistics* **45**, 285–300.
- [5] Conover, W.J. (1998). *Practical Nonparametric Statistics*, 3rd Edition, John Wiley & Sons, New York.
- [6] Daniel, W.W. (2001). *Applied Nonparametric Statistics*, 2nd Edition, Duxbury Press, Pacific Grove.
- [7] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [8] Fisher, R.A. (1934). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [9] Good, P.I. (2000). *Permutation Methods: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd Edition, Springer-Verlag, New York.
- [10] Good, P.I. (2001). *Resampling Methods*, 2nd Edition. Birkhauser, Boston.
- [11] Irwin, J.O. (1935). Tests of significance for differences between percentages based on small numbers, *Metron* **12**, 83–94.
- [12] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [13] Maritz, J.S. (1995). *Distribution-free Statistical Methods*, 2nd Edition, Chapman & Hall, London.
- [14] Mehta, C.R. & Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in unordered $r \times c$ contingency tables, *Journal of the American Statistical Association* **78**, 427–434.
- [15] Pitman, E.J.G. (1937). Significance tests that may be applied to samples from any population, *Journal of the Royal Statistical Society Supplement* **4**, 119–130.
- [16] Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer-Verlag, New York.

6 Exact Methods for Categorical Data

- [17] Sprent, P. (1998). *Data Driven Statistical Methods*, Chapman & Hall, London.
- [18] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Edition, Chapman & Hall, Boca Raton.
- [19] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum, Hillsdale.
- [20] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society Supplement* **1**, 217–245.

SCOTT L. HERSHBERGER

Expectancy Effect by Experimenters

ROBERT ROSENTHAL

Volume 2, pp. 581–582

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Expectancy Effect by Experimenters

Some expectation of how the research will turn out is virtually a constant in science. Social scientists, like other scientists generally, conduct research specifically to examine hypotheses or expectations about the nature of things. In the social and behavioral sciences, the hypothesis held by the investigators can lead them unintentionally to alter their behavior toward the research participants in such a way as to increase the likelihood that participants will respond so as to confirm the investigator's hypothesis or expectations. We are speaking, then, of the investigator's hypothesis as a self-fulfilling prophecy. One prophesies an event, and the expectation of the event then changes the behavior of the prophet in such a way as to make the prophesied event more likely. The history of science documents the occurrence of this phenomenon with the case of Clever Hans as a prime example [3].

The first experiments designed specifically to investigate the effects of experimenters' expectations on the results of their research employed human research participants. Graduate students and advanced undergraduates in the field of Psychology were employed to collect data from introductory psychology students. The experimenters showed a series of photographs of faces to research participants and asked participants to rate the degree of success or failure reflected in the photographs. Half the experimenters, chosen at random, were led to expect that their research participants would rate the photos as being of more successful people. The remaining half of the experimenters were given the opposite expectation – that their research participants would rate the photos as being of less successful people. Despite the fact that all experimenters were instructed to conduct a perfectly standard experiment, reading only the same printed instructions to all their participants, those experimenters who had been led to expect ratings of faces as being of more successful people obtained such ratings from their randomly assigned participants. Those experimenters who had been led to expect results in the opposite direction tended to obtain results in the opposite direction.

These results were replicated dozens of times employing other human research participants [7]. They were also replicated employing animal research

subjects. In the first of these experiments, experimenters who were employed were told that their laboratory was collaborating with another laboratory that had been developing genetic strains of maze-bright and maze-dull rats. The task was explained as simply observing and recording the maze-learning performance of the maze-bright and maze-dull rats. Half the experimenters were told that they had been assigned rats that were maze-bright while the remaining experimenters were told that they had been assigned rats that were maze-dull. None of the rats had really been bred for maze-brightness or maze-dullness, and experimenters were told purely at random what type of rats they had been assigned. Despite the

Table 1 Strategies for the control of experimenter expectancy effects

1. Increasing the number of experimenters:
 - decreases learning of influence techniques,
 - helps to maintain blindness,
 - minimizes effects of early data returns,
 - increases generality of results,
 - randomizes expectancies,
 - permits the method of collaborative disagreement, and
 - permits statistical correction of expectancy effects.
2. Observing the behavior of experimenters:
 - sometimes reduces expectancy effects,
 - permits correction for unprogrammed behavior, and
 - facilitates greater standardization of experimenter behavior.
3. Analyzing experiments for order effects:
 - permits inference about changes in experimenter behavior.
4. Analyzing experiments for computational errors:
 - permits inference about expectancy effects.
5. Developing selection procedures:
 - permits prediction of expectancy effects.
6. Developing training procedures:
 - permits prediction of expectancy effects.
7. Developing a new profession of psychological experimenter:
 - maximizes applicability of controls for expectancy effects, and
 - reduces motivational bases for expectancy effects.
8. Maintaining blind contact:
 - minimizes expectancy effects (see Table 2).
9. Minimizing experimenter-participant contact:
 - minimizes expectancy effects (see Table 2).
10. Employing expectancy control groups:
 - permits assessment of expectancy effects.

2 Expectancy Effect by Experimenters

fact that the only differences between the allegedly bright and dull rats were in the mind of the experimenter, those who believed their rats were brighter obtained brighter performance from their rats than did the experimenters who believed their rats were duller. Essentially the same results were obtained in a replication of this experiment employing Skinner boxes instead of mazes. Tables 1 and 2 give a brief overview of some of the procedures that have been proposed to help control the methodological problems created by experimenter expectancy effects [4].

The research literature of the experimenter expectancy effect falls at the intersection of two distinct domains of research. One of these domains is the domain of artifacts in behavioral research [4, 6, 8], including such experimenter-based artifacts

as observer error, interpreter error, intentional error, effects of biosocial and psychosocial attributes, and modeling effects; and such participant-based artifacts as the perceived **demand characteristics** of the experimental situation, **Hawthorne effects** and volunteer bias.

The other domain into which the experimenter expectancy effect simultaneously falls is the more substantive domain of interpersonal expectation effects. This domain includes the more general, social psychology of the interpersonal self-fulfilling prophecy. Examples of major subliterations of this domain include the work on managerial expectation effects in business and military contexts [2] and the effects of teachers' expectations on the intellectual performance of their students [1, 5].

Table 2 Blind and minimized contact as controls for expectancy effects

Blind contact	
A.	Sources of breakdown of blindness
1.	Principal investigator
2.	Participant ('side effects')
B.	Procedures facilitating maintenance of blindness
1.	The 'total-blind' procedure
2.	Avoiding feedback from the principal investigator
3.	Avoiding feedback from the participant
Minimized contact	
A.	Automated data collection systems
1.	Written instructions
2.	Tape-recorded instructions
3.	Filmed instructions
4.	Televised instructions
5.	Telephoned instructions
6.	Computer-based instructions
B.	Restricting unintended cues to participants and experimenters
1.	Interposing screen between participants and experimenters
2.	Contacting fewer participants per experimenter
3.	Having participants or computers record responses

References

- [1] Blanck, P.D., ed. (1994). *Interpersonal Expectations: Theory, Research, and Applications*, Cambridge University Press, New York.
- [2] Eden, D. (1990). *Pygmalion in Management: Productivity as a Self-fulfilling Prophecy*, D. C. Heath, Lexington.
- [3] Pfungst, O. (1965). *Clever Hans*, Translated by Rahn, C.L., Holt, New York, 1911; Holt, Rinehart and Winston.
- [4] Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York, (Enlarged edition, 1976; Irvington Publishers).
- [5] Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the Classroom*, Holt, Rinehart and Winston, New York.
- [6] Rosenthal, R. & Rosnow, R.L., eds (1969). *Artifact in Behavioral Research*, Academic Press, New York.
- [7] Rosenthal, R. & Rubin, D.B. (1978). Interpersonal expectancy effects: the first 345 studies, *Behavioral and Brain Sciences* 3, 377-386.
- [8] Rosnow, R.L. & Rosenthal, R. (1997). *People Studying People: Artifacts and Ethics in Behavioral Research*, W. H. Freeman, New York.

ROBERT ROSENTHAL

Expectation

REBECCA WALWYN

Volume 2, pp. 582–584

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Expectation

The expectation or expected value of a random variable is also referred to as its mean or ‘average value.’ The terms *expectation*, *expected value*, and *mean* can be used interchangeably. The expectation can be thought of as a measure of the ‘center’ or ‘location’ of the probability distribution of the random variable. Two other measures of the center or location are the median and the mode. If the random variable is denoted by the letter X , the expectation of X is usually denoted by $E(X)$, said ‘E of X .’

The form of the definition of the expectation of a random variable is slightly different depending on the nature of the probability distribution of the random variable (see **Catalogue of Probability Density Functions**). The expectation of a discrete random variable is defined as the sum of each value of the random variable weighted by the probability that the random variable is equal to that value. Thus,

$$E(X) = \sum_x x f(x), \quad (1)$$

where $f(x)$ denotes the probability distribution of the discrete random variable, X . So, for example, suppose that the random variable X takes on five discrete values, -2 , -1 , 0 , 1 , and 2 . It could be an item on a questionnaire, for instance. Also suppose that the probability that the random variable takes on the value -2 (i.e., $(X = -2)$) is equal to 0.1 , and that $P(X = -1) = 0.2$, $P(X = 0) = 0.3$, $P(X = 1) = 0.25$, and $P(X = 2) = 0.15$. Then,

$$E(X) = (-2 * 0.1) + (-1 * 0.2) + (0 * 0.3) + (1 * 0.25) + (2 * 0.15) = 0.15. \quad (2)$$

Note that the individual probabilities sum to 1. The expectation is meant to summarize a ‘typical’ observation of the random variable but, as can be seen from the example above, the expectation of a random variable is not necessarily equal to one of the values of that random variable. It can also be very sensitive to small changes in the probability assigned to a large value of X .

It is possible that the expectation of a random variable does not exist. In the case of a discrete random variable, this occurs when the sum defining the expectation does not converge or is not equal to infinity.

Consider the *St. Petersburg paradox* discovered by the Swiss eighteenth-century mathematician Daniel Bernoulli (see **Bernoulli Family**) [1]. A fair coin is tossed repeatedly until it shows tails. The total number of tosses determines the prize, which equals $\$2^n$, where n is the number of tosses. The expected value of the prize for each toss is $\$1$ ($2 * 0.5 + 0 * 0.5$). The expected value of the prize for the game is the sum of the expected values for each toss. As there are an infinite number of tosses, the expected value of the prize for the game is an infinite number of dollars.

The expectation of a continuous random variable is defined as the integral of the individual values of the random variable weighted according to their probability distribution. Thus,

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \quad (3)$$

where $f(x)$ is the probability distribution of the continuous random variable, X . This achieves the same end for a continuous random variable as (1) did for a discrete random variable; it sums each value of the random variable weighted by the probability that the random variable is equal to that value.

Again it is possible that the expectation of the continuous random variable does not exist. This occurs when the integral does not converge or equals infinity. A classic example of a random variable whose expected value does not exist is a Cauchy random variable.

An expectation can also be calculated for the function of a random variable. This is done by applying (1) or (3) to the distribution of a function of the random variable. For example, a function of the random variable X is $(X - \mu_X)^2$, where μ_X denotes the mean of the random variable X . The expectation of this function is referred to as the variance of X .

Expectations have a number of useful properties.

1. The expectation of a constant is equal to that constant.
2. The expectation of a constant multiplied by a random variable is equal to the constant multiplied by the expectation of the random variable.
3. The expectation of the sum of a group of random variables is equal to the sum of their individual expectations.
4. The expectation of the product of a group of random variables is equal to the product of their

2 Expectation

individual expectations if the random variables are independent.

More information on the topic of expectation is given in [2], [3] and [4].

References

- [1] Bernoulli, D. (1738). Exposition of a new theory on the measurement of risk, *Econometrica* **22**, 23–36.

- [2] Casella, G. & Berger, R.L. (1990). *Statistical Inference*, Duxbury, California.
- [3] DeGroot, M.H. (1986). *Probability and Statistics*, 2nd Edition, Addison-Wesley, Massachusetts.
- [4] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, McGraw-Hill, Singapore.

REBECCA WALWYN

Experimental Design

HAROLD D. DELANEY

Volume 2, pp. 584–586

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Experimental Design

Experiments are, as **Sir Ronald Fisher** remarked, ‘only experience carefully planned in advance and designed to form a secure basis of new knowledge’ [3, p. 8]. The goal of experimental design is to allow inferences to be drawn that are logically compelled by the data about the effects of treatments. Efficiency and generalizability are also important concerns, but are logically secondary to the interpretability of the experiment.

The primary goal of experimental design is assessing causal relationships. This is a different focus from those methodologies, central to epidemiology or sociology, that attempt to determine the prevalence of a phenomenon in a population. When one wants to determine the prevalence of a disease or a political opinion, for example, the primary concern must be with the representativeness of the sample, and hence *random sampling* and the accompanying statistical theory regarding various sampling procedures are critical concerns (see **Randomization**) [1]. In experimental design, however, the methodological feature regarded, at least since the time of Fisher, as most critical is *random assignment* to conditions. Random assignment assures that no uncontrolled factor will, in the long run, bias the comparison of conditions, and hence provides the secure basis for causal inferences that Fisher was seeking. The use of random assignment also provides the justification for using mathematical models involving random processes and for making probabilistic statements that guide the interpretation of the experiment, as seen most explicitly in what are called **randomization tests** or **permutation tests** [2, 4].

When random assignment is not feasible, one of a variety of **quasi-experimental** or nonrandomized designs can be employed (see **observational studies**) [6], though threats to the validity of inferences abound in such studies. For example, in a design with nonequivalent groups, any differences on a posttest may be due to preexisting differences or selection effects rather than the treatment. Biases induced by nonrandom assignment may be exacerbated when participants self-select into conditions, whereas group similarity can sometimes be increased by matching. While attempts to control for **confounding variables** via matching or **analysis of covariance** are of some value, one cannot be assured that all the relevant

differences will be taken into account or that those that one has attempted to adjust for have been perfectly measured [5].

In the behavioral sciences, dependent variables usually are reasonably continuous. Experimenters typically attempt to account for or predict the variability of the dependent variables by factors that they have manipulated, such as group assignment, and by factors that they have measured, such as a participant’s level of depression prior to the study. Manipulated factors are almost always discrete variables, whereas those that are measured, although sometimes discrete, are more often continuous. To account for variability in the dependent variable, the most important factors to include in a statistical model are usually continuous measures of preexisting individual differences among participants (see **Nuisance Variables**). However, such ‘effects’ may not be causal, whereas the effects of manipulated factors, though perhaps smaller, will have a clearer theoretical interpretation and practical application. This is especially true when the participants are randomly assigned to levels of those factors.

Randomized designs differ in a variety of ways: (a) the number of factors investigated, (b) the number of levels of each factor and how those levels are selected, (c) how the levels of different factors are combined, and (d) whether the participants in the study experience only one treatment or more than one treatment (see **Analysis of Variance: Classification**). The simplest experimental designs have only a single factor. In the most extreme case, a single group of participants may experience an experimental treatment, but there is no similar control group. This design constitutes a *one-shot case study* [6], and permits only limited testing of hypotheses. If normative data on a psychological measure of depression are available, for example, one can compare posttreatment scores on depression for a single group to those norms. However, discrepancies from the norms could be caused either by the treatment or by differences between the individuals in the study and the characteristics of the participants in the norm group. A preferable design includes one or more control groups whose performance can be compared with that of the group of interest. When more than two groups are involved, one will typically be interested not only in whether there are any differences among the groups, but also in specific comparisons among combinations

2 Experimental Design

of group means. Designs in which the groups experience different levels of a single factor are referred to as *one-way designs*, because the groups differ in one way or along one dimension.

For practical or theoretical reasons, an experimenter may prefer to include multiple factors in a single study rather than performing a separate experiment to investigate each factor. An added factor could represent some characteristic of the participants such as gender. Including gender as a second factor along with the factor of treatment condition typically increases the power of the *F* test to detect the effect of the treatment factor as well as allowing a check on the consistency of the effect of that factor across male and female subgroups. When the various conditions included in a study represent combinations of levels of two different factors, the design is referred to as a *two-way* or **factorial design**. One-way designs can be represented with a schematic involving a group of a cells differing along one dimension. In the usual case, two-way designs can be represented as a two-dimensional table.

Experimental designs with multiple factors differ in which combinations of levels of the different factors are utilized. In most cases, all possible combinations of levels of the factors occur. The factors in such a design are said to be *crossed*, with all levels of one factor occurring in conjunction with every level of the other factor or factors. Thus, if there are *a* levels of Factor A and *b* levels of Factor B, there are *a* × *b* combinations of levels in the design. Each combination of levels corresponds to a different cell of the rectangular schematic of the design. Alternatively, a design may not include all of the possible combinations of levels. The most common example of such an incomplete design is one where nonoverlapping subsets of levels of one factor occur in conjunction with the different levels of the other factor. For example, in a comparison of psychoanalytic and cognitive behavioral therapies, the therapists may be qualified to deliver one method but not the other method. In such a case, the therapists would be *nested* within therapy methods. In contrast, if all therapists used both methods, therapists would be crossed with methods. Diagrams of these designs are shown in Figure 1.

Experimental designs also differ in terms of the way the levels of a particular factor are selected for inclusion in the experiment. In most instances, the levels are included because of an inherent interest in

Crossed design

	Psychoanalytic	Cognitive behavioral
Therapist 1		
Therapist 2		
Therapist 3		
Therapist 4		
Therapist 5		
Therapist 6		

Nested design

	Psychoanalytic	Cognitive behavioral
Therapist 1		missing
Therapist 2		missing
Therapist 3		missing
Therapist 4	missing	
Therapist 5	missing	
Therapist 6	missing	

Figure 1 Diagrams of crossed and nested designs

the specific levels. One might be interested in particular drug treatments or particular patient groups. In any replication of the experiment, the same treatments or groups would be included. Such factors are said to be *fixed* (see **Fisherian Tradition in Behavioral Genetics** and **Fixed and Random Effects**). Any generalization to other levels or conditions besides those included in the study must be made on nonstatistical grounds. Alternatively, a researcher can select the levels of a factor for inclusion in a study at random from some larger set of levels. Such factors are said to be random, and the random selection procedure permits a statistical argument to be made supporting the generalization of findings to the larger set of levels. The nature of the factors, fixed versus random, affects the way the statistical analysis of the data is carried out as well as the interpretation of the results.

Perhaps the most important distinction among experimental designs is whether the designs are *between-subjects* designs or *within-subjects* designs. The distinction is based on whether each subject experiences only one experimental condition or multiple experimental conditions. The advantage of between-subjects designs is that one does not need to be concerned about possible **carryover effects** from other conditions because the subject experiences only one condition. On the other hand, a researcher may want to use the same participants under different

conditions. For example, each participant can be used as his or her own control by contrasting that participant's performance under one condition with his or her performance under another condition. In many cases in psychology, the various conditions experienced by a given participant will correspond to observations at different points in time. For example, a test of clinical treatments may assess clients at each of several follow-up times. In this case, the same participants are observed multiple times. While there are obvious advantages to this procedure in terms of efficiency, conventional **analysis-of-variance** approaches to within-subjects designs (see **Repeated Measures Analysis of Variance** and **Educational Psychology: Measuring Change Over Time**) require that additional assumptions be made about the data. These assumptions are often violated. Furthermore, within-subjects designs require that participants have no **missing data** [5]. A variety of methods for dealing with these issues have been developed in recent years, including

various imputation methods and hierarchical linear modeling procedures.

References

- [1] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition, Wiley, New York.
- [2] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Dekker, New York.
- [3] Fisher, R.A. (1971). *Design of Experiments*, Hafner, New York. (Original work published 1935).
- [4] Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer, New York.
- [5] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data; A Model Comparison Perspective*, Erlbaum, Mahwah, NJ.
- [6] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

HAROLD D. DELANEY

Exploratory Data Analysis

SANDY LOVIE

Volume 2, pp. 586–588

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Exploratory Data Analysis

For most statisticians in the late 1960s and early 1970s, the really hot topics in their subject were definitely not simple descriptive statistics or basic plots. Instead, the fireworks came from the continuing debates over the nature of statistical inference and the foundations of probability, followed closely by the modeling of increasingly complex experimental designs within a general **analysis of variance** setting, an area that was boosted by the growing availability of cheap and powerful computing facilities. However, in an operation reminiscent of the *samizdat* in the latter days of the Soviet Union (which secretly circulated the banned novels of Aleksandr Solzhenitsyn, amongst other things), mimeographed copies of a revolutionary recasting of statistics began their hand-to-hand journey around a selected number of universities from early 1970 onwards.

This heavily thumbed, three-volume text was not the work of a dissident junior researcher impatient with the conservatism of their elders, but was by one of America's most distinguished mathematicians, **John Tukey**. What Tukey did was to reestablish the relationship between these neglected descriptive measures (and plots) and many of the currently active areas in statistics, including computing, modeling, and inference. In doing so, he also resuscitated interest in the development of new summary measures and displays, and in the almost moribund topic of regression. The central notion behind what he termed *Exploratory Data Analysis*, or EDA, is simple but fruitful: knowledge about the world comes from many wells and by many routes. One legitimate and productive way is to treat data not as merely supportive of already existing knowledge, but primarily as a source of new ideas and hypotheses. For Tukey, like a detective looking for the criminal, the world is full of potentially valuable material that could yield new insights into what might be the case. EDA is the active reworking of this material with the aim of extracting new meanings from it: 'The greatest value of a picture is when it *forces* us to notice what we never expected to see' [9, p. vi]. In addition, while EDA has very little to say about how data might be gathered (although there is a subdivision of the movement whose concentration on model evaluation could be said to have *indirect* implications for data collection), the movement is very concerned that the

tools to extract meaning from data do so with minimum ambiguity. What this indicates, using a slightly more technical term, is that EDA is concerned with *robustness*, that is, with the study of all the factors that disturb or distort conclusions drawn from data, and how to draw their teeth (see also [11] for an introduction to EDA, which nicely mixes the elementary with the sophisticated aspects of the topic).

A simple example here is the use of the median versus the arithmetic mean as a measure of *location*, where the former is said to be robust resistant to discrepant values or **outliers** because it is based only on the *position* of the middle number in the (ordered) sample, while the latter uses all the numbers and their values, however unrepresentative, in its calculation (see [5] for an introduction to outliers and their treatment and [7] for an account of outliers as collective social phenomena). Further, a **trimmed mean**, calculated from an *ordered* sample where, say, 10% of the top and bottom numbers have been removed, is also more robust resistant than the raw arithmetic mean based on all the readings, since trimming automatically removes any outliers, which, by definition, lie in the upper and lower tails of such a sample, and are few in number. In general, robust resistant measures are preferable to nonrobust ones because, while both give pretty much the same answer in samples where there are no outliers, the former yields a less distorted measure in samples where outliers are present. Tukey and some of his colleagues [1] set themselves the task of investigating over 35 separate measures of location for their robustness. Although the median and the percentage trimmed mean came out well, Tukey also promoted the computer-dependent *biweight* as his preferred measure of location. Comparable arguments can also be made for choosing a robust measure of *spread* (or *scale*) over a nonrobust one, with the *midspread*, that is, the middle 50% of a sample (or the difference between the upper and lower quartiles or *hinges*, to use Tukey's term), being more acceptable than the variance or the standard deviation. Since the experienced data explorer is unlikely to know beforehand what to expect with a given *batch* (Tukey's neologism for a *sample*), descriptive tools that are able to cope with messy data sets are preferable to the usual, but more outlier-prone, alternatives.

Plots that use robust rather than outlier-sensitive measures were also part of the EDA doctrine, thus the

ubiquitous **box and whisker plot** is based on both the median and the midspread, while the latter measure helps determine the whisker length, which can then be useful in identifying potential outliers. Similar novelties such as the **stem and leaf plot** are also designed for the median rather than the mean to be read off easily, as can the upper and lower quartiles, and hence the midspread. Rather more exotic birds such as *hanging or suspended rootograms* and *half normal plots* (all used for checking the Gaussian nature of the data) were either Tukey's own creation or that of workers influenced by him, for instance, Daniel and Wood [2], whose classic study of data modeling owed much to EDA. Tukey did not neglect earlier and simpler displays such as **scatterplots**, although, true to his radical program, these became transformed into windows onto data shape and symmetry, robustness, and even robust differences of location and spread. A subset of these displays, **residual** and **leverage plots**, for instance, also played a key role in the revival of interest in regression, particularly in the area termed *regression diagnostics* (see [6]). In addition, while EDA has concentrated on relatively simple data sets and data structures, there are less well known but still provocative incursions into the partitioning of multiway and multifactor tables [3], including Tukey's rethinking of the analysis of variance (see [4]).

While EDA offered a flexibility and adaptability to knowledge-building missing from the more formal processes of statistical inference, this was achieved with a loss of certainty about the knowledge gained – provided, of course, that one believes that more formal methods do generate truth, or your money back. Tukey was less sure on this latter point in that while formal methods of inference are bolted onto EDA in the form of Confirmatory Data Analysis, or CDA (which favored the Bayesian route – see Mosteller and Tukey's early account of EDA and CDA in [8]), he never expended as much effort on CDA as he did on EDA, although he often argued that both should run in harness, with the one complementing the other. Thus, for Tukey (and EDA), truth gained by such an epistemologically *inductive* method was always going to be partial, local and relative, and liable to fundamental challenge and change. On the other hand, without taking such risks, nothing new could emerge. What seemed to be on offer, therefore, was an *entrepreneurial* view of statistics, and *science*, as permanent revolution.

Although EDA broke cover over twenty-five years ago with the simultaneous publication of the two texts on EDA and regression [10, 9], it is still too early to say what has been the real impact of EDA on data analysis. On the one hand, many of EDA's graphical novelties have been incorporated into most modern texts and computer programs in statistics, as have the raft of robust measures on offer. On the other, the risky approach to statistics adopted by EDA has been less popular, with deduction-based inference taken to be the only way to discover the world. However, while one could point to underlying and often contradictory cultural movements in scientific belief and practice to account for the less than wholehearted embrace of Tukey's ideas, the future of statistics lies with EDA.

References

- [1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton.
- [2] Daniel, C. & Wood, F.S. (1980). *Fitting Equations to Data*, 2nd Edition, Wiley, New York.
- [3] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1985). *Exploring Data Tables, Trends and Shapes*, John Wiley, New York.
- [4] Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1992). *Fundamentals of Exploratory Analysis of Variance*, John Wiley, New York.
- [5] Lovie, P. (1986). Identifying outliers, in *New Developments in Statistics for Psychology and the Social Sciences*, A.D. Lovie, ed., BPS Books & Routledge, London.
- [6] Lovie, P. (1991). Regression diagnostics: a rough guide to safer regression, in *New Developments in Statistics for Psychology and the Social Sciences*, Vol. 2, P. Lovie & A.D. Lovie, eds, BPS Books & Routledge, London.
- [7] Lovie, A.D. & Lovie, P. (1998). The social construction of outliers, in *The Politics of Constructionism*, I. Velody & R. Williams, eds, Sage, London.
- [8] Mosteller, F. & Tukey, J.W. (1968). Data analysis including statistics, in *Handbook of Social Psychology*, G. Lindzey & E. Aronson, eds, Addison-Wesley, Reading.
- [9] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading.
- [10] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.
- [11] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

SANDY LOVIE

External Validity

DAVID C. HOWELL

Volume 2, pp. 588–591

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

External Validity

Cook and Campbell [2] identified many threats to our ability to generalize from an experiment to the larger population of people, settings, and methods. They distinguished between construct validity, which refers to the generalizability from the measures of an experiment to the constructs that were under investigation, and external validity, which refers to the degree to which any causal relationship that is found between an independent and a dependent variable can be generalized across people, settings, and times. For example, we often wonder whether the results we obtain in a tightly controlled social psychology experiment, with people communicating over an intercom rather than in person, would generalize to the 'real world'. In this case, we are talking about external validity.

Whereas **internal validity** refers to the degree to which any effect that is found for an experiment can be attributed to the causal action of the manipulated independent variable, external validity deals with generalizability. Campbell and Stanley considered internal validity to be the *sine qua non* of good experimental design. Without internal validity, it is not worth worrying about external validity. But without external validity, the interpretation of our results is severely limited, and perhaps meaningless for most practical situations.

Campbell and Stanley [1] were some of the first to discuss threats to external validity. This work was expanded by Cook and Campbell [2] in 1979, and, most recently, by Shadish, Cook, and Campbell [3] in 2002. This entry discusses five threats to external validity.

Threats to External Validity

- Interaction of selection and treatment
 - It is not uncommon to find that those who agree to participate in a particular experimental treatment differ from those who will not participate or drop out before the end of treatment. If this is the case, the results of those in the treatment condition may not generalize to the population of interest. Suppose,

for example, that we have a fairly unpleasant, though potentially effective, treatment for alcoholism. We create a treatment and a control condition by random assignment to conditions. Because the treatment is unpleasant, we have a substantial dropout rate in the treatment condition. Thus, at the end of the experiment, only the very highly motivated participants remain in that condition. In this case, the results, though perhaps internally valid, are unlikely to generalize to the population we wish to address all adults with alcohol problems. The treatment may be very good for the highly motivated, but it will not work for the poorly motivated because they will not stay with it.

- A similar kind of problem arises with special populations. For example, an experimental manipulation that has a particular effect on the ubiquitous 'college sophomore' may not apply, or may apply differently, to other populations. To show that a 'token economy' treatment works with college students does not necessarily mean that it will work with prison populations, to whom we might like to generalize.
- Interaction of setting and treatment
 - This threat refers to the fact that results obtained in one setting may not generalize to other settings. For example, when I was in graduate school in the mid-1960s, we conducted a verbal learning study on recruits on an air force base. The study was particularly boring for participants, but when their sergeant told them to participate, they participated! I have since wondered whether I would have obtained similar data if participation, using the same population of military personnel, was completely voluntary.
- Interaction of treatment effects with outcome measures
 - In the social sciences, we have a wealth of dependent variables to choose from, and we often choose the one that is the easiest to collect. For example, training programs for adolescents with behavior problems often target school attendance because that variable is easily obtained from school records. Intervention programs can be devised to improve

2 External Validity

attendance, but that does not necessarily mean that they improve the student's behavior in school, whether the student pays attention in class, whether the student masters the material, and so on. When we cannot generalize from one reasonable and desirable outcome variable to another, we compromise the external validity of the study.

- Interaction of treatment outcome with treatment variation
 - Often, the independent variable that we would like to study in an experiment is not easy to clearly define, and we select what we hope is an intelligent operationalization of that variable. (For example, we might wish to show that increasing the attention an adolescent receives from his or her peers will modify an undesirable behavior. However, you just need to think of the numerous ways we can 'pay attention' to someone to understand the problem.) Similarly an experimental study might use a multifaceted treatment, but when others in the future attempt to apply that in their particular setting, they may find that they only have the resources to implement some of the facets of the treatment. In these cases, the external validity of the study involving its ability to generalize to other settings, may be compromised.
 - Context-dependent mediation
 - Many causal relationships between an independent and a dependent variable are mediated by the presence or absence of another variable. For example, the degree to which your parents allowed you autonomy when you were growing up might affect your level of self-confidence, and that self-confidence might in turn influence the way you bring up your own children. In this situation, self-confidence is a mediating variable. The danger in generalizing from one experimental context to another involves the possibility that the mediating variable does not have the same influence in all contexts. For example, it is easy to believe that the mediating role of self-confidence may be quite different in children brought up under conditions of severe economic deprivation than in children brought up in a middle-class family.
- Other writers have proposed additional threats to external validity, and these are listed below. In general, any factors that can compromise our ability to generalize from the results of an experimental manipulation are threats to external validity.
- Interaction of history and treatment
 - Occasionally, the events taking place outside of the experiment influence the results. For example, experiments that happened to be conducted on September 11, 2001 may very well have results that differ from the results expected on any other day. Similarly, a study of the effects of airport noise may be affected by whether that issue has recently been widely discussed in the daily press.
 - Pretest-treatment interaction
 - Many experiments are designed with a pretest, an intervention, and a posttest. In some situations, it is reasonable to expect that the pretest will sensitize participants to the experimental treatment or cause them to behave in a particular way (perhaps giving them practice on items to be included in the posttest.). In this case, we would have difficulty generalizing to those who received the intervention but had not had a pretest.
 - Multiple-treatment interference
 - Some experiments are designed to have participants experience more than one treatment (hopefully in random order). In this situation, the response to one treatment may depend on the other treatments the individual has experienced, thus limiting generalizability.
 - Specificity of variables
 - Unless variables are clearly described and operationalized, it may be difficult to replicate the setting and procedures in a subsequent implementation of the intervention. This is one reason why good clinical psychological research often involves a very complete manual on the implementation of the intervention.
 - Experimenter bias
 - This threat is a threat to both **internal validity** and external validity. If even good experimenters have a tendency to see what they

expect to see, the results that they find in one setting, with one set of expectations, may not generalize to other settings.

- Reactivity effects
 - A classic study covered in almost any course on experimental design involves what is known as the **Hawthorne effect**. This is often taken to refer to the fact that even knowing that you are participating in an experiment may alter your performance. **Reactivity** effects refer to the fact that participants often ‘react’ to the very existence of an experiment in ways that they would not otherwise react.

For other threats to invalidity, see the discussion in the entry on **internal validity**. For approaches to

dealing with these threats, see **Nonequivalent Group Design**, **Regression Discontinuity Design**, and, particularly, **Quasi-experimental Designs**.

References

- [1] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Skokie.
- [2] Cook, T. & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston.
- [3] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.

DAVID C. HOWELL

Face-to-Face Surveys

JAMES K. DOYLE

Volume 2, pp. 593–595

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Face-to-Face Surveys

A face-to-face survey is a **telephone survey** without the telephone. The interviewer physically travels to the respondent's location to conduct a personal interview. Unlike the freewheeling type of interview one sees on *60 Minutes*, where the interviewer adapts the questions on the fly based on previous responses (or lack thereof), face-to-face surveys follow a standardized script without deviation, just like a mail or telephone survey. From the respondent's point of view, the process could not be easier: the interviewer arrives at a convenient, prearranged time, reads the survey for you, deals with any questions or problems that arise, records your answers, and is shown the door. No one calls you during supper and there are no envelopes to lick. This ease of response in fact makes face-to-face surveys ideally suited for populations that have difficulty answering mail or telephone surveys due to poor reading or writing skills, disability, or infirmity.

Compared with mail and telephone surveys, face-to-face surveys offer significant advantages in terms of the amount and complexity of the data that can be collected. For example, face-to-face surveys can be significantly longer. Most people will allow an interviewer to occupy their living room couch for up to an hour, whereas respondents will typically not tolerate telephone interviews that extend much beyond half an hour or mail surveys that require more than 15 or 20 min of effort. The additional length allows researchers the opportunity to ask more questions, longer questions, more detailed questions, more open-ended questions, and more complicated or technical questions. Skip patterns, in which different respondents navigate different paths through the survey depending on their answers, also can be more complicated. In addition, the use of graphic or visual aids, impossible by telephone and costly by mail, can be easily and economically incorporated into face-to-face surveys.

Face-to-face surveys also offer advantages in terms of data quality. More than any other survey delivery mode, a face-to-face survey allows researchers a high degree of control over the data collection process and environment. Interviewers can ensure, for example, that respondents do not skip ahead or 'phone a friend', as they might do when filling out a mail survey, or that they do not watch

TV or surf the internet during the interview, as they might do during a telephone survey. Since the interviewer elicits and records the data, the problems of missing data, ambiguous markings, and illegible handwriting that plague mail surveys are eliminated. If the respondent finds a question to be confusing or ambiguous, the interviewer can immediately clarify it. Similarly, the respondent can be asked to clarify any answers that the interviewer cannot interpret.

Perhaps the most important procedural variable affecting data quality in a survey study is the response rate, that is, the number of completed questionnaires obtained divided by the number of people who were asked to complete them. Since it is much more difficult for people to shut the door in the face of a live human being than hang up on a disembodied voice or toss a written survey into the recycling bin with the junk mail, face-to-face surveys typically offer the highest response rates obtainable (over 90% in some cases). Like telephone surveys, face-to-face surveys also avoid a type of response bias typical of mail surveys, namely, the tendency for respondents, on average, to be more highly educated than those who fail to respond.

Of course, all of these benefits typically come at a great cost to the researchers, who must carefully hire, train, and monitor the interviewers and pay them to travel from one neighborhood to the next (and sometimes back again) knocking on doors. Largely due to the nature and cost of the travel involved, face-to-face surveys can end up costing more than twice as much and taking more than three times as long to complete as an equivalent telephone survey. Face-to-face surveys can also have additional disadvantages. For example, budgetary constraints typically limit them to a comparatively small geographical area. Also, some populations can be difficult to reach in person because they are rarely at home (e.g., college students), access to their home or apartment is restricted, or traveling in their neighborhood places interviewers at risk. There is also evidence that questions of a personal nature are less likely to be answered fully and honestly in a face-to-face survey. This is probably because respondents lose the feeling of anonymity that is easily maintained when the researcher is safely ensconced in an office building miles away. In addition, because face-to-face interviews put people on the spot by requiring an immediate answer, questions that require a lot

2 Face-to-Face Surveys

of reflection or a search for personal records are better handled by the self-paced format of a mail survey.

Perhaps the largest 'cost' associated with a face-to-face survey is the increased burden placed on the researcher to ensure that the interviewers who are collecting the data do not introduce 'interviewer bias', that is, do not, through their words or actions, unintentionally influence respondents to answer in a particular way. While interviewer bias is also a concern in telephone surveys, it poses even more of a problem in face-to-face surveys for two reasons. First, the interviewer is exposed to the potentially biasing effect of the respondent's appearance and environment in addition to their voice. Second, the interviewer may inadvertently give respondents nonverbal as well as verbal cues about how they should respond. Interviewing skills do not come naturally to people because a standardized interview violates some of the normative rules of efficient conversation. For instance, interviewers must read all questions and response options exactly as written rather than paraphrasing them, since even small changes in wording have the potential to influence survey outcomes. Interviewers also have to ask a question even when the respondent has already volunteered the answer. To reduce bias as well as to avoid interviewer effects, that is, the tendency for the data collected by different interviewers to differ due to procedural inconsistency, large investments must typically be made in providing interviewers the necessary training and practice. Data analyses of face-to-face surveys should also examine and report on any significant interviewer effects identified in the data.

In summary, face-to-face surveys offer many advantages over mail and telephone surveys in terms of the complexity and quality of the data collected, but these advantages come with significantly increased logistical costs as well as additional potential sources of response bias. The costs are in fact so prohibitive that face-to-face surveys are typically employed only when telephone surveys are impractical, for example, when the questionnaire is too long or complex to deliver over the phone or when a significant proportion of the population of interest lacks telephone access.

Further Reading

- Czaja, R. & Blair, J. (1996). *Designing Surveys: A Guide to Decisions and Procedures*, Pine Forge Press, Thousand Oaks.
- De Leeuw, E.D. & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis, in *Telephone Survey Methodology*, R.M. Groves, P.N. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nichols II & J. Waksberg, eds, Wiley, New York, pp. 283–299.
- Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*, 2nd Edition, Wiley, New York.
- Fowler, F.J. (1990). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Sage, Newbury Park.
- Fowler Jr, F.J. (2002). *Survey Research Methods*, 3rd Edition, Sage, Thousand Oaks.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*, Wiley, New York.
- Groves, R.M. & Kahn, R.L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*, Academic Press, Orlando.
- Hyman, H., Feldman, J. & Stember, C. (1954). *Interviewing in Social Research*, University of Chicago Press, Chicago.

JAMES K. DOYLE

Facet Theory

INGWER BORG

Volume 2, pp. 595–599

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Facet Theory

Facet theory (FT) is a methodology that may be considered an extension of methods for the design and analysis of experiments. Experiments use one or more factors that characterize the conditions under

$\mathbf{R}$ is recorded for each such person-question crossing, we have the mapping $\mathbf{P} \times \mathbf{Q} \rightarrow \mathbf{R}$.

Things become more interesting when $\mathbf{Q}$ (and $\mathbf{P}$) are *facetized* and embedded into the relational context of a *mapping sentence*. For example, one may want to assess the intelligence of students by different types of test items that satisfy the following definition:

Student (p) solves a		
$\mathbf{Q}_1 = \text{language}$		$\mathbf{Q}_2 = \text{operation}$
(numerical)		(finding)
(geometrical)	task that requires	(applying)
(verbal)		(learning)
	$\mathbf{R} = \text{range}$	
	(right)	
	(to)	in the sense of the objective rule
	(wrong)	

which the variables of interest are observed. Data analysis, then, studies the effects of these factors on the dependent variables. FT extends these notions to the social sciences, proposing to systematically design the researcher's questions and items over a framework of facets, and then study the resulting data with data analytic tools that ask how these facets show up in the observations [2].

Facet Design

Designing observations in FT means, first of all, characterizing them in terms of *facets*. A facet is a variable that allows one to sort objects of interest into different classes. For example, 'gender' is a facet that sorts persons into the classes male and female. Similarly, 'mode of attitude' serves to classify attitudinal behavior as either emotional, cognitive, or overt action.

In survey research, facets are used routinely to stratify a population $\mathbf{P}$ in order to generate samples with guaranteed distributions on important background variables. The same idea can be carried over to the *universe of questions* ($\mathbf{Q}$): introducing *content facets* and systematically crossing them defines various types of questions. Finally, the third component in an empirical query is the set of responses ($\mathbf{R}$) that are considered relevant when $\mathbf{P}$ is confronted with $\mathbf{Q}$. In the usual case, where each person in $\mathbf{P}$ is given every question in $\mathbf{Q}$, and exactly one response out of

A mapping sentence interconnects the facets and clarifies their roles within the context of a particular *content domain*. The example shows that $\mathbf{Q}_2 = \text{'operation'}$ is the *stem* facet, while $\mathbf{Q}_1 = \text{'language'}$ is a *modifying* facet for IQ tasks belonging to the design $\mathbf{Q}_1 \times \mathbf{Q}_2$. The population facet $\mathbf{P}$ is not further facetized here. The mapping sentence for $\mathbf{P} \times \mathbf{Q}_1 \times \mathbf{Q}_2 \rightarrow \mathbf{R}$ serves as a blueprint for systematically constructing concrete test items. The two $\mathbf{Q}$ -facets can be fully crossed, and so there are nine types of tasks. For example, we may want to construct tasks that assess the student's ability to apply (A) or to find (F) rules, both with numerical (N) and geometrical (G) tasks. This yields four types of tasks, each characterized by a combination (*structuple*) such as AN, AG, FN, or FG.

The structuple, however, only describes an item's question part. The item is incomplete without its range of admissible answers. In the above example, the *range facet* ($\mathbf{R}$) is of particular interest, because it represents a *common* range for all questions that can be constructed within this definitional framework. Indeed, according to Guttman [6], an intelligence item can be distinguished from other items (e.g., such as attitude items) by satisfying exactly this range. That is, an item is an intelligence item if and only if the answers to its question are mapped onto a right-to-wrong (according to a logical, scientific, or factual rule) continuum.

Mapping sentences are useful devices for conceptually organizing a domain of research questions. They enable the researcher to structure a universe

2 Facet Theory

of items and to systematically construct a sample of items that belong to this universe. Mapping sentences are, however, not a tool that automatically yields meaningful results. One needs solid substantive knowledge and clear semantics to make them work.

Correspondence Hypotheses Relating Design to Data

A common range of items gives rise to certain *monotonicity hypotheses*. For intelligence items, Guttman [6] predicts that they correlate nonnegatively among each other, which is confirmed for the case shown in Table 1. This ‘first law of intelligence’ is a well-established empirical regularity for the universe of intelligence items. A similar law holds for attitude items.

A more common-place hypothesis in FT is to check whether the various facets of the study’s design show up, in one way or another, in the structure of the data. More specifically, one often uses the *discrimination hypothesis* that the distinctions a facet makes for the types of observations should be reflected in differences of the data for these types. Probably, the most successful variant of this hypothesis is that the **Q**-facets should allow one to partition a space

that represents the items’ empirical similarities into simple *regions*. Figure 1 shows three prototypical patterns that often result in this context. The space here could be an **multidimensional scaling** (MDS) representation of the intercorrelations of a battery of items. The plots are *facet diagrams*, where the points are replaced by the element (*struct*) that the item represented by a particular point has on a particular facet. Hence, if we look at the configuration in terms of facet 1 (left panel), the points form a pattern that allows us to cut the plane into parallel stripes. If the facet were ordered so that $a < b < c$, this *axial* pattern leads to an embryonic dimensional interpretation of the X-axis. The other two panels of Figure 1 show patterns that are also often found in real applications, that is, a *modular* and a *polar* regionalization, respectively. Combined, these prototypes give rise to various partitionings such as the *radex* (a modular combined with a polar facet) or the *duplex* (two axial facets). Each such pattern is found by partitioning the space facet by facet.

A third type of correspondence hypothesis used in FT exists for design (or data) structuples whose facets are all ordered in a common sense. Elizur [3], for example, asked persons to indicate whether they were

Table 1 Intercorrelations of eight intelligence test items (Guttman, 1965)

Structuple	Task	1	2	3	4	5	6	7	8
NA	1	1.00							
NA	2	0.67	1.00						
NF	3	0.40	0.50	1.00					
GF	4	0.19	0.26	0.52	1.00				
GF	5	0.12	0.20	0.39	0.55	1.00			
GA	6	0.25	0.28	0.31	0.49	0.46	1.00		
GA	7	0.26	0.26	0.18	0.25	0.29	0.42	1.00	
GA	8	0.39	0.38	0.24	0.22	0.14	0.38	0.40	1.00

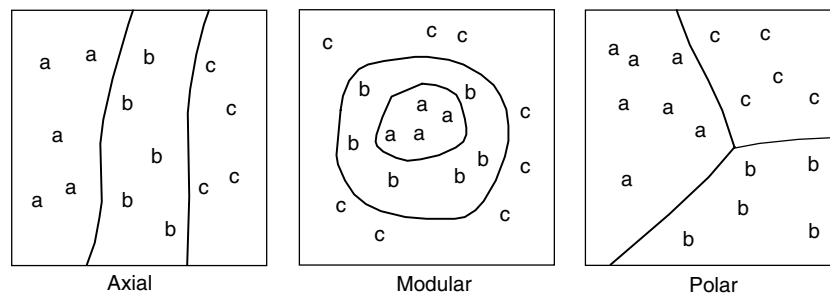


Figure 1 Prototypical partitionings of facet diagrams over an MDS plane

Table 2 Profiles of concern [3]

Profile	A = interest	B = expertise	C = stability	D = employment	Freq.
1	0	0	0	0	85
2	0	0	0	1	38
3	0	0	1	0	28
4	1	0	0	0	3
5	0	1	0	1	18
6	0	0	1	1	37
7	1	0	1	0	2
8	1	1	0	0	5
9	0	1	1	1	5
10	1	0	1	1	2
11	1	1	1	0	2
12	1	1	1	1	53

concerned or not about losing certain features of their job after the introduction of computers to their workplace. The design of this study was ‘Person (p) is concerned about losing $\rightarrow \mathbf{A} \times \mathbf{B} \times \mathbf{C} \times \mathbf{D}$ ’, where **A** = ‘interest (yes/no)’, **B** = ‘experience (yes/no)’, **C** = ‘stability (yes/no)’, and **D** = ‘employment (yes/no)’. Each person generated an answer profile with four elements such as 1101 or 0100, where 1 = yes and 0 = no. Table 2 lists 98% of the observed profiles. We now ask whether these profiles form a Guttman scale [4] or, if not, whether they form a partial order with nontrivial properties. For example, the partial order may be flat in the sense that it can be represented in a plane such that its Hasse diagram has no paths that cross each other outside of common points (*diamond*). If so, it can be explained by two dimensions [10].

A soft variant of relating structuples to data is the hypothesis that underlies *multidimensional scalogram analysis* or MSA [9]. It predicts that given design (or data) structuples can be placed into in a multidimensional space of given dimensionality so that this space can be partitioned, facet by facet, into simple regions as shown, for example, in Figure 1. (Note though that MSA does not involve first computing any overall proximities. It also places no requirements on the scale level of the facets of the structuples.)

Facet-theoretical Data Analysis

Data that are generated within a facet-theoretical design are often analyzed by special data analytic methods or by traditional statistics used in particular ways. An example for the latter is given in Figure 2.

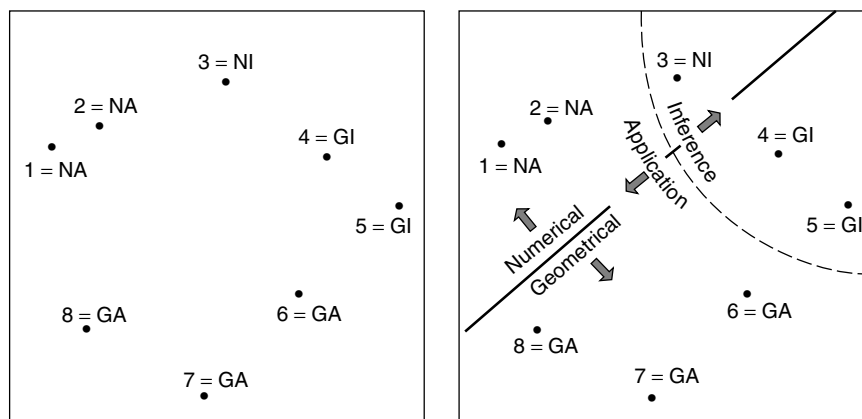


Figure 2 MDS representation of correlations of Table 1 (left panel) and partitioning by facets ‘language’ and ‘operation’ (right panel)

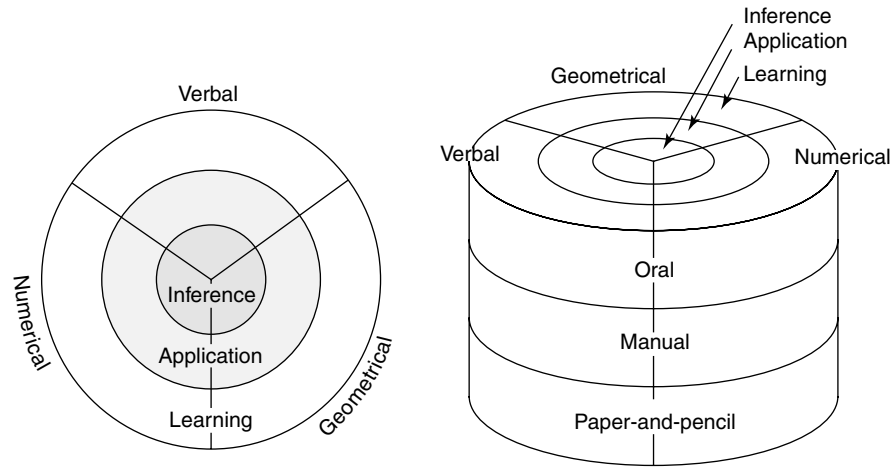


Figure 3 Structure of the universe of intelligence test items

Its left panel shows a two-dimensional multidimensional scaling space for the data in Table 1. The right panel of this figure demonstrates that the space can be divided by each of the two design facets, ‘language’ and ‘operation’, into two regions. Each region contains only items of one particular type. The resulting pattern partially confirms a structure found for the universe of typical paper-and-pencil intelligence test items. The universe has an additional facet (‘mode of communicating’) that shows up as the axis of the *cylindrex* shown in Figure 3. One also notes that the facet ‘operation’ turns out to be ordered, which stimulates theoretical thinking aimed at explaining this more formally.

Partial order hypotheses for structuples can be checked by POSAC (partial order structuple analysis with coordinates; [8]). POSAC computer programs are available within SYSTAT or HUDAP [1]. Figure 4 shows a POSAC solution for the data in Table 2. The 12 profiles here are represented as points. Two points are connected by a line if their profiles can be ordered such that profile x is ‘higher’ (in the sense of the common range ‘concern’) than profile y on at least one facet and not lower on any other facet. One notes that the profiles on each path that is monotonic with respect to the joint direction – as, for example, the chain 1-2-5-9-12 – form a perfect Guttman scale (see **Unidimensional Scaling**). One also notes that the various paths in the partial order do not cross each other except possibly in points that they share. Hence, the partial order is

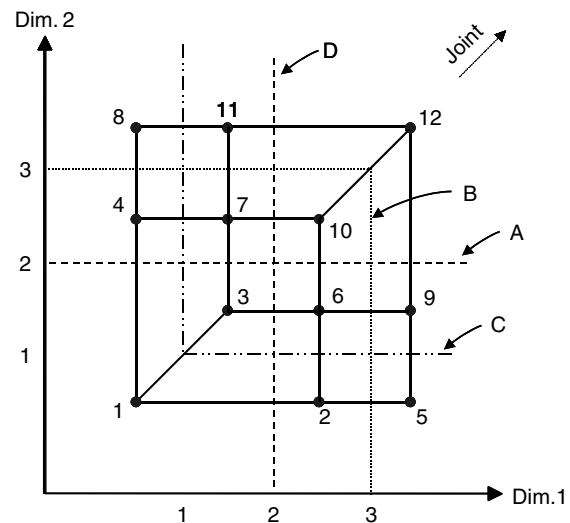


Figure 4 POSAC solution for the data in Table 2

flat and can be explained by two dimensions. Dimension 1 has a simple relation to facet **D**: all profiles that have a low value on that dimension have a zero on **D**, that is, these persons are not concerned about losing their employment; all profiles with high values on dimension 1 are high on **D**. Dimension 2 has a similar relation to facet **A**. Hence, facets **D** and **A** explain the dimensions. The remaining two facets play secondary roles in this context: **B** is ‘accentuating’ [10] in the sense that persons who agree to

B are very high on at least one of the base dimensions. **C**, in contrast, is ‘attenuating’ so that persons high on **C** are relatively similar on their *X* and *Y* coordinates. Both secondary facets thus generate additional cutting points and thereby more intervals on the base dimensions.

HUDAP also contains programs for doing MSA. MSA is seldom used in practice, because its solutions are rather indeterminate [2]. They can be transformed in many ways that can radically change their appearance, which makes it difficult to interpret and to replicate them. One can, however, make MSA solutions more robust by enforcing additional constraints such as linearity onto the boundary lines or planes [9]. Such constraints are not related to content and for that reason rejected by many facet theorists who prefer data analysis that is as ‘intrinsic’ [7] as possible. On the other hand, if a good MSA solution is found under ‘extrinsic’ side constraints, it certainly also exists for the softer intrinsic model.

References

- [1] Amar, R. & Toledano, S. (2001). *HUDAP Manual*, Hebrew University of Jerusalem, Jerusalem.
- [2] Borg, I. & Shye, S. (1995). *Facet Theory: form and Content*, Sage, Newbury Park.
- [3] Elizur, D. (1970). *Adapting to Innovation: A Facet Analysis of the Case of the Computer*, Jerusalem Academic Press, Jerusalem.
- [4] Guttman, L. (1944). A basis for scaling qualitative data, *American Sociological Review* **9**, 139–150.
- [5] Guttman, L. (1965). A faceted definition of intelligence, *Scripta Hierosolymitana* **14**, 166–181.
- [6] Guttman, L. (1991a). Two structural laws for intelligence tests, *Intelligence* **15**, 79–103.
- [7] Guttman, L. (1991b). *Louis Guttman: In Memoriam—chapters from an Unfinished Textbook on Facet Theory*, Israel Academy of Sciences and Humanities, Jerusalem.
- [8] Levy, S. & Guttman, L. (1985). The partial-order of severity of thyroid cancer with the prognosis of survival, in *Ins and Outs of Solving Problems*, J.F. Marchiorchino, J.-M. Proth & J. Jansen, eds, Elsevier, Amsterdam, pp. 111–119.
- [9] Lingoes, J.C. (1968). The multivariate analysis of qualitative data, *Multivariate Behavioral Research* **1**, 61–94.
- [10] Shye, S. (1985). *Multiple Scaling*, North-Holland, Amsterdam.

(See also **Multidimensional Unfolding**)

INGWER BORG

Factor Analysis: Confirmatory

BARBARA M. BYRNE

Volume 2, pp. 599–606

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Analysis: Confirmatory

Of primary import to **factor analysis**, in general, is the notion that some variables of theoretical interest cannot be observed directly; these unobserved variables are termed **latent variables** or *facto-
tors*. Although latent variables cannot be measured directly, information related to them can be obtained indirectly by noting their effects on observed variables that are believed to represent them. The oldest and best-known statistical procedure for investigating relations between sets of observed and latent variables is that of factor analysis. In using this approach to data analyses, researchers examine the covariation among a set of observed variables in order to gather information on the latent constructs (i.e., factors) that underlie them. In factor analysis models, each observed variable is hypothesized to be determined by two types of influences: (a) the latent variables (factors) and (b) unique variables (called either *residual* or *error variables*). The strength of the relation between a factor and an observed variable is usually termed the *loading of the variable* on the factor.

Exploratory versus Confirmatory Factor Analysis

There are two basic types of factor analyses: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is most appropriately used when the links between the observed variables and their underlying factors are unknown or uncertain. It is considered to be exploratory in the sense that the researcher has no prior knowledge that the observed variables do, indeed, measure the intended factors. Essentially, the researcher uses EFA to determine factor structure. In contrast, CFA is appropriately used when the researcher has some knowledge of the underlying latent variable structure. On the basis of theory and/or empirical research, he or she postulates relations between the observed measures and the underlying factors *a priori*, and then tests this hypothesized structure statistically. More specifically, the CFA approach examines the extent to which a highly constrained *a priori* factor structure is consistent with the sample data. In summarizing the primary distinction between the two methodologies, we can say

that whereas EFA operates inductively in allowing the observed data to determine the underlying factor structure *a posteriori*, CFA operates deductively in postulating the factor structure *a priori* [5].

Of the two factor analytic approaches, CFA is by far the more rigorous procedure. Indeed, it enables the researcher to overcome many limitations associated with the EFA model; these are as follows: First, whereas the EFA model assumes that all common factors are either correlated or uncorrelated, the CFA model makes no such assumptions. Rather, the researcher specifies, *a priori*, only those factor correlations that are considered to be substantively meaningful. Second, with the EFA model, all observed variables are directly influenced by all common factors. With CFA, each factor influences only those observed variables with which it is purported to be linked. Third, whereas in EFA, the unique factors are assumed to be uncorrelated, in CFA, specified covariation among particular uniquenesses can be tapped. Finally, provided with a malfitting model in EFA, there is no mechanism for identifying which areas of the model are contributing most to the misfit. In CFA, on the other hand, the researcher is guided to a more appropriately specified model via indices of misfit provided by the statistical program.

Hypothesizing a CFA Model

Given the *a priori* knowledge of a factor structure and the testing of this factor structure based on the analysis of covariance structures, CFA belongs to a class of methodology known as **structural equation modeling** (SEM). The term structural equation modeling conveys two important notions: (a) that structural relations can be modeled pictorially to enable a clearer conceptualization of the theory under study, and (b) that the causal processes under study are represented by a series of structural (i.e., regression) equations. To assist the reader in conceptualizing a CFA model, I now describe the specification of a hypothesized CFA model in two ways; first, as a graphical representation of the hypothesized structure and, second, in terms of its structural equations.

Graphical Specification of the Model

CFA models are schematically portrayed as path diagrams (see **Path Analysis and Path Diagrams**) through the incorporation of four geometric symbols: a circle (or ellipse) representing unobserved

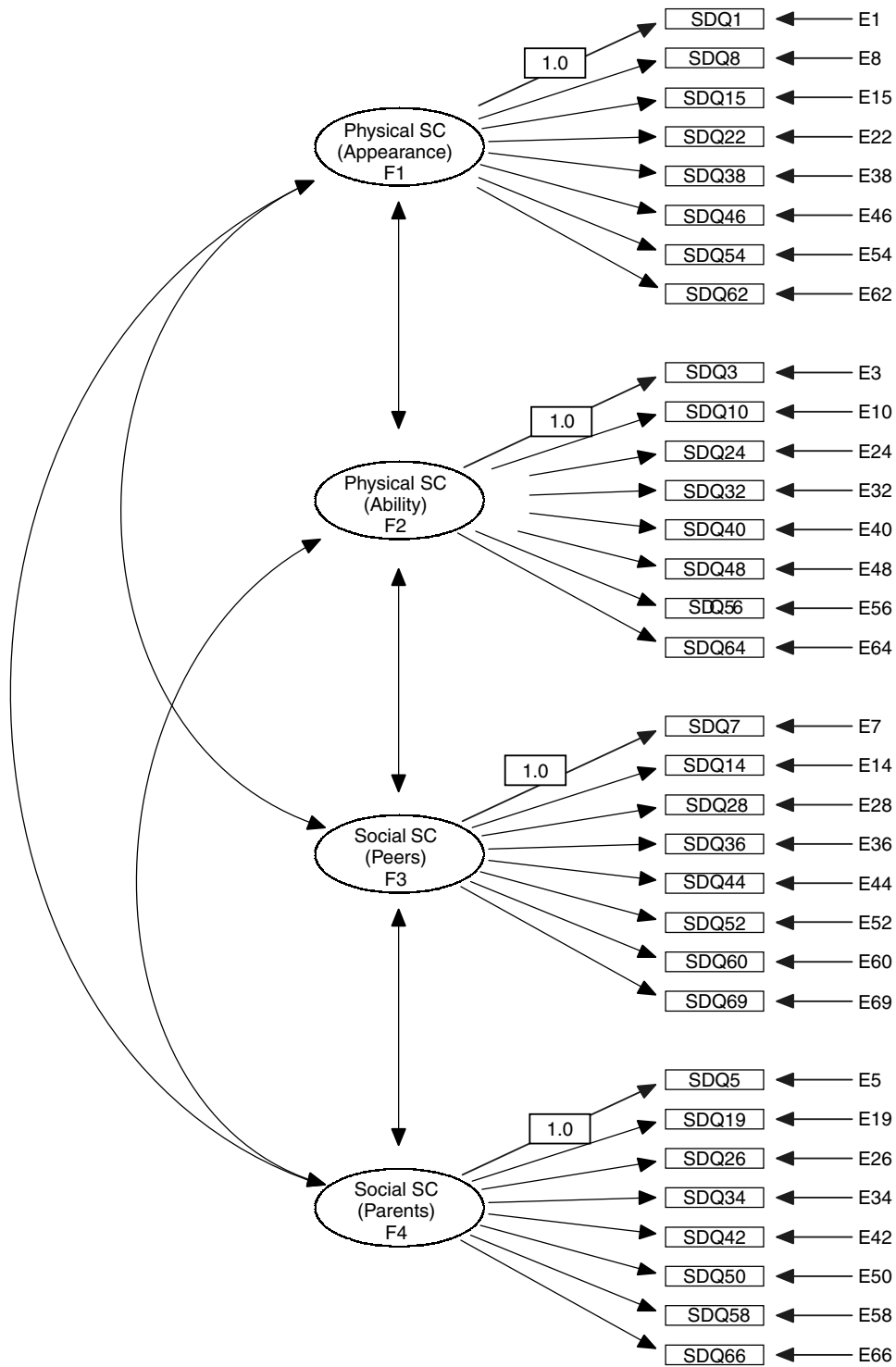


Figure 1 Hypothesized CFA model

latent factors, a square (or rectangle) representing observed variables, a single-headed arrow ($\rightarrow$) representing the impact of one variable on another, and a double-headed arrow ($\leftrightarrow$) representing covariance between pairs of variables. In building a CFA model, researchers use these symbols within the framework of three basic configurations, each of which represents an important component in the analytic process. We turn now to the CFA model presented in Figure 1, which represents the postulated four-factor structure of nonacademic self-concept (SC) as tapped by items comprising the Self Description Questionnaire-I (SDQ-I; [15]). As defined by the SDQ-I, nonacademic SC embraces the constructs of physical and social SCs.

On the basis of the geometric configurations noted above, decomposition of this CFA model conveys the following information: (a) there are four factors, as indicated by the four ellipses labeled Physical SC (Appearance; F1), Physical SC (Ability; F2), Social SC (Peers; F3), and Social SC (Parents; F4); (b) the four factors are intercorrelated, as indicated by the six two-headed arrows; (c) there are 32 observed variables, as indicated by the 32 rectangles (SDQ1-SDQ66); each represents one item from the SDQ-I; (d) the observed variables measure the factors in the following pattern: Items 1, 8, 15, 22, 38, 46, 54, and 62 measure Factor 1, Items 3, 10, 24, 32, 40, 48, 56, and 64 measure Factor 2, Items 7, 14, 28, 36, 44, 52, 60, and 69 measure Factor 3, and Items 5, 19, 26, 34, 42, 50, 58, and 66 measure Factor 4; (e) each observed variable measures one and only one factor; and (f) errors of measurement associated with each observed variable (E1-E66) are uncorrelated (i.e., there are no double-headed arrows connecting any two error terms. Although the error variables, technically speaking, are unobserved variables, and should have ellipses around them, common convention in such diagrams omits them in the interest of clarity.

In summary, a more formal description of the CFA model in Figure 1 argues that: (a) responses to the SDQ-I are explained by four factors; (b) each item has a nonzero loading on the nonacademic SC factor it was designed to measure (termed *target loadings*), and zero loadings on all other factors (termed *nontarget loadings*); (c) the four factors are correlated; and (d) measurement error terms are uncorrelated.

Structural Equation Specification of the Model

From a review of Figure 1, you will note that each observed variable is linked to its related factor by a single-headed arrow pointing *from* the factor *to* the observed variable. These arrows represent regression paths and, as such, imply the influence of each factor in predicting its set of observed variables. Take, for example, the arrow pointing from Physical SC (Ability) to SDQ1. This symbol conveys the notion that responses to Item 1 of the SDQ-I assessment measure are 'caused' by the underlying construct of physical SC, as it reflects one's perception of his or her physical ability. In CFA, these symbolized regression paths represent factor loadings and, as with all factor analyses, their strength is of primary interest. Thus, specification of a hypothesized model focuses on the formulation of equations that represent these structural regression paths. Of secondary importance are any covariances between the factors and/or between the measurement errors.

The building of these equations, in SEM, embraces two important notions: (a) that any variable in the model having an arrow pointing at it represents a dependent variable, and (b) dependent variables are always explained (i.e., accounted for) by other variables in the model. One relatively simple approach to formulating these structural equations, then, is first to note each dependent variable in the model and then to summarize all influences on these variables. Turning again to Figure 1, we see that there are 32 variables with arrows pointing toward them; all represent observed variables (SDQ1-SDQ66). Accordingly, these regression paths can be summarized in terms of 32 separate equations as follows:

$$\begin{aligned}
 \text{SDQ1} &= \text{F1} + \text{E1} \\
 \text{SDQ8} &= \text{F1} + \text{E8} \\
 \text{SDQ15} &= \text{F1} + \text{E15} \\
 &\vdots \\
 \text{SDQ62} &= \text{F1} + \text{E62} \\
 \text{SDQ3} &= \text{F2} + \text{E3} \\
 \text{SDQ10} &= \text{F2} + \text{E10} \\
 &\vdots \\
 \text{SDQ64} &= \text{F2} + \text{E64} \\
 \text{SDQ7} &= \text{F3} + \text{E7} \\
 \text{SDQ14} &= \text{F3} + \text{E14} \\
 &\vdots \\
 \text{SDQ69} &= \text{F3} + \text{E69}
 \end{aligned}$$

$$\begin{aligned}
 \text{SDQ5} &= \text{F4} + \text{E5} \\
 \text{SDQ19} &= \text{F4} + \text{E19} \\
 &\vdots \\
 \text{SDQ66} &= \text{F4} + \text{E66}
 \end{aligned}
 \tag{1}$$

Although, in principle, there is a one-to-one correspondence between the schematic presentation of a model and its translation into a set of structural equations, it is important to note that neither one of these representations tells the whole story. Some parameters, critical to the estimation of the model, are not explicitly shown and thus may not be obvious to the novice CFA analyst. For example, in both the schematic model (see Figure 1) and the linear structural equations cited above, there is no indication that either the factor variances or the error variances are parameters in the model. However, such parameters are essential to all structural equation models and therefore must be included in the model specification. Typically, this specification is made via a separate program command statement, although some programs may incorporate default values. Likewise, it is equally important to draw your attention to the specified nonexistence of certain parameters in a model. For example, in Figure 1, we detect no curved arrow between E1 and E8, which would suggest the lack of covariance between the error terms associated with the observed variables SDQ1 and SDQ8. (Error covariances can reflect overlapping item content and, as such, represent the same question being asked, but with a slightly different wording.)

Testing a Hypothesized CFA Model

Testing for the validity of a hypothesized CFA model requires the satisfaction of certain statistical assumptions and entails a series of analytic steps. Although a detailed review of this testing process is beyond the scope of the present chapter, a brief outline is now presented in an attempt to provide readers with at least a flavor of the steps involved. (For a non-mathematical and paradigmatic introduction to SEM based on three different programmatic approaches to the specification and testing of a variety of basic CFA models, readers are referred to [6–9]; for a more detailed and analytic approach to SEM, readers are referred to [3], [14], [16] and [17].)

Statistical Assumptions

As with other multivariate methodologies, SEM assumes that certain statistical conditions have been met. Of primary importance is the assumption that the data are multivariate normal (see **Catalogue of Probability Density Functions**). In essence, the concept of multivariate normality embraces three requirements: (a) that the univariate distributions are normal; (b) that the joint distributions of all variable combinations are normal; and (c) that all bivariate scatterplots are linear and homoscedastic [14]. Violations of multivariate normality can lead to the distortion of goodness-of-fit indices related to the model as a whole (see e.g., [12]; [10]; and (see **Goodness of Fit**) to positively biased tests of significance related to the individual parameter estimates [14]).

Estimating the Model

Once the researcher determines that the statistical assumptions have been met, the hypothesized model can then be tested statistically in a simultaneous analysis of the entire system of variables. As such, some parameters are freely estimated while others remain fixed to zero or some other nonzero value. (Nonzero values such as the 1's specified in Figure 1 are typically assigned to certain parameters for purposes of model identification and latent factor scaling.) For example, as shown in Figure 1, and in the structural equation above, the factor loading of SDQ8 on Factor 1 is freely estimated, as indicated by the single-headed arrow leading from Factor 1 to SDQ8. By contrast, the factor loading of SDQ10 on Factor 1 is not estimated (i.e., there is no single-headed arrow leading from Factor 1 to SDQ10); this factor loading is automatically fixed to zero by the program. Although there are four main methods for estimating parameters in CFA models, **maximum likelihood estimation** remains the one most commonly used and is the default method for all SEM programs.

Evaluating Model Fit

Once the CFA model has been estimated, the next task is to determine the extent to which its specifications are consistent with the data. This evaluative process focuses on two aspects: (a) goodness-of-fit of the model as a whole, and (b) goodness-of-fit of individual parameter estimates. Global assessment of fit

is determined through the examination of various fit indices and other important criteria. In the event that goodness-of-fit is adequate, the model argues for the plausibility of postulated relations among variables; if it is inadequate, the tenability of such relations is rejected. Although there is now a wide array of fit indices from which to choose, typically only one or two need be reported, along with other fit-related indicators. A typical combination of these evaluative criteria might include the Comparative Fit Index (CFI; Bentler, [1]), the standardized root mean square residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA; [18]), along with its 90% confidence interval. Indicators of a well-fitting model would be evidenced from a CFI value equal to or greater than .93 [11], an SRMR value of less than .08 [11], and an RMSEA value of less than .05 [4].

Goodness-of-fit related to individual parameters of the model focuses on both the appropriateness (i.e., no negative variances, no correlations >1.00) and statistical significance (i.e., estimate divided by standard error >1.96) of their estimates. For parameters to remain specified in a model, their estimates must be statistically significant.

Post Hoc Model-fitting

Presented with evidence of a poorly fitting model, the hypothesized CFA model would be rejected. Analyses then proceed in an exploratory fashion as the researcher seeks to determine which parameters in the model are misspecified. Such information is gleaned from program output that focuses on modification indices (MIs), estimates that derive from testing for the meaningfulness of all constrained (or fixed) parameters in the model. For example, the constraint that the loading of SDQ10 on Factor 1 is zero, as per Figure 1 would be tested. If the MI related to this fixed parameter is large, compared to all other MIs, then this finding would argue for its specification as a freely estimated parameter. In this case, the new parameter would represent a loading of SDQ10 on both Factor 1 and Factor 2. Of critical importance in *post hoc* model-fitting, however, is the requirement that only substantively meaningful parameters be added to the original model specification.

Interpreting Estimates

Shown in Figure 2 are standardized parameter estimates resulting from the testing of the hypothesized CFA model portrayed in Figure 1. Standardization transforms the solution so that all variables have a variance of 1; factor loadings will still be related in the same proportions as in the original solution, but parameters that were originally fixed will no longer have the same values. In a standardized solution, factor loadings should generally be less than 1.0 [14].

Turning first to the factor loadings and their associated errors of measurement, we see that, for example, the regression of Item SDQ15 on Factor 1 (Physical SC; Appearance) is .82. Because SDQ15 loads only on Factor 1, we can interpret this estimate as indicating that Factor 1 accounts for approximately 67% ($100 \times .82^2$) of the variance in this item. The measurement error coefficient associated with SDQ15 is .58, thereby indicating that some 34% (as a result of decimal rounding) of the variance associated with this item remains unexplained by Factor 1. (It is important to note that, unlike the LISREL program [13], which does not standardize errors in variables, the EQS program [2] used here does provide these standardized estimated values; *see Structural Equation Modeling: Software*.)

Finally, values associated with the double-headed arrows represent latent factor correlations. Thus, for example, the value of .41 represents the correlation between Factor 1 (Physical SC; Appearance) and Factor 2 (Physical SC; Ability). These factor correlations should be consistent with the theory within which the CFA model is grounded.

In conclusion, it is important to emphasize that only issues related to the specification of first-order CFA models, and only a cursory overview of the steps involved in testing these models has been included here. Indeed, sound application of SEM procedures in testing CFA models requires that researchers have a comprehensive understanding of the analytic process. Of particular importance are issues related to the assessment of multivariate normality, appropriateness of sample size, use of incomplete data, correction for nonnormality, model specification, identification, and estimation, evaluation of model fit, and post hoc model-fitting. Some of these topics are covered in other entries, as well as the books and journal articles cited herein.

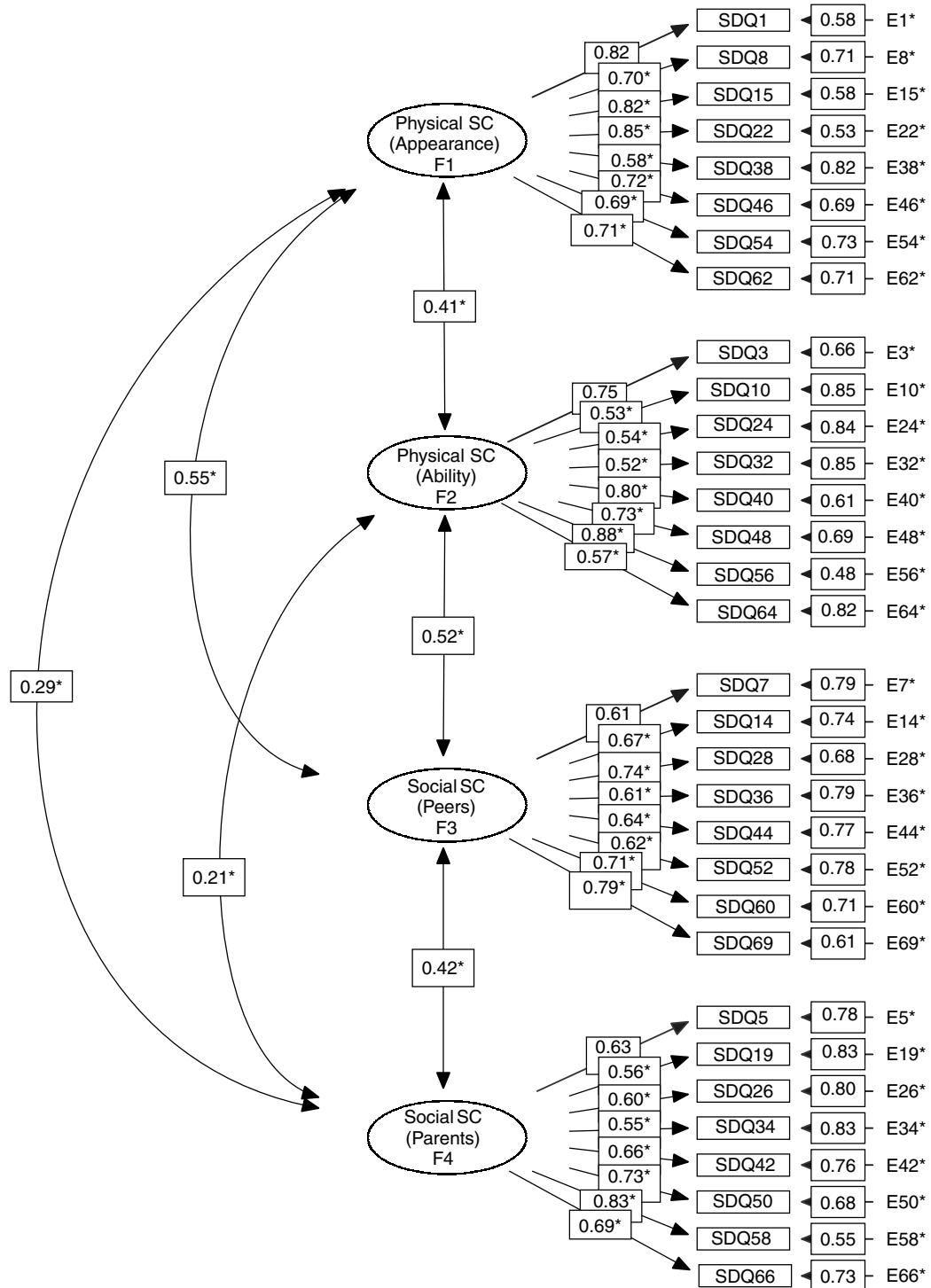


Figure 2 Standardized estimates for hypothesized CFA model

References

- [1] Bentler, P.M. (1990). Comparative fit indexes in structural models, *Psychological Bulletin* **107**, 238–246.
- [2] Bentler, P.M. (2004). *EQS 6.1: Structural Equations Program Manual*, Multivariate Software Inc, Encino.
- [3] Bollen, K. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [4] Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long eds, Sage, Newbury Park, pp. 136–162.
- [5] Bryant, F.B. & Yarnold, P.R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis, in *Reading and Understanding Multivariate Statistics*, L.G. Grimm & P.R. Yarnold eds, American Psychological Association, Washington.
- [6] Byrne, B.M. (1994). *Structural Equation Modeling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming*, Sage, Thousand Oaks.
- [7] Byrne, B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*, Erlbaum, Mahwah.
- [8] Byrne, B.M. (2001a). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*, Erlbaum, Mahwah.
- [9] Byrne, B.M. (2001b). Structural equation modeling with AMOS, EQS, and LISREL: comparative approaches to testing for the factorial validity of a measuring instrument, *International Journal of Testing* **1**, 55–86.
- [10] Curran, P.J., West, S.G. & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis, *Psychological Methods* **1**, 16–29.
- [11] Hu, L.-T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives, *Structural Equation Modeling* **6**, 1–55.
- [12] Hu, L.-T., Bentler, P.M. & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin* **112**, 351–362.
- [13] Jöreskog, K.G. & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*, Scientific Software International, Chicago.
- [14] Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, Guildwood Press, New York.
- [15] Marsh, H.W. (1992). *Self Description Questionnaire (SDQ) I: A Theoretical and Empirical Basis for the Measurement of Multiple Dimensions of Preadolescent Self-concept: A Test Manual and Research Monograph*, Faculty of Education, University of Western Sydney, Macarthur, New South Wales.
- [16] Maruyama, G.M. (1998). *Basics of Structural Equation Modeling*, Sage, Thousand Oaks.
- [17] Raykov, T. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*, Erlbaum, Mahwah.
- [18] Steiger, J.H. (1990). Structural model evaluation and modification: an interval estimation approach, *Multivariate Behavioral Research* **25**, 173–180.

(See also **History of Path Analysis; Linear Statistical Models for Causation: A Critical Review; Residuals in Structural Equation, Factor Analysis, and Path Analysis Models; Structural Equation Modeling: Checking Substantive Plausibility**)

BARBARA M. BYRNE

Factor Analysis: Exploratory

ROBERT PRUZEK

Volume 2, pp. 606–617

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Analysis: Exploratory

Introduction

This year marks the one hundredth anniversary for exploratory factor analysis (EFA), a method introduced by **Charles Spearman** in 1904 [21]. It is testimony to the deep insights of Spearman as well as many who followed that EFA continues to be central to **multivariate analysis** so many years after its introduction. In a recent search of electronic sources, where I restricted attention to the psychological and social sciences (using PsychINFO), more than 20 000 articles and books were identified in which the term ‘factor analysis’ had been used in the summary, well over a thousand citations from the last decade alone.

EFA, as it is known today, was for many years called common factor analysis. The method is in some respects similar to another well-known method called **principal component analysis** (PCA) and because of various similarities, these methods are frequently confused. One of the purposes of this article will be to try to dispel at least some of the confusion.

The general methodology currently seen as an umbrella for both exploratory factor analysis and confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) is called **structural equation modeling** (SEM) Although EFA can be described as an exploratory or unrestricted structural equation model, it would be a shame to categorize EFA as nothing more than a SEM, as doing so does an injustice to its long history as the most used and most studied latent variable method in the social and behavioral sciences. This is somewhat like saying that **analysis of variance** (ANOVA) which has been on the scene for more than seventy-five years and which is prominently related to experimental design, is just a **multiple linear regression** model. There is some truth to each statement, but it is unfair to the rich histories of EFA and ANOVA to portray their boundaries so narrowly.

A deeper point about the relationships between EFA and SEM is that these methods appeal to very different operational philosophies of science. While SEMs are standardly seen as founded on rather strict hypothetico-deductive logic, EFAs are

not. Rather, EFA generally invokes an exploratory search for structure that is open to new structures not imagined prior to analysis. Rozeboom [20] has carefully examined the logic of EFA, using the label *explanatory induction* to describe it; this term neatly summarizes EFA’s reliance on data to induce hypotheses about structure, and its general concern for explanation.

Several recent books, excellent reviews, and constructive critiques of EFA have become available to help understand its long history and its potential for effective use in modern times [6, 8, 15, 16, 23, 25]. A key aim of this article is to provide guidance with respect to literature about factor analysis, as well as to software to aid applications.

Basic Ideas of EFA Illustrated

Given a matrix of correlations or covariances (*see* **Correlation and Covariance Matrices**) among a set of manifest or observed variables, EFA entails a model whose aim is to explain or account for correlations using a smaller number of ‘underlying variables’ called common factors. EFA postulates common factors as **latent variables** so they are unobservable in principle. Spearman’s initial model, developed in the context of studying relations among psychological measurements, used a single common factor to account for all correlations among a battery of tests of intellectual ability. Starting in the 1930s, **Thurstone** generalized the ‘two-factor’ method of Spearman so that EFA became a multiple (common) factor method [22]. In so doing, Thurstone effectively broadened the range of prospective applications in science. The basic model for EFA today remains largely that of Thurstone. EFA entails an assumption that there exist uniqueness factors as well as common factors, and that these two kinds of factors complement one another in mutually orthogonal spaces. An example will help clarify the central ideas.

Table 1 below contains a correlation matrix for all pairs of five variables, the first four of which correspond to ratings by the seventeenth century art critic de Piles (using a 20 point scale) of 54 painters for whom data were complete [7]. Works of these painters were rated on four characteristics: composition, drawing, color, and expression. Moreover, each painter was associated with a particular ‘School.’ For current purposes, all information about Schools is

2 Factor Analysis: Exploratory

Table 1 Correlations among pairs of variables, painter data of [8]

	Composition	Drawing	Color	Expression	School D
Composition	1.00				
Drawing	0.42	1.00			
Color	-0.10	-0.52	1.00		
Expression	0.66	0.57	-0.20	1.00	
School D	-0.29	-0.36	0.53	-0.45	1.00

ignored except for distinguishing the most distinctive School D (Venetian) from the rest using a binary variable. For more details, see the file ‘painters’ in the Modern Applied Statistics with S (MASS) library in R or Splus software (see **Software for Statistical Analyses**), and note that the original data and several further analyses can be found in the MASS library [24].

Table 1 exhibits correlations among the painter variables, where upper triangle entries are ignored since the matrix is symmetric. Table 2 exhibits a common factor coefficients matrix (of order 5×2) that corresponds to the initial correlations, where entries of highest magnitude are in bold print. The final column of Table 2 is labeled h^2 , the standard notation for variable communalities. Because these factor coefficients correspond to an orthogonal factor solution, that is, uncorrelated common factors, each communality can be reproduced as a (row) sum of squares of the two factor coefficients to its left; for example $(0.76)^2 + (-0.09)^2 = 0.59$. The columns labeled 1 and 2 are factor loadings, each of which is properly interpreted as a (product-moment) correlation between one of the original manifest variables (rows) and a derived common factor (columns). Post-multiplying the factor coefficient matrix by its transpose yields numbers that

approximate the corresponding entries in the correlation matrix. For example, the *inner product* of the rows for Composition and Drawing is $0.76 \times 0.50 + (-0.09) \times (-0.56) = 0.43$, which is close to 0.42, the observed correlation; so the corresponding residual equals -0.01 . Pairwise products for all rows reproduce the observed correlations in Table 1 quite well as only one residual fit exceeds 0.05 in magnitude, and the mean residual is 0.01.

The final row of Table 2 contains the average sum of squares for the first two columns; the third entry is the average of the communalities in the final column, as well as the sum of the two average sums of squares to its left: $0.31 + 0.28 \approx 0.60$. These results demonstrate an additive decomposition of common variance in the solution matrix where 60 percent of the total variance is common among these five variables, and 40 percent is uniqueness variance.

Users of EFA have often confused communality with reliability, but these two concepts are quite distinct. Classical common factor and psychometric test theory entail the notion that the uniqueness is the sum of two (orthogonal) parts, specificity and error. Consequently, uniqueness variance is properly seen as an upper bound for error variance; alternatively, communality is in principle a lower bound for reliability. It might help to understand this by noting that each EFA entails analysis of just a sample of observed variables or measurements in some domain, and that the addition of more variables within the general domain will generally increase shared variance as well as individual communalities. As battery size is increased, individual communalities increase toward upper limits that are in principle close to variable reliabilities. See [15] for a more elaborate discussion.

To visualize results for my example, I plot the common factor coefficients in a plane, after making some modifications in signs for selected rows and the second column. Specifically, I reverse the signs of the 3rd and 5th rows, as well as in the second column, so that all values in the factor coefficients matrix become

Table 2 Factor loadings for 2-factor EFA solution, painter data

	Factor		
Variable name	1	2	h^2
Composition	0.76	-0.09	0.59
Drawing	0.50	-0.56	0.56
Color	-0.03	0.80	0.64
Expression	0.81	-0.26	0.72
School D	-0.30	0.62	0.47
Avg. Col. SS	0.31	0.28	0.60

positive. Changes of this sort are always permissible, but we need to keep track of the changes, in this case by renaming the third variable to 'Color[-1]' and the final binary variable to 'School.D[-1]'. Plotting the revised coefficients by rows yields the five labeled points of Figure 1.

In addition to plotting points, I have inserted vectors to correspond to 'transformed' factors; the arrows show an 'Expression-Composition' factor and a second, correlated, 'Drawing-Color[-1]' factor. That the School.D variable also loads highly on this second factor, and is also related to, that is, not orthogonal to, the point for Expression, shows that mean ratings, especially for the Drawing, Expression, and Color variates (the latter in an opposite direction), are notably different between Venetian School artists and painters from the collection of other schools. This can be verified by examination of the correlations (sometimes called point biserials) between the School.D variable and all the ratings variables in Table 1; the skeptical reader can easily acquire these data and study details. In fact, one of the reasons for choosing this example was to show that EFA as an exploratory data analytic method can help in studies of relations among quantitative and categorical variables. Some connections of EFA with other methods will be discussed briefly in the final section.

In modern applications of factor analysis, investigators ordinarily try to name factors in terms of dimensions of individual difference variation, to

identify latent variables that in some sense appear to 'underlie' observed variables. In this case, my ignorance of the works of these classical painters, not to mention of the thinking of de Piles as related to his ratings, led to my literal, noninventive factor names.

Before going on, it should be made explicit that insertion of the factor-vectors into this plot, and the attempt to name factors, are best regarded as discretionary parts of the EFA enterprise. The key output of such an analysis is the identification of the subspace defined by the common factors, within which variables can be seen to have certain distinctive structural relationships with one another. In other words, it is the configuration of points in the derived space that provides the key information for interpreting factor results; a relatively low-dimensional subspace provides insights into structure, as well as quantification of how much variance variables have in common. Positioning or naming of factors is generally optional, however common. When the common number of derived factors exceeds two or three, factor transformation is an almost indispensable part of an EFA, regardless of whether attempts are made to name factors.

Communalities generally provide information as to how much variance variables have in common or share, and can sometimes be indicative of how highly predictable variables are from one another. In fact, the squared multiple correlation of each variable with all others in the battery is often recommended as an initial estimate of communality for each variable. Communalities can also signal (un)reliability, depending on the composition of the battery of variables, and the number of factors; recall the foregoing discussion on this matter.

Note that there are no assumptions that point configurations for variables must have any particular form. In this sense, EFA is more general than many of its counterparts. Its exploratory nature also means that prior structural information is usually not part of an EFA, although this idea will eventually be qualified in the context of reviewing factor transformations. Even so, clusters or hierarchies of either variables or entities may sometimes be identified in EFA solutions. In our example, application of the common factor method yields a relatively parsimonious model in the sense that two common factors account for all relationships among variables. However, EFA was, and is usually, antiparsimonious in another sense as there is one uniqueness factor for each variable as

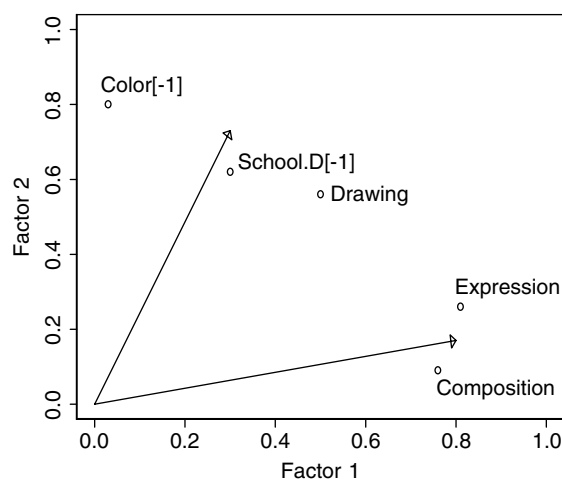


Figure 1 Plot of variables-as-points in 2-factor space, painter data

well as common factors to account for all entries in the correlation table.

Some Relationships Between EFA and PCA

As noted earlier, EFA is often confused with PCA. In fact, misunderstanding occurs so often in reports, published articles, and textbooks that it will be useful to describe how these methods compare, at least in a general way. More detailed or more technical discussions concerning such differences is available in [15].

As noted, the key aim of EFA is usually to derive a relatively small number of common factors to explain or account for (off-diagonal) covariances or correlations among a set of observed variables. However, despite being an exploratory method, EFA entails use of a falsifiable *model* at the level of manifest observations or correlations (covariances). For such a model to make sense, relationships among manifest variables should be approximately linear. When approximate linearity does not characterize relationships among variables, attempts can be made to transform (at least some of) the initial variables to ‘remove bends’ in their relationships with other variables, or perhaps to remove outliers. Use of square root, logarithmic, reciprocal, and other nonlinear transformations are often effective for such purposes. Some investigators question such steps, but rather than asking why nonlinear transformations should be considered, a better question usually is, ‘Why should the analyst believe the metric used at the outset for particular variables should be expected to render relationships linear, without reexpressions or transformations?’ Given at least approximate linearity among all pairs of variables – the inquiry about which is greatly facilitated by examining pairwise scatterplots among all pairs of variables – common factor analysis can often facilitate explorations of relationships among variables. The prospects for effective or productive applications of EFA are also dependent on thoughtful efforts at the stage of study design, a matter to be briefly examined below. With reference to our example, the pairwise relationships between the various pairs of de Pile’s ratings of painters were found to be approximately linear.

In contrast to EFA, principal components analysis does *not* engage a model. PCA generally entails

an algebraic decomposition of an initial data matrix into mutually orthogonal derived variables called *components*. Alternatively, PCA can be viewed as a linear transformation of the initial data vectors into uncorrelated variates with certain optimality properties. Data are usually centered at the outset by subtracting means for each variable and then scaled so that all variances are equal, after which the (rectangular) data matrix is resolved using a method called *singular value decomposition* (SVD). Components from a SVD are usually ordered so that the first component accounts for the largest amount of variance, the second the next largest amount, subject to the constraint that it be uncorrelated with the first, and so forth. The first few components will often summarize the majority of variation in the data, as these are principal components. When used in this way, PCA is justifiably called a *data reduction* method and it has often been successful in showing that a rather large number of variables can be summarized quite well using a relatively small number of derived components.

Conventional PCA can be completed by simply computing a table of correlations of each of the original variables with the chosen principal components; indeed doing so yields a PCA counterpart of the EFA coefficients matrix in Table 2 if two components are selected. Furthermore, sums of squares of correlations in this table, across variables, show the total variance each component explains. These component-level variances are the eigenvalues produced when the correlation matrix associated with the data matrix is resolved into eigenvalues and eigenvectors. Alternatively, given the original (centered and scaled) data matrix, and the eigenvalues and vectors of the associated correlation matrix, it is straightforward to compute principal components. As in EFA, derived PCA coefficient matrices can be rotated or transformed, and for purposes of interpretation this has become routine.

Given its algebraic nature, there is no particular reason for transforming variables at the outset so that their pairwise relationships are even approximately linear. This can be done, of course, but absent a model, or any particular justification for concentrating on pairwise *linear* relationships among variables, principal components analysis of correlation matrices is somewhat arbitrary. Because PCA is just an algebraic decomposition of data, it can be used for any kind of data; no constraints are made about the

dimensionality of the data matrix, no constraints on data values, and no constraints on how many components to use in analyses. These points imply that for PCA, assumptions are also optional regarding statistical distributions, either individually or collectively. Accordingly, PCA is a highly general method, with potential for use for a wide range of data types or forms. Given their basic form, PCA methods provide little guidance for answering model-based questions, such as those central to EFA. For example, PCA generally offers little support for assessing how many components ('factors') to generate, or try to interpret; nor is there assistance for choosing samples or extrapolating beyond extant data for purposes of statistical or psychometric generalization. The latter concerns are generally better dealt with using models, and EFA provides what in certain respects is one of the most general classes of models available.

To make certain other central points about PCA more concrete, I return to the correlation matrix for the painter data. I also conducted a PCA with two components (but to save space I do not present the table of 'loadings').

That is, I constructed the first two principal component variables, and found their correlations with the initial variables. A plot (not shown) of the principal component loadings analogous to that of Figure 1 shows the variables to be configured similarly, but all points are further from the origin. The row sums of squares of the component loadings matrix were 0.81, 0.64, 0.86, 0.83, and 0.63, values that correspond to communality estimates in the third column of the common factor matrix in Table 2. Across all five variables, PCA row sums of squares (which should not be called communalities) range from 14 to 37 percent larger than the h^2 entries in Table 2, an average of 27 percent; this means that component loadings are substantially larger in magnitude than their EFA counterparts, as will be true quite generally. For any data system, given the same number of components as common factors, component solutions yield row sums of squares that tend to be at least somewhat, and often markedly, larger than corresponding communalities.

In fact, these differences between characteristics of the PCA loadings and common factor loadings signify a broad point worthy of discussion. Given that principal components are themselves linear combinations of the original data vectors, each of the data variables tends to be part of the linear combination

with which it is correlated. The largest weights for each linear combination correspond to variables that most strongly define the corresponding linear combination, and so the corresponding correlations in the Principal Component (PC) loading matrix tend to be highest, and indeed to have spuriously high magnitudes. In other words, each PC coefficient in the matrix that constitutes the focal point for interpretation of results, tends to have a magnitude that is 'too large' because the corresponding variable is correlated partly with itself, the more so for variables that are largest parts of corresponding components. Also, this effect tends to be exacerbated when principal components are rotated. Contrastingly, common factors are latent variables, outside of the space of the data vectors, and common factor loadings are not similarly spurious. For example, EFA loadings in Table 2, being correlations of observed variables with latent variables, do not reflect self-correlations, as do their PCA counterparts.

The Central EFA Questions: How Many Factors? What Communalities?

Each application of EFA requires a decision about how many common factors to select. Since the common factor model is at best an approximation to the real situation, questions such as how many factors, or what communalities, are inevitably answered with some degree of uncertainty. Furthermore, particular features of given data can make formal fitting of an EFA model tenuous. My purpose here is to present EFA as a true exploratory method based on common factor principles with the understanding that formal 'fitting' of the EFA model is secondary to 'useful' results in applications; moreover, I accept that certain decisions made in contexts of real data analysis are inevitably somewhat arbitrary and that any given analysis will be incomplete. A wider perspective on relevant literature will be provided in the final section.

The history of EFA is replete with studies of how to select the number of factors; hundreds of both theoretical and empirical approaches have been suggested for the number of factors question, as this issue has been seen as basic for much of the past century. I shall summarize some of what I regard as the most enduring principles or methods, while trying to shed light on when particular methods are likely

6 Factor Analysis: Exploratory

to work effectively, and how the better methods can be attuned to reveal relevant features of extant data.

Suppose scores have been obtained on some number of correlated variables, say p , for n entities, perhaps persons. To entertain a factor analysis (EFA) for these variables generally means to undertake an exploratory structural analysis of linear relations among the p variables by analyzing a $p \times p$ covariance or correlation matrix. Standard outputs of such an analysis are a factor loading matrix for orthogonal or correlated common factors as well as communality estimates, and perhaps factor score estimates. All such results are conditioned on the number, m , of common factors selected for analysis. I shall assume that in deciding to use EFA, there is at least some doubt, *a priori*, as to how many factors to retain, so extant data will be the key basis for deciding on the number of factors. (I shall also presume that the data have been properly prepared for analysis, appropriate nonlinear transformations made, and so on, with the understanding that even outwardly small changes in the data can affect criteria bearing on the number of factors, and more.)

The reader who is even casually familiar with EFA is likely to have learned that one way to select the number of factors is to see how many **eigenvalues** (of the correlation matrix; recall PCA) exceed a certain criterion. Indeed, the ‘roots-greater-than-one’ rule has become a default in many programs. Alas, rules of this sort are generally too rigid to serve reliably for their intended purpose; they can lead either to overestimates or underestimates of the number of common factors. Far better than using any fixed cutoff is to understand certain key principles and then learn some elementary methods and strategies for choosing m . In some cases, however, two or more values of m may be warranted, in different solutions, to serve distinctive purposes for different EFAs of the same data.

A second thing even a nonspecialist may have learned is to employ a ‘scree’ plot (SP) to choose the number of factors in EFA. An SP entails plotting eigenvalues, ordered from largest to smallest, against their ordinal position, 1, 2, ..., and so on. Ordinarily, the SP is based on eigenvalues of a correlation matrix [5]. While the usual SP sometimes works reasonably well for choosing m , there is a mismatch between such a standard SP, and another relevant fact: a tacit assumption of this method is that all p communalities are the same. But to assume equal

communalities is usually to make a rather strong assumption, one quite possibly not supported by data in hand.

A better idea for SP entails computing the original correlation matrix, R , as well as its inverse R^{-1} . Then, denoting the *diagonal* of the inverse as D^2 (entries of which exceed unity), rescale the initial correlation matrix to DRD , and then compute eigenvalues of this rescaled correlation matrix. Since the largest entries in D^2 correspond to variables that are most predictable from all others, and vice versa, the effect is to weigh variables more if they are more predictable, less if they are less predictable from other variables in the battery. (The complement of the reciprocal of any D^2 entry is in fact the squared multiple correlation (SMC) of that variable with all others in the set.) An SP based on eigenvalues of DRD allows for variability of communalities, and is usually realistic in assuming that communalities are at least roughly proportional to SMC values.

Figure 2 provides illustrations of two scree plots based on DRD , as applied to *two simulated* random samples. Although real data were used as the starting point for each simulation, both samples are just simulation sets of (the same size as) the original data set, where four factors had consistently been identified as the ‘best’ number to interpret.

Each of these two samples yields a scree plot, and both are given in Figure 2 to provide some sense of sampling variation inherent in such data; in this case, each plot leads to breaks after four common factors – where the break is found by reading the plot from right to left. But the slope between four and five

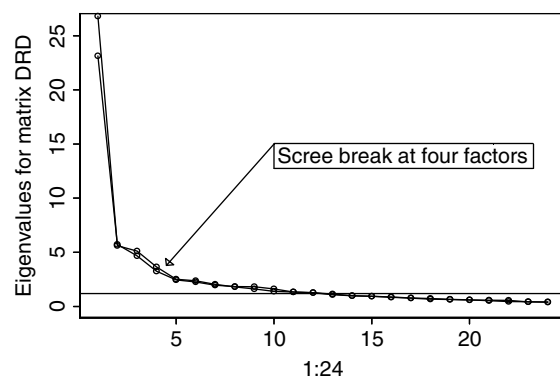


Figure 2 Two scree plots, for two simulated data sets, each $n = 145$

factors is somewhat greater for one sample than the other, so one sample identifies m as four with slightly more clarity than the other. In fact, for some other samples examined in preparing these scree plots, breaks came after three or five factors, not just four. Note that for smaller samples greater variation can be expected in the eigenvalues, and hence the scree breaks will generally be less reliable indicators of the number of common factors for smaller samples.

So what is the principle behind the scree method? The answer is that the *variance* of the $p - m$ smallest eigenvalues is closely related to the *variance* of residual correlations associated with fitting off-diagonals of the observed correlation matrix in successive choices for m , the number of common factors. When a break occurs in the eigenvalue plot, it signifies a notable drop in the sum of squares of residual correlations after fitting the common factor model to the observed correlation matrix for a particular value of m . I have constructed a horizontal line in Figure 2 to correspond to the *mean* of the 20 smallest eigenvalues (24–4) of DRD, to help see the variation these so-called ‘*rejected*’ eigenvalues have around their mean. In general, it is the variation around such a mean of rejected eigenvalues that one seeks to reduce to a ‘reasonable’ level when choosing m in the EFA solution, since a ‘good’ EFA solution accounts well for the off-diagonals of the correlation matrix. Methods such as bootstrapping – wherein multiple versions of DRD are generated over a series of bootstrap samples of the original data matrix – can be used to get a clearer sense of sampling variation, and probably should become part of standard practice in EFA both at the level of selecting the number of factors, and assessing variation in various derived EFA results.

When covariances or correlations are well fit by some relatively small number of common factors, then scree plots often provide flexible, informative, and quite possibly persuasive evidence about the number of common factors. However, SPs alone can be misleading, and further examination of data may be helpful. The issue in selecting m vis-à-vis the SP concerns the nature or reliability of the information in **eigenvectors** associated with corresponding eigenvalues. Suppose some number m^* is seen as a possible underestimate for m ; then deciding to add one more factor to have $m^* + 1$ factors, is to decide that the additional eigenvector adds useful or meaningful structural information to the EFA solution. It is

possible that m^* is an ‘underestimate’ solely because a single correlation coefficient is poorly fit, and that adding a common factor merely reduces a single ‘large’ residual correlation. But especially if the use of $m^* + 1$ factors yields a factor loading matrix that upon rotation (see below) improves interpretability in general, there may be ex post facto evidence that m^* was indeed an underestimate. Similar reasoning may be applied when moving to $m^* + 2$ factors, etc. Note that sampling variation can also result in sample reordering of so-called *population eigenvectors* too.

An adjunct to an SP that is too rarely used is simply to plot the distribution of the residual correlations, either as a histogram, or in relation to the original correlations, for, say, m , $m + 1$, and $m + 2$ factors in the vicinity of the scree break; outliers or other anomalies in such plots can provide evidence that goes usefully beyond the SP when selecting m . Factor transformation(s) (see below) may be essential to one’s final decision. Recall that it may be a folly even to think there is a single ‘correct’ value for m for some data sets.

Were one to use a different selection of variables to compose the data matrix for analysis, or perhaps make changes in the sample (deleting or adding cases), or try various different factoring algorithms, further modifications may be expected about the number of common factors. Finally, there is always the possibility that there are simply too many distinctive dimensions of individual difference variation, that is, common factors, for the EFA method to work effectively in some situations. It is not unusual that more variables, larger samples, or generally more investigative effort, are required to resolve some basic questions such as how many factors to use in analysis.

Given some choice for m , the next decision is usually that of deciding what factoring method to use. The foregoing idea of computing DRD, finding its eigenvalues, and producing an SP based on those, can be linked directly to an EFA method called *image factor analysis* (IFA) [13], which has probably been underused, in that several studies have found it to be a generally sound and effective method. IFA is a noniterative method that produces common factor coefficients and communalities directly. IFA is based on the m largest eigenvalues, say, the diagonal entries of Γ_m , and corresponding eigenvectors, say Q_m , of the matrix denoted DRD, above. Given a particular factor method, communality estimates follow directly from selection of the number of common factors.

The analysis usually commences from a correlation matrix, so communality estimates are simply row sums of squares of the (orthogonal) factor coefficients matrix that for m common factors is computed as $\Lambda_m = D^{-1}Q_m(\Gamma_m - \phi I)^{1/2}$, where ϕ is the average of the $p - m$ smallest eigenvalues. IFA may be especially defensible for EFA when sample size is limited; more details are provided in [17], including a sensible way to modify the diagonal D^2 when the number of variables is a ‘substantial fraction’ of sample size.

A more commonly used EFA method is called *maximum likelihood factor analysis* (MLFA) for which algorithms and software are readily available, and generally well understood. The theory for this method has been studied perhaps more than any other and it tends to work effectively when the EFA problem has been well-defined and the data are ‘well-behaved.’ Specialists regularly advocate use of the MLFA method [1, 2, 16, 23], and it is often seen as the common factor method of choice when the sample is relatively large. Still, MLFA is an iterative method that can lead to poor solutions, so one must be alert in case it fails in some way. Maximum likelihood EFA methods generally call for large n ’s, using an assumption that the sample has been drawn randomly from a parent population for which multivariate normality (see **Catalogue of Probability Density Functions**) holds, at least approximately; when this assumption is violated seriously, or when sample size is not ‘large,’ MLFA may not serve its exploratory purpose well. Statistical tests may sometimes be helpful, but the sample size issue is vital if EFA is used for testing statistical hypotheses. There can be a mismatch between exploratory use of EFA and statistical testing because small samples may not be sufficiently informative to reject any factor model, while large samples may lead to rejection of every model in some domains of application. Generally scree methods for choosing the number of factors are superior to statistical testing procedures.

Given a choice of factoring methods – and of course there are many algorithms in addition to IFA and MLFA – the generation of communality estimates follows directly from the choice of m , the number of common factors. However, some EFA methods or algorithms can yield numerically unstable results, particularly if m is a substantial fraction of p , the number of variables, or when n is not large in relation to p . Choice of factor methods, like many

other methodological decisions, is often best made in consultation with an expert.

Factor Transformations to Support EFA Interpretation

Given at least a tentative choice for m , EFA methods such as IFA or MLFA can be used straightforwardly to produce matrices of factor coefficients to account for structural relations among variables. However, attempts to interpret factor coefficient matrices without further efforts to transform factors usually fall short unless $m = 1$ or 2 , as in our illustration. For larger values of m , factor transformation can bring order out of apparent chaos, with the understanding that order can take many forms. Factor transformation algorithms normally take one of three forms: Procrustes fitting to a prespecified target (see **Procrustes Analysis**), orthogonal simple structure, or oblique simple structure. All modern methods entail use of specialized algorithms. I shall begin with Procrustean methods and review each class of methods briefly.

Procrustean methods owe their name to a figure of ancient Greek mythology, Procrustes, who made a practice of robbing highway travelers, tying them up, and stretching them, or cutting off their feet to make them fit a rigid iron bed. In the context of EFA, Procrustes methods are more benign; they merely invite the investigator to prespecify his or her beliefs about structural relations among variables in the form of a target matrix, and then transform an initial factor coefficients matrix to put it in relatively close conformance with the target. Prespecification of configurations of points in m -space, preferably on the basis of hypothesized structural relations that are meaningful to the investigator, is a wise step for most EFAs even if Procrustes methods are not to be used explicitly for transformations. This is because explication of beliefs about structures can afford (one or more) reference system(s) for interpretation of empirical data structures however they were initially derived. It is a long-respected principle that prior information, specified independently of extant empirical data, generally helps to support scientific interpretations of many kinds, and EFA should be no exception. In recent times, however, methods such as confirmatory factor analysis (CFA), are usually seen as making Procrustean EFA methods obsolete because CFA methods offer generally more

sophisticated numerical and statistical machinery to aid analyses. Still, as a matter of principle, it is useful to recognize that general methodology of EFA has for over sixty years permitted, and in some respects encouraged, incorporation of sharp prior questions in structural analyses.

Orthogonal rotation algorithms provide relatively simple ways for transforming factors and these have been available for nearly forty years. Most commonly, an ‘orthomax’ criterion is optimized, using methods that have been dubbed ‘quartimax’, ‘varimax’, or ‘equamax.’ Dispensing with quotations, we merely note that in general, equamax solutions tend to produce simple structure solutions for which different factors account for nearly equal amounts of common variance; quartimax, contrastingly, typically generates one broad or general factor followed by $m - 1$ ‘smaller’ ones; varimax produces results intermediate between these extremes. The last, varimax, is the most used of the orthogonal simple structure rotations, but choice of a solution should not be based too strongly on generic popularity, as particular features of a data set can make other methods more effective. Orthogonal solutions offer the appealing feature that squared common factor coefficients show directly how much of each variable’s common variance is associated with each factor. This property is lost when factors are transformed obliquely. Also, the factor coefficients matrix alone is sufficient to interpret orthogonal factors; not so when derived factors are mutually correlated. Still, forcing factors to be uncorrelated can be a weakness when the constraint of orthogonality limits factor coefficient configurations unrealistically, and this is a common occurrence when several factors are under study.

Oblique transformation methods allow factors to be mutually correlated. For this reason, they are more complex and have a more complicated history. A problem for many years was that by allowing factors to be correlated, oblique transformation methods often allowed the m -factor space to collapse; successful methods avoided this unsatisfactory situation while tending to work well for wide varieties of data. While no methods are entirely acceptable by these standards, several, notably those of Jennrich and Sampson (direct quartimin) [12], Harris and Kaiser (obliquimax), Rozeboom (Hyball) [18], Yates (geomin) [25], and Hendrikson and White (promax) [9] are especially worthy of consideration for

applications. Browne [2], in a recent overview of analytic rotation methods for EFA, stated that Jennrich and Sampson [12] ‘solved’ the problems of oblique rotation; however, he went on to note that ‘... we are not at a point where we can rely on mechanical exploratory rotation by a computer program if the complexity of most variables is not close to one [2, p. 145].’ Methods such as Hyball [19] facilitate random starting positions in m -space of transformation algorithms to produce multiple solutions that can then be compared for interpretability. The promax method is notable not only because it often works well, but also because it combines elements of Procrustean logic with analytical orthogonal transformations. Yates’ geomin [25] is also a particularly attractive method in that the author went back to Thurstone’s basic ideas for achieving simple structure and developed ways for them to be played out in modern EFA applications. A special reason to favor simple structure transformations is provided in [10, 11] where the author noted that standard errors of factor loadings will often be substantially smaller when population structures are simple than when they are not; of course this calls attention to the design of the battery of variables.

Estimation of Factor Scores

It was noted earlier that latent variables, that is, common factors, are basic to any EFA model. A strong distinction is made between observable variables and the underlying latent variables seen in EFA as accounting for manifest correlations or covariances between all pairs of manifest variables. The latent variables are by definition never observed or observable in a real data analysis, and this is not related to the fact that we ordinarily see our data as a sample (of cases, or rows); latent variables are in principle not observable, either for statistically defined samples, or for their population counterparts. Nevertheless, it is not difficult to *estimate* the postulated latent variables, using linear combinations of the observed data. Indeed, many different kinds of factor score estimates have been devised over the years (*see Factor Score Estimation*).

Most methods for estimating factor scores are not worth mentioning because of one or another kind of technical weakness. But there are two methods that are worthy of consideration for practical applications in EFA where factor score estimates seem

needed. These are called *regression estimates* and *Bartlett* (also, *maximum likelihood*) *estimates* of factor scores, and both are easily computed in the context of IFA. Recalling that D^2 was defined as the diagonal of the inverse of the correlation matrix, now suppose the initial data matrix has been centered and scaled as Z where $Z'Z = R$; then, using the notation given earlier in the discussion of IFA, Bartlett estimates of factor scores can be computed as $X_{m-\text{Bartlett}} = Z D Q_m (\Gamma_m - \phi I)^{-1/2}$. The discerning reader may recognize that these factor scores estimates can be further simplified using the singular value decomposition of matrix $Z D$; indeed, these score estimates are just rescaled versions of the first m principal components of $Z D$. Regression estimates, in turn, are further column rescalings of the same m columns in $X_{m-\text{Bartlett}}$. MLFA factor score estimates are easily computed, but to discuss them goes beyond our scope; see [15]. Rotated or transformed versions of factor score estimates are also not complicated; the reader can go to **factor score estimation (FSE)** for details.

EFA in Practice: Some Guidelines and Resources

Software packages such as CEFA [3], which implements MLFA as well as geomin among other methods, and Hyball [18], can be downloaded from the web without cost, and they facilitate use of most of the methods for factor extraction as well as factor transformation. These packages are based on modern methods, they are comprehensive, and they tend to offer advantages that most commercial software for EFA do not. What these methods lack, to some extent, is mechanisms to facilitate modern graphical displays. Splus and R software, the latter of which is also freely available from the web [r-project.org], provide excellent modern graphical methods as well as a number of functions to implement many of the methods available in CEFA, and several in Hyball. A small function for IFA is provided at the end of this article; it works in both R and Splus. In general, however, no one source provides all methods, mechanisms, and management capabilities for a fully operational EFA system – nor should this be expected since what one specialist means by ‘fully operational’ necessarily differs from that of others.

Nearly all real-life applications of EFA require decisions bearing on how and how many cases are

selected, how variables are to be selected and transformed to help ensure approximate linearity between variates; next, choices about factoring algorithms or methods, the number(s) of common factors and factor transformation methods must be made. That there be no notably weak links in this chain is important if an EFA project is to be most informative. Virtually all questions are contextually bound, but the literature of EFA can provide guidance at every step.

Major references on EFA application, such as that of Carroll [4], point up many of the possibilities and a perspective on related issues. Carroll suggests that special value can come from side-by-side analyses of the same data using EFA methods and those based on structural equation modeling (SEM). McDonald [15] discusses EFA methods in relation to SEM. Several authors have made connections between EFA and other multivariate methods such as basic regression; see [14, 17] for examples.

– – an S function for Image Factor Analysis – –

```
'ifa'<-function(rr,mm) {
# routine is based on image factor
# analysis;
# it generates an unrotated common
# factor coefficients matrix & a scree
# plot; in R, follow w/ promax or
# varimax; in Splus follow w/ rotate.
# rr is taken to be symmetric matrix
# of correlations or covariances;
# mm is no. of factors. For additional
# functions or assistance, contact:
# rpruzek@uamail.albany.edu
  rinv <- solve(rr) #takes inverse
  # of R; so R must be nonsingular
  sm2i <- diag(rinv)
  smrt <- sqrt(sm2i)
  dsmrt <- diag(smrt)
  rsr <- dsmrt %*% rr %*% dsmrt
  reig <- eigen(rsr, sym = T)
  vlamd <- reig$va
  vlamdm <- vlamd[1:mm]
  qqm <- as.matrix(reig$ve[, 1:mm])
  theta <- mean(vlamd[(mm + 1)
:nrow(qqm)])
  dg <- sqrt(vlamdm - theta)
  if(mm == 1)
    fac <- dg[1] * diag(1/smrt)
    %*% qqm
  else fac <- diag(1/smrt) %*% qqm
    %*% diag(dg)
```

```

plot(1:nrow(rr), vlamd, type
= "o")
abline(h = theta, lty = 3)
title("Scree plot for IFA")
print("Common factor coefficients
matrix is: fac")
print(fac)
list(vlamd = vlamd, theta = theta,
fac = fac)
}

```

References

- [1] Browne, M.W. (1968). A comparison of factor analytic techniques, *Psychometrika* **33**, 267–334.
- [2] Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis, *Multivariate Behavioral Research* **36**, 111–150.
- [3] Browne, M.W., Cudeck, R., Tateneni, K. & Mels, G. (1998). CEFA: Comprehensive Exploratory Factor Analysis (computer software and manual). [<http://quantrm2.psy.ohio-state.edu/browne/>]
- [4] Carroll, J.B. (1993). *Human Cognitive Abilities: A Survey of Factor Analytic Studies*, Cambridge University Press, New York.
- [5] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.
- [6] Darlington, R. (2000). Factor Analysis (Instructional Essay on Factor Analysis). [<http://comp9.psych.cornell.edu/Darlington/factor.htm>]
- [7] Davenport, M. & Studdert-Kennedy, G. (1972). The statistical analysis of aesthetic judgement: an exploration, *Applied Statistics* **21**, 324–333.
- [8] Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research, *Psychological Methods* **3**, 272–299.
- [9] Hendrickson, A.E. & White, P.O. (1964). PROMAX: a quick method for transformation to simple structure, *Brit. Jour. of Statistical Psychology* **17**, 65–70.
- [10] Jennrich, R.I. (1973). Standard errors for obliquely rotated factor loadings, *Psychometrika* **38**, 593–604.
- [11] Jennrich, R.I. (1974). On the stability of rotated factor loadings: the Wexler phenomenon, *Brit. J. Math. Stat. Psychology* **26**, 167–176.
- [12] Jennrich, R.I. & Sampson, P.F. (1966). Rotation for simple loadings, *Psychometrika* **31**, 313–323.
- [13] Jöreskog, K.G. (1969). Efficient estimation in image factor analysis, *Psychometrika* **34**, 51–75.
- [14] Lawley, D.N. & Maxwell, A.E. (1973). Regression and factor analysis, *Biometrika* **60**, 331–338.
- [15] McDonald, R.P. (1984). *Factor Analysis and Related Methods*, Lawrence Erlbaum Associates, Hillsdale.
- [16] Preacher, K.J. & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine, *Understanding Statistics* **2**, 13–43. [<http://www.geocities.com/Athens/Acropolis/8950/tomswift.pdf>]
- [17] Pruzek, R.M. & Lepak, G.M. (1992). Weighted structural regression: a broad class of adaptive methods for improving linear prediction, *Multivariate Behavioral Research* **27**, 95–130.
- [18] Rozeboom, W.W. (1991). HYBALL: a method for subspace-constrained oblique factor rotation, *Multivariate Behavioral Research* **26**, 163–177. [<http://web.psych.ualberta.ca/~rozeboom/>]
- [19] Rozeboom, W.W. (1992). The glory of suboptimal factor rotation: why local minima in analytic optimization of simple structure are more blessing than curse, *Multivariate Behavioral Research* **27**, 585–599.
- [20] Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive, in *What if there were no significance tests?*, Chapter 13, L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum Associates, Hillsdale, NJ.
- [21] Spearman, C. (1904). General intelligence objectively determined and measured, *American Jour. of Psychology* **15**, 201–293.
- [22] Thurstone, L.L. (1947). *Multiple-factor Analysis: A Development and Expansion of the Vectors of Mind*, University of Chicago Press, Chicago.
- [23] Tucker, L. & MacCallum, R.C. (1997). *Exploratory factor analysis*. [unpublished, but available: <http://www.unc.edu/~rcm/book/factornew.htm>]
- [24] Venables, W.N. & Ripley, B.D. (2002). *Modern Applied Statistics with S*, Springer, New York.
- [25] Yates, A. (1987). *Multivariate Exploratory Data Analysis: A Perspective on Exploratory Factor Analysis*, State University of New York Press, Albany.

ROBERT PRUZEK

Factor Analysis: Multiple Groups

TODD D. LITTLE AND DAVID W. SLEGGERS

Volume 2, pp. 617–623

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Analysis: Multiple Groups

Factor Analysis: Multiple Groups with Means

The confirmatory factor analysis (CFA) (*see* **Factor Analysis: Confirmatory**) model is a very effective approach to modeling multivariate relationships across multiple groups. The CFA approach to factorial invariance has its antecedents in exploratory factor analysis. **Cattell** [4] developed a set of principles by which to judge the rotated solutions from two populations with the goal being simultaneous simple structure. Further advancements were made by Horst & Schaie [7] and culminated with work by Meredith [13] in which he gave methods for rotating solutions from different populations to achieve one best fitting factor pattern. Confirmatory factor analytic techniques have made exploratory methods of testing for invariance obsolete by allowing an exact structure to be hypothesized. The multiple-group CFA approach is particularly useful when making cross-group comparisons because it allows for (a) simultaneous estimation of all parameters (including mean-level information) for all groups and (b) direct statistical comparisons of the estimated parameters across the groups. The theoretical basis for selecting groups can vary from nominal variables such as gender, race, clinical treatment group, or nationality to continuous variables that can be easily categorized such as age-group or grade level. When making comparisons across distinct groups, however, it is critical to determine that the constructs of interest have the same meaning in each group (i.e., they are said to be measurement equivalent, or have strong factorial invariance; see below). This condition is necessary in order to make meaningful comparisons across groups [1].

In order to determine measurement equivalence, the analyses should go beyond the standard covariance structures information of the traditional CFA model to also include the mean structure information [9, 14, 16, 21]. We refer to such integrated modeling as mean and covariance structure (MACS) modeling. MACS analyses are well suited to establish construct comparability (i.e., factorial

invariance) and, at the same time, detect possible between-group differences because they allow: (a) simultaneous model fitting of an hypothesized factorial structure in two or more groups (i.e., the expected pattern of indicator-to-construct relations *for both the intercepts and factor loadings*), (b) tests of the cross-group equivalence of *both intercepts and loadings*, (c) corrections for measurement error whereby estimates of the latent constructs' means and covariances are disattenuated (i.e., estimated as true and reliable values), and (d) strong tests of substantive hypotheses about possible cross-group differences on the constructs [11, 14].

The General Factor Model. To understand the logic and steps involved in multiple-group MACS modeling, we begin with the matrix algebra notations for the general factor model, which, for multiple populations $g = 1, 2, \dots, G$, is represented by:

$$X_g = \tau_g + \Lambda_g \xi_g + \delta_g \quad (1)$$

$$E(X_g) = \mu_{xg} = \tau_g + \Lambda_g \kappa_g \quad (2)$$

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g \quad (3)$$

where x is a vector of observed or manifest indicators, ξ is a vector of latent constructs, τ is a vector of intercepts of the manifest indicators, Λ is the factor pattern or loading matrix of the indicators, κ represents the means of the latent constructs, Φ is the variance-covariance matrix of the latent constructs, and Θ is a symmetric matrix with the variances of the error term, δ , along the diagonal and possible covariances among the residuals in the off diagonal. All of the parameter matrices are subscripted with a g to indicate that the parameters may take different values in each population. For the common factor model, we assume that the indicators (i.e., items, parcels, scales, responses, etc.) are continuous variables that are multivariate normal (*see Catalogue of Probability Density Functions*) in the population and the elements of Θ have a mean of zero and are independent of the estimated elements in the other parameter matrices.

In a MACS framework, there are six types of parameter estimate that can be evaluated for equivalence across groups. The first three components refer to the measurement level: (a) Λ , the unstandardized regression coefficients of the indicators on the latent constructs (the loadings of the indicators), (b) τ , the

2 Factor Analysis: Multiple Groups

intercepts or means of the indicators, and (c) Θ , the residual variances of each indicator, which is the aggregate of the unique factor variance and the unreliable variance of an indicator. The other three types of parameter refer to the latent construct level: (d) κ , the mean of the latent constructs, (e) ϕ_{ii} latent variances, and (f) ϕ_{ij} latent covariances or correlations [9, 12, 14].

Taxonomy of Invariance

A key aspect of multiple-group MACS modeling is the ability to assess the degree of factorial invariance of the constructs across groups. Factorial invariance addresses whether the constructs' measurement properties (i.e., the intercepts and loadings, which reflect the reliable components of the measurement space) are the same in two or more populations. This question is distinct from whether the latent aspects of the constructs are the same (e.g., the constructs' mean levels or covariances). This latter question deals with particular substantive hypotheses about possible group differences on the constructs (i.e., the reliable and true properties of the constructs). The concept of invariance is typically thought of and described as a hierarchical sequence of invariance starting with the weakest form and working up to the strictest form. Although we will often discuss the modeling procedures in terms of two groups, the extension to three or more groups is straightforward (see e.g., [9]).

Configural Invariance. The most basic form of factorial invariance is ensuring that the groups have the same basic factor structure. The groups should have the same number of latent constructs, the same number of manifest indicators, and the same pattern of fixed and freed (i.e., estimated) parameters. If these conditions are met, the groups are said to have configural invariance. As the weakest form of invariance, configural invariance only requires the same pattern of fixed and freed estimates among the manifest and latent variables, but does not require the coefficients be equal across groups.

Weak Factorial Invariance. Although termed 'weak factorial invariance', this level of invariance is more restricted than configural invariance. Specifically, in addition to the requirement of having the same pattern of fixed and freed parameters across groups, the

loadings are equated across groups. The manifest means and residual variances are free to vary. This condition is also referred to as pattern invariance [15] or metric invariance [6]. Because the factor variances are free to vary across groups, the factor loadings are, technically speaking, proportionally equivalent (i.e., weighted by the differences in latent variances). If weak factorial invariance is found to be untenable (see 'testing' below) then only configural invariance holds across groups. Under this condition, one has little basis to suppose that the constructs are the same in each group and systematic comparisons of the constructs would be difficult to justify. If invariance of the loadings holds then one has a weak empirical basis to consider the constructs to be equivalent and would allow cross-group comparisons of the latent variances and covariances, but not the latent means.

Strong Factorial Invariance. As Meredith [14] compellingly argued, any test of factorial invariance *should* include the manifest means – weak factorial invariance is not a complete test of invariance. With strong factorial invariance, the loadings and the intercepts are equated (and like the variances of the constructs, the latent means are allowed to vary in the second and all subsequent groups). This strong form of factorial invariance, also referred to as scalar invariance [22], is required in order for individuals with the same ability in separate groups to have the same score on the instrument. With any less stringent condition, two individuals with the same true level of ability would not have the same expected value on the measure. This circumstance would be problematic because, for example, when comparing groups based on gender on a measure of mathematical ability one would want to ensure that a male and a female with the same level of ability would receive the same score.

An important advantage of strong factorial invariance is that it establishes the measurement equivalence (or construct comparability) of the measures. In this case, constructs are defined in precisely the same operational manner in each group; as a result, they can be compared meaningfully and with quantitative precision. Measurement equivalence indicates that (a) the constructs are generalizable entities in each subpopulation, (b) sources of bias and error (e.g., cultural bias, translation errors, varying conditions of administration) are minimal, (c) subgroup differences

have not differentially affected the constructs underlying measurement characteristics (i.e., constructs are comparable because the indicators' specific variances are independent of cultural influences after conditioning on the construct-defining common variance; [14]), and (d) between-group differences in the constructs' mean, variance, and covariance relations are quantitative in nature (i.e., the nature of group differences can be assessed as mean-level, variance, and covariance or correlational effects) at the construct level. In other words, with strong factorial invariance, the broadest spectrum of hypotheses about the primary construct moments (means, variances, covariances, correlations) can be tested while simultaneously establishing measurement equivalence (i.e., two constructs can demonstrate different latent relations across subgroups, yet still be defined equivalently at the measurement level).

Strict Factorial Invariance. With strict factorial invariance, all conditions are the same as for strong invariance but, in addition, the residual variances are equated across groups. This level of invariance is not required for making veridical cross-group comparisons because the residuals are where the aggregate of the true measurement error variance and the indicator-specific variance is represented. Here, the factors that influence unreliability are not typically expected to operate in an equivalent manner across the subgroups of interest. In addition, the residuals reflect the unique factors of the measured indicators (i.e., variance that is reliable but unique to the particular indicator). If the unique factors differ trivially with regard to subgroup influences, this violation of selection theorem [14] can be effectively tolerated, if sufficiently small, by allowing the residuals to vary across the subgroups. In other words, strong factorial invariance is less biasing than strict factorial invariance because, even though the degree of random error may be quite similar across groups, if it is not exactly equal, the nonequal portions of the random error are forced into other parameters of a given model, thereby introducing potential sources of bias. Moreover, in practical applications of cross-group research such as cross-cultural studies, some systematic bias (e.g., translation bias) may influence the reliable component of a given residual. Assuming these sources of bias and error are negligible (see 'testing' below), they could be represented as unconstrained residual variance terms across groups in order to examine

the theoretically meaningful common-variance components as unbiasedly as possible.

Partial Invariance. Widaman and Reise [23] and others have also introduced the concept of partial invariance, which is the condition when a constraint of invariance is not warranted for one or a few of the loading parameters. When invariance is untenable, one may then attempt to determine which indicators contribute significantly to the misfit ([3] [5]). It is likely that only a few of the indicators deviate significantly across groups, giving rise to the condition known as partial invariance. When partial invariance is discovered there are a variety of ways to proceed. (a) One can leave the estimate in the model, but not constrain it to be invariant across groups and argue that the invariant indicators are sufficient to establish comparability of the constructs [23]; (b) one can argue that the differences between indicators are small enough that they would not make a substantive difference and proceed with invariance constraints in place [9]; (c) one could decide to reduce the number of indicators by only using indicators that are invariant across groups [16]; (d) one could conclude that because invariance cannot be attained that the instrument must be measuring different constructs across the multiple groups and, therefore, not use the instrument at all [16]. Milsap and Kwok [16] also describe a method to assess the severity of the violations of invariance by evaluating the sensitivity and specificity at various selection points.

Selection Theorem Basis for Expecting Invariance. The loadings and intercepts of a constructs indicators can be expected to be invariant across groups under a basic tenet of selection theorem – namely, conditional independence ([8, 14]; see also [18]). In particular, if subpopulation influences (i.e., the basis for selecting the groups) and the *specific* components (unique factors) of the construct's manifest indicators are independent when conditioned on the common construct components, then an invariant measurement space can be specified even under extreme selection conditions. When conditional independence between the indicators' unique factors and the selection basis hold, the construct information (i.e., common variance) contains, or carries, information about subpopulation influences. This expectation is quite reasonable if one assumes that the subpopulations derive from a common population from which the subpopulations can be described as 'selected' on the basis of one or

more criteria (e.g., experimental treatment, economic affluence, degree of industrialization, degree of individualism etc.). This expectation is also reasonable if one assumes on the basis of a specific theoretical view that the constructs should exist in each assessed subpopulation and that the constructs' indicators reflect generally equivalent domain representations.

Because manifest indicators reflect both common and specific sources of variance, cross-group effects may influence not only the common construct-related variance of a set of indicators but also the specific variance of one or more of them [17]. Measurement equivalence will hold if these effects have influenced only the common-variance components of a set of construct indicators and not their unique-specific components [8, 14, 18]. If cross-group influences differentially and strongly affect the specific components of indicators, nonequivalence would emerge. Although measurement nonequivalence can be a meaningful analytic outcome, it disallows, when sufficiently strong, quantitative construct comparisons.

Identification Constraints

There are three methods of placing constraints on the model parameters in order to identify the constructs and model (*see Identification*). When a mean structure is used, the location must be identified in addition to the scale of the other estimated parameters.

The first method to identification and scale setting is to fix a parameter in the latent model. For example, to set the scale for the location parameters, one can fix the latent factor mean, κ , to zero (or a nonzero value). Similarly, to set the scale for the variance-covariance and loading parameters one can fix the variances, ϕ_{ii} to 1.0 (or any other nonzero value). The advantages of this approach are that the estimated latent means in each subsequent group are relative mean differences from the first group. Because this first group is fixed at zero, the significance of the latent mean estimates in the subsequent groups is the significance of the difference from the first group. Fixing the latent variances to 1.0 has the advantage of providing estimates of the associations among the latent constructs in correlational metric as opposed to an arbitrary covariance metric.

The second common method is known as the marker-variable method. To set the location parameters, one element of τ is set to zero (or a nonzero

value) for each construct. To set the scale, one element of λ is fixed to 1.0 (or any other nonzero value) for each construct. This method of identification is less desirable than the 1st and 3rd methods because the location and scale of the latent construct is determined arbitrarily on the basis of which indicator is chosen. Reise, Widaman, and Pugh [19] recommend that if one chooses this approach the marker variables should be supported by previous research or selected on the basis of strong theory.

A third possible identification method is to constrain the sum of τ for each factor to zero [20]. For the scale identification, the λ s for a factor should sum to p , the number of manifest variables. This method forces the mean and variance of the latent construct to be the weighted average of all of its indicators' means and loadings. The method has the advantage of providing a nonarbitrary scale that can legitimately vary across constructs and groups. It would be feasible, in fact, to compare the differences in means of two different constructs if one was theoretically motivated to do so (see [20], for more details of this method).

Testing for Measurement Invariance and Latent Construct Differences

In conducting cross-group tests of equality, either a statistical or a modeling rationale can be used for evaluating the tenability of the cross-group restrictions [9]. With a statistical rationale, an equivalence test is conducted as a nested-model comparison between a model in which specific parameters are constrained to equality across groups and one in which these parameters (and all others) are freely estimated in all groups. The difference in χ^2 between the two models is a test of the equality restrictions (with degrees of freedom equal to the difference in their degrees of freedom). If the test is nonsignificant then the statistical evidence indicates no cross-group differences between the equated parameters. If it is significant, then evidence of cross-group inequality exists.

The other rationale is termed a *modeling rationale* [9]. Here, model constraints are evaluated using practical fit indices to determine the overall adequacy of a fitted model. This rationale is used for large models with numerous constrained parameters because the χ^2 statistic is an overly sensitive index of model fit, particularly for large numbers of constraints and when estimated on large sample sizes (e.g., [10]).

From this viewpoint, if a model with *numerous* constraints evinces adequate levels of practical fit, then the *set* of constraints are reasonable approximations of the data.

Both rationales could be used in testing the *measurement level* and the *latent level* parameters. Because these two levels represent distinctly and qualitatively different empirical and theoretical goals, however, their corresponding rationale could also be different. Specifically, testing measurement equivalence involves evaluating the general tenability of an imposed indicator-to-construct structure via overall model fit indices. Here, various sources of model misfit (random or systematic) may be deemed substantively trivial if model fit is acceptable (i.e., if the model provides a reasonable approximation of the data; [2, 9]). The conglomerate effects of these sources of misfit, when sufficiently small, can be depicted parsimoniously as residual variances and general lack of fit, with little or no loss to theoretical meaningfulness (i.e., the trade-off between empirical accuracy and theoretical parsimony; [11]). When compared to a non-invariance model, an invariance model differs substantially in interpretability and parsimony (i.e., fewer parameter estimates than a non-invariance model), and it provides the theoretical and mathematical basis for quantitative between-group comparisons.

In contrast to the measurement level, the latent level reflects interpretable, error-free effects among constructs. Here, testing them for evidence of systematic differences (i.e., the hypothesis-testing phase of an analysis) is probably best done using a statistical rationale because it is a precise criteria for testing the specific theoretically driven questions about the constructs and because such substantive tests are typically narrower in scope (i.e., fewer parameters are involved). However, such tests should carefully consider issues such as error rate and effect size.

Numerous examples of the application of MACS modeling can be found in the literature, however, Little [9] offers a detailed didactic discussion of the issues and steps involved when making cross-group comparisons (including the LISREL source code used to estimate the models and a detailed Figural representation). His data came from a cross-cultural study of personal agency beliefs about school performance that included 2493 boys and girls from Los Angeles, Moscow, Berlin, and Prague. Little conducted an 8-group MACS comparison of boys and girls across the

four sociocultural settings. His analyses demonstrated that the constructs were measurement equivalent (i.e., had strong factorial invariance) across all groups indicating that the translation process did not unduly influence the measurement properties of the instrument. However, the constructs themselves revealed a number of theoretically meaningful differences, including striking differences in the mean levels and the variances across the groups, but no differences in the strength of association between the two primary constructs examined.

Extensions to Longitudinal MACS Modeling

The issues related to cross-group comparisons with MACS models are directly applicable to longitudinal MACS modeling. That is, establishing the measurement equivalence (strong metric invariance) of a construct's indicators over time is just as important as establishing their equivalence across subgroups. One additional component of longitudinal MACS modeling that needs to be addressed is the fact that the specific variances of the indicators of a construct will have some degree of association across time. Here, independence of the residuals is not assumed, but rather dependence of the unique factors is expected. In this regard, the *a priori* factor model, when fit across time, would specify and estimate all possible residual correlations of an indicator with itself across each measurement occasion.

Summary

MACS models are a powerful tool for cross-group and longitudinal comparisons. Because the means or intercepts of measured indicators are included explicitly in MACS models, they provide a very strong test of the validity of construct comparisons (i.e., measurement equivalence). Moreover, the form of the group- or time-related differences can be tested on many aspects of the constructs (i.e., means, variances, and covariances or correlations). As outlined here, the tenability of measurement equivalence (i.e., construct comparability) can be tested using model fit indices (i.e., the modeling rationale), whereas specific hypotheses about the nature of possible group differences on the constructs can be tested using precise statistical criteria. A measurement equivalent model is advantageous for three reasons: (a) it is theoretically very parsimonious and, thus, a reasonable *a*

priori hypothesis to entertain, (b) it is empirically very parsimonious, requiring fewer estimates than a non-invariance model, and (c) it provides the mathematical and theoretical basis by which quantitative cross-group or cross-time comparisons can be conducted. In other words, strong factorial invariance indicates that constructs are fundamentally similar in each group or across time (i.e., comparable) and hypotheses about the nature of possible group- or time-related influences can be meaningfully tested on any of the constructs' basic moments across time or across each group whether the groups are defined on the basis of culture, gender, or any other grouping criteria.

References

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [2] Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage Publications, Newbury Park, pp. 136–162.
- [3] Byrne, B.M., Shavelson, R.J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance, *Psychological Bulletin* **105**, 456–466.
- [4] Cattell, R.B. (1944). Parallel proportional profiles and other principles for determining the choice of factors by rotation, *Psychometrika* **9**, 267–283.
- [5] Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method, *Journal of Management* **25**, 1–27.
- [6] Horn, J.L. & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research, *Experimental Aging Research* **18**, 117–144.
- [7] Horst, P. & Schaie, K.W. (1956). The multiple group method of factor analysis and rotation to a simple structure hypothesis, *Journal of Experimental Education* **24**, 231–237.
- [8] Lawley, D.N. & Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd Edition, Butterworth, London.
- [9] Little, T.D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues, *Multivariate Behavioral Research* **32**, 53–76.
- [10] Marsh, H.W., Balla, J.R. & McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size, *Psychological Bulletin* **103**, 391–410.
- [11] McArdle, J.J. (1996). Current directions in structural factor analysis, *Current Directions* **5**, 11–18.
- [12] McArdle, J.J. & Cattell, R.B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems, *Multivariate Behavioral Research* **29**, 63–113.
- [13] Meredith, W. (1964). Rotation to achieve factorial invariance, *Psychometrika* **29**, 187–206.
- [14] Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance, *Psychometrika* **58**, 525–543.
- [15] Millsap, R.E. (1997). Invariance in measurement and prediction: their relationship in the single-factor case, *Psychological Methods* **2**, 248–260.
- [16] Millsap, R.E. & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations, *Psychological Methods* **9**, 93–115.
- [17] Mulaik, S.A. (1972). *The Foundations of Factor Analysis*, McGraw-Hill, New York.
- [18] Muthén, B.O. (1989). Factor structure in groups selected on observed scores, *British Journal of Mathematical and Statistical Psychology* **42**, 81–90.
- [19] Reise, S.P., Widaman, K.F. & Pugh, R.H. (1995). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance, *Psychological Bulletin* **114**, 552–566.
- [20] Slegers, D.W. & Little, T.D. (in press). Evaluating contextual influences using multiple-group, longitudinal mean and covariance structures (MACS) methods, in *Modeling contextual influences in longitudinal data*, T.D. Little, J.A. Bovaird & J. Marquis, eds, Lawrence Erlbaum, Mahwah.
- [21] Sörbom, D. (1982). Structural equation models with structured means, in *Systems Under Direct Observation*, K.G. Jöreskog & H. Wold, eds, Praeger, New York, pp. 183–195.
- [22] Steenkamp, J.B. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research, *Journal of Consumer Research* **25**, 78–90.
- [23] Widaman, K.F. & Reise, S.P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain, in *The science of prevention: Methodological advances from alcohol and substance abuse research*, K.J. Bryant & M. Windle et al., eds, American Psychological Association, Washington, pp. 281–324.

Further Reading

- MacCallum, R.C., Browne, M.W. & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling, *Psychological Methods* **1**, 130–149.
- McGaw, B. & Jöreskog, K.G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status, *British Journal of Mathematical and Statistical Psychology* **24**, 154–168.

(See also **Structural Equation Modeling: Latent Growth Curve Analysis**)

TODD D. LITTLE AND DAVID W. SLEGERS

Factor Analysis: Multitrait–Multimethod

WERNER WOTHKE

Volume 2, pp. 623–628

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Analysis: Multitrait–Multimethod

The Multitrait–Multimethod Matrix

The well-known paper by Campbell and Fiske [6] proposed the multitrait–multimethod (MTMM) matrix as a measurement design to study trait validity across assessment methods. Their central idea was that traits should be independent of and detectable by a variety of measurement methods. In particular, the magnitude of a trait should not change just because a different assessment method is used. Campbell and Fiske's main distinction was between two forms of validity, identified as *convergent* and *discriminant*. Convergent validity assures that measures of the same trait are statistically related to each other and that their error and unique components are relatively small. Discriminant validity postulates that measures of one trait are not too highly correlated with measures of different traits and particularly not too highly correlated just because they share the same assessment method.

The variables of an MTMM matrix follow a crossed-factorial measurement design whereby each of t traits is assessed with each of m measurement methods. Table 1 gives an example of how the observed variables and their correlation coefficients are arranged in the correlation matrix, conventionally ordering traits within methods. Because the matrix is symmetric, only entries in its lower half have been marked. Particular types of correlations are marked symbolically:

- V – validity diagonals, correlations of measures of the same traits assessed with different methods.
- M – monomethod triangles, correlations of measures of different traits that share the same methods.
- H – heterotrait–heteromethod triangles, correlations of measures of different traits obtained with different methods.
- 1 – main diagonal, usually containing unit entries. It is not uncommon to see the unit values replaced by reliability estimates.

Campbell and Fiske [6] proposed four qualitative criteria for evaluating convergent and discriminant validity by the MTMM matrix.

CF1 (convergent validity): ‘...the entries in the validity diagonal [V] should be significantly different from zero and sufficiently large ...’

CF2 (discriminant validity): ‘...a validity diagonal value [V] should be higher than the values lying in its column and row in the heterotrait–heteromethod triangles [H].’

CF3 (discriminant validity): ‘...a variable correlate higher with an independent effort to measure the same trait [V] than with measure designed to get at different traits which happen to employ the same method [M].’

CF4 (discriminant validity): ‘...the same pattern of trait interrelationship be shown in all of the heterotrait triangles of both the monomethod [M] and heteromethod [H] blocks.’

Depending on which of the criteria were satisfied, convergent or discriminant validity of assessment instruments would then be ascertained or rejected.

Confirmatory Factor Analysis Approach to MTMM

Confirmatory factor analysis (CFA) (*see Factor Analysis: Confirmatory*) was proposed as a model-oriented approach to MTMM matrix analysis by [1], [11], [12], and [13]. Among the several competing multivariate models for MTMM matrix analysis reviewed by [17] and [18], CFA is the only approach with an appreciable following in the literature.

Under the factor model (*see Factor Analysis: Exploratory*), the $n \times p$ observed data matrix $\mathbf{X}$ of n observations on p variables arises as a linear combination of $n \times k$, $k < p$ factor scores, with factor loading matrix $\mathbf{\Lambda}$, and uncorrelated residuals $\mathbf{E}$. The covariance structure of the observed data is

$$\mathbf{\Sigma}_x = \mathbf{\Lambda} \mathbf{\Phi} \mathbf{\Lambda}' + \mathbf{\Theta}, \quad (1)$$

where $\mathbf{\Phi}$ is the covariance matrix of the k latent factors and $\mathbf{\Theta}$ the diagonal covariance matrix of the residuals. There are two prominent models for MTMM factor analysis: the *trait-only* model [11, 12] expressing the observed variables in terms of t correlated trait factors and the *trait-method* factor model [1, 12, 13] with t trait and m method factors.

2 Factor Analysis: Multitrait–Multimethod

Table 1 Components of a 3-trait–3-method correlation matrix

		Method 1			Method 2			Method 3		
		Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
Method 1	Trait 1	1								
	Trait 2	M	1							
	Trait 3	M	M	1						
Method 2	Trait 1	V	H	H	1					
	Trait 2	H	V	H	M	1				
	Trait 3	H	H	V	M	M	1			
Method 3	Trait 1	V	H	H	V	H	H	1		
	Trait 2	H	V	H	H	V	H	M	1	
	Trait 3	H	H	V	H	H	V	M	M	1

Confirmatory Factor Analysis – Trait-only Model

The trait-only model allows one factor per trait. Trait factors are usually permitted to correlate. For the nine-variable MTMM matrix shown in Table 1, assuming the same variable order, the loading matrix Λ_τ has the following simple structure:

$$\Lambda_\tau = \begin{pmatrix} \lambda_{1,1} & 0 & 0 \\ 0 & \lambda_{2,2} & 0 \\ 0 & 0 & \lambda_{3,3} \\ \lambda_{4,1} & 0 & 0 \\ 0 & \lambda_{5,2} & 0 \\ 0 & 0 & \lambda_{6,3} \\ \lambda_{7,1} & 0 & 0 \\ 0 & \lambda_{8,2} & 0 \\ 0 & 0 & \lambda_{9,3} \end{pmatrix}. \quad (2)$$

and the matrix of factor correlations is

$$\Phi_\tau = \begin{pmatrix} 1 & \phi_{21} & \phi_{31} \\ \phi_{21} & 1 & \phi_{32} \\ \phi_{31} & \phi_{32} & 1 \end{pmatrix}. \quad (3)$$

All zero entries in Λ_τ and the diagonal entries in Φ_τ are fixed (predetermined) parameters; the p factor loading parameters $\lambda_{i,j}$, $t(t-1)/2$ factor correlations, and p uniqueness coefficients in the diagonal of Θ are estimated from the data. The model is identified when three or more methods are included in the measurement design. For the special case that all intertrait correlations are *nonzero*, model identification requires only two methods (two-indicator rule [2]).

The worked example uses the MTMM matrix of Table 2 on the basis of data by Flamer [8], also published in [9] and [22]. The traits are *Attitude toward Discipline in Children* (ADC), *Attitude toward Mathematics* (AM), and *Attitude toward the Law* (AL). The methods are all paper-and-pencil, differing by response format: dichotomous Likert (L) scales, Thurstone (Th) scales, and the semantic differential (SD) technique. Distinctly larger entries in the validity diagonals (in bold face) and similar patterns of small off-diagonal correlations in the monomethod triangles and heterotrait–heteromethod

Table 2 Flamer (1978) attitude data, sample A ($N = 105$)^a

	ADC_L	AM_L	AL_L	ADC_Th	AM_Th	AL_Th	ADC_SD	AM_SD	AL_SD
ADC_L	1.00								
AM_L	−0.15	1.00							
AL_L	0.19	−0.12	1.00						
ADC_Th	0.72	−0.11	0.19	1.00					
AM_Th	−0.01	0.61	−0.03	−0.02	1.00				
AL_Th	0.26	−0.04	0.34	0.27	0.01	1.00			
ADC_SD	0.42	−0.15	0.21	0.40	0.01	0.34	1.00		
AM_SD	−0.06	0.72	−0.05	−0.03	0.75	−0.03	0.00	1.00	
AL_SD	0.13	−0.12	0.46	0.17	−0.01	0.44	0.33	0.00	1.00

^aReproduced with permission from materials held in the University of Minnesota Libraries.

Table 3 Trait-only factor analysis of the Flamer attitude data

Factor loading matrix $\hat{\Lambda}_\tau$					
Method	Trait	Trait factors			Uniqueness estimates $\hat{\theta}$
		ADC	AM	AL	
Likert	ADC	0.85	0.0	0.0	0.28
	AM	0.0	0.77	0.0	0.41
	AL	0.0	0.0	0.61	0.63
Thurstone	ADC	0.84	0.0	0.0	0.29
	AM	0.0	0.80	0.0	0.36
	AL	0.0	0.0	0.62	0.62
Semantic diff	ADC	0.50	0.0	0.0	0.75
	AM	0.0	0.95	0.0	0.12
	AL	0.0	0.0	0.71	0.50
Factor correlations $\hat{\Phi}_\tau$					
		ADC	AM	AL	
ADC		1.0			
AM		−0.07	1.0		
AL		0.39	−0.05	1.0	
$\chi^2 = 23.28$					$P = 0.503$
df = 24					$N = 105$

blocks suggest some stability of the traits across the three methods.

The parameter estimates for the trait-only factor model are shown in Table 3. The solution is admissible and its low maximum-likelihood χ^2 -value signals acceptable statistical model fit. No additional model terms are called for. This factor model postulates considerable generality of traits across methods, although the large uniqueness estimates of some of the attitude measures indicate low factorial validity, limiting their practical use.

Performance of the trait-only factor model with other empirical MTMM data is mixed. In Wothke's [21] reanalyses of 23 published MTMM matrices, the model estimates were inadmissible or failed to converge in 10 cases. Statistically acceptable model fit was found with only 2 of the 23 data sets.

Confirmatory Factor Analysis – Traits Plus Methods Model

Measures may not only be correlated because they reflect the same trait but also because they share the same assessment method. Several authors [1, 12]

have therefore proposed the less restrictive *trait-method* factor model, permitting systematic variation due to shared methods as well as shared traits. The factor loading matrix of the expanded model simply has several columns of method factor loadings appended to the right, one column for each method:

$$\Lambda_{\tau\mu} = \left(\begin{array}{ccc|ccc} \lambda_{1,1} & 0 & 0 & \lambda_{1,4} & 0 & 0 \\ 0 & \lambda_{2,2} & 0 & \lambda_{2,4} & 0 & 0 \\ 0 & 0 & \lambda_{3,3} & \lambda_{3,4} & 0 & 0 \\ \lambda_{4,1} & 0 & 0 & 0 & \lambda_{4,5} & 0 \\ 0 & \lambda_{5,2} & 0 & 0 & \lambda_{5,5} & 0 \\ 0 & 0 & \lambda_{6,3} & 0 & \lambda_{6,5} & 0 \\ \lambda_{7,1} & 0 & 0 & 0 & 0 & \lambda_{7,6} \\ 0 & \lambda_{8,2} & 0 & 0 & 0 & \lambda_{8,6} \\ 0 & 0 & \lambda_{9,3} & 0 & 0 & \lambda_{9,6} \end{array} \right). \quad (4)$$

A particularly interesting form of factor correlation matrix is the block-diagonal model, which implies independence between trait and method factors:

$$\Phi_{\tau\mu} = \begin{pmatrix} \Phi_\tau & 0 \\ 0 & \Phi_\mu \end{pmatrix}. \quad (5)$$

4 Factor Analysis: Multitrait–Multimethod

In the structured correlation matrix (5), the submatrix Φ_τ contains the correlations among traits and the submatrix Φ_μ contains the correlations among methods.

While the block-diagonal trait-method model appeared attractive when first proposed, there has been growing evidence that its parameterization is inherently flawed. Inadmissible or unidentified model solutions are nearly universal with both simulated and empirical MTMM data [3, 15, 21]. In addition, identification problems of several aspects of the trait-method factor model have been demonstrated formally [10, 14, 16, 20]. For instance, consider factor loading structures whose nonzero entries are *proportional* by rows and columns:

$$\Lambda_{\tau\mu}^{(p)} = \left(\begin{array}{ccc|ccc} \lambda_1 & 0 & 0 & \lambda_4 & 0 & 0 \\ 0 & \delta_2 \cdot \lambda_2 & 0 & \delta_2 \cdot \lambda_4 & 0 & 0 \\ 0 & 0 & \delta_3 \cdot \lambda_3 & \delta_3 \cdot \lambda_4 & 0 & 0 \\ \delta_4 \cdot \lambda_1 & 0 & 0 & 0 & \delta_4 \cdot \lambda_5 & 0 \\ 0 & \delta_5 \cdot \lambda_2 & 0 & 0 & \delta_5 \cdot \lambda_5 & 0 \\ 0 & 0 & \delta_6 \cdot \lambda_3 & 0 & \delta_6 \cdot \lambda_5 & 0 \\ \delta_7 \cdot \lambda_1 & 0 & 0 & 0 & 0 & \delta_7 \cdot \lambda_6 \\ 0 & \delta_8 \cdot \lambda_2 & 0 & 0 & 0 & \delta_8 \cdot \lambda_6 \\ 0 & 0 & \delta_9 \cdot \lambda_3 & 0 & 0 & \delta_9 \cdot \lambda_6 \end{array} \right), \quad (6)$$

where the δ_i are $mt - 1$ nonzero scale parameters for the rows of $\Lambda_{\tau\mu}^{(p)}$, with $\delta_1 = 1$ fixed and all other δ_i estimated, and the λ_k are a set of $m + t$ nonzero scale parameters for the columns of $\Lambda_{\tau\mu}^{(p)}$, with all λ_k estimated. Grayson and Marsh [10] proved algebraically that factor models with loading matrix (6) and factor correlation structure (5) are unidentified no matter how many traits and methods are analyzed. Even if $\Lambda_{\tau\mu}^{(p)}$ is further constrained by setting all (row) scale parameters to unity ($\delta_i = 1$), the factor model will remain unidentified [20].

Currently, identification conditions for the general form of the trait-method model are not completely known. Identification and admissibility problems appear to be the rule with empirical MTMM data, although an identified, admissible, and fitting solution has been reported for one particular dataset [2]. However, in order to be identified, the estimated factor loadings must necessarily be different from the proportional structure in (6) – a difference that would complicate the evaluation of trait validity. Estimation itself can also be difficult: The usually iterative estimation process often approaches an intermediate solution of the form (6) and cannot

continue because the matrix of second derivatives of the fit function becomes rank deficient at that point. This is a serious practical problem because condition (6) is so general that it ‘slices’ the identified solution space into many disjoint subregions so that the model estimates can become extremely sensitive to the choice of start values. Kenny and Kashy [14] noted that ‘... estimation problems increase as the factor loadings become increasingly similar.’

There are several alternative modeling approaches that the interested reader may want to consult: (a) CFA with alternative factor correlation structures [19]; (b) CFA with correlated uniqueness coefficients [4, 14, 15]; (c) covariance components

analysis [22]; and (d) the direct product model [5]. Practical implementation issues for several of these models are reviewed in [14] and [22].

Conclusion

About thirty years of experience with confirmatory factor analysis of MTMM data have proven less than satisfactory. Trait-only factor analysis suffers from poor fit to most MTMM data, while the block-diagonal trait-method model is usually troubled by identification, convergence, or admissibility problems, or by combinations thereof. In the presence of method effects, there is no generally accepted multivariate model to yield summative measures of convergent and discriminant validity. In the absence of such a model, ‘(t)here remains the basic eyeball analysis as in the original article [6]. It is not always dependable; but it is cheap’ [7].

References

- [1] Althausen, R.P. & Heberlein, T.A. (1970). Validity and the multitrait-multimethod matrix, in *Sociological*

- Methodology* 1970, E.F. Borgatta, ed., Jossey-Bass, San Francisco.
- [2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
 - [3] Brannick, M.T. & Spector, P.E. (1990). Estimation problems in the block-diagonal model of the multitrait-multimethod matrix, *Applied Psychological Measurement* **14**(4), 325–339.
 - [4] Browne, M.W. (1980). Factor analysis for multiple batteries by maximum likelihood, *British Journal of Mathematical and Statistical Psychology* **33**, 184–199.
 - [5] Browne, M.W. (1984). The decomposition of multitrait-multimethod matrices, *British Journal of Mathematical and Statistical Psychology* **37**, 1–21.
 - [6] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
 - [7] Fiske, D.W. (1995). Reprise, new themes and steps forward, in *Personality, Research Methods and Theory. A Festschrift Honoring Donald W. Fiske*, P.E. Shrout & S.T. Fiske, eds, Lawrence Erlbaum Associates, Hillsdale.
 - [8] Flamer, S. (1978). The effects of number of scale alternatives and number of items on the multitrait-multimethod matrix validity of Likert scales, Unpublished Dissertation, University of Minnesota.
 - [9] Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of Likert scales via confirmatory factor analysis, *Multivariate Behavioral Research* **18**, 275–308.
 - [10] Grayson, D. & Marsh, H.W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis multitrait-multimethod models, *Psychometrika* **59**, 121–134.
 - [11] Jöreskog, K.G. (1966). Testing a simple structure hypothesis in factor analysis, *Psychometrika* **31**, 165–178.
 - [12] Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests, *Psychometrika* **36**(2), 109–133.
 - [13] Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices, *Psychometrika* **43**(4), 443–477.
 - [14] Kenny, D.A. & Kashy, D.A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis, *Psychological Bulletin* **112**(1), 165–172.
 - [15] Marsh, H.W. & Bailey, M. (1991). Confirmatory factor analysis of multitrait-multimethod data: A comparison of alternative models, *Applied Psychological Measurement* **15**(1), 47–70.
 - [16] Millsap, R.E. (1992). Sufficient conditions for rotational uniqueness in the additive MTMM model, *British Journal of Mathematical and Statistical Psychology* **45**, 125–138.
 - [17] Millsap, R.E. (1995). The statistical analysis of method effects in multitrait-multimethod data: a review, in *Personality, Research Methods and Theory. A Festschrift Honoring Donald W. Fiske*, P.E. Shrout & S.T. Fiske, eds, Lawrence Erlbaum Associates, Hillsdale.
 - [18] Schmitt, N. & Stults, D.M. (1986). Methodology review: analysis of multitrait-multimethod matrices, *Applied Psychological Measurement* **10**, 1–22.
 - [19] Widaman, K.F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data, *Applied Psychological Measurement* **9**, 1–26.
 - [20] Wothke, W. (1984). The estimation of trait and method components in multitrait-multimethod measurement, Unpublished doctoral dissertation, University of Chicago.
 - [21] Wothke, W. (1987). Multivariate linear models of the multitrait-multimethod matrix, in *Paper Presented at the Annual Meeting of the American Educational Research Association*, Washington, (paper available through ERIC).
 - [22] Wothke, W. (1996). Models for multitrait-multimethod matrix analysis, in *Advanced Structural Equation Modeling. Issues and Techniques*, G.A. Marcoulides & R.E. Schumacker, eds, Lawrence Erlbaum Associates, Mahwah.

(See also **History of Path Analysis; Residuals in Structural Equation, Factor Analysis, and Path Analysis Models; Structural Equation Modeling: Overview**)

WERNER WOTHKE

Factor Analysis of Personality Measures

WILLIAM F. CHAPLIN

Volume 2, pp. 628–636

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Analysis of Personality Measures

The technique of **factor analysis** was developed about 100 years ago by **Charles Spearman** [12] who applied the technique to the observed correlations among measures of mental abilities. Briefly, factor analysis is a statistical technique that derives aggregates of variables (typically called ‘factors’) from the observed relations (typically indexed by correlations) among those variables. The result of Spearman’s analysis was the identification of a single factor that seemed to underlie observed scores on a large number of measures of human mental ability. Subsequently, further applications of factor analysis to the mental ability domain indicated that the one factor model was too simple. In particular, **Louis Thurstone** suggested seven primary mental ability factors rather than the single factor claimed by Spearman. Interestingly, Thurstone’s 1933 American Psychological Association presidential address, *Vectors of the Mind*, [13] in which he presented this alternate view of the structure of mental abilities focused as much or more on the application of factor analysis to personality data, and this represents the first presentation of a major factor analysis of personality measures. Thurstone, however, later dropped this line of investigation to focus on mental abilities.

Numerous other personality scientists soon followed Thurstone’s initial lead and began using factor analytic techniques to identify, evaluate, and refine the major dimensions of personality. The personality theories and measures of **Raymond Cattell** and Hans Eysenck represent two major early applications and more recently the factor analyses of Jack Digman, Lewis Goldberg, Paul Costa, and Jeff McCrae, and a host of others have laid the foundation for a widely used, though not by any means universally accepted, five-factor structure of personality often called the *Big Five*. Today there are a variety of structural models of personality that are based on factor analyses. A number of these are summarized in Table 1.

This table is intended to be illustrative rather than comprehensive or definitive. There are other systems, other variants on the systems shown here,

and other scientists who might be listed. More comprehensive tables can be found in [4, 9, and 10]. Although Table 1 represents only a portion of the factor analytic models of personality, it is sufficient to raise the fundamental issue that will be the focus of this contribution: Why does the same general analytic strategy (factor analysis) result in structural models of personality that are so diverse? In addressing this issue, I will consider the variety of factor analytic procedures that result from different subjective decisions about the conduct of a factor analysis. A more thorough discussion of these issues can be found in [6] and [7]. These decisions include (a) the sample of observed items to be factored, (b) the method of factor extraction, (c) the criteria for deciding the number of factors to be extracted, (d) the type of factor rotation if any, and (e) the naming of the factors. Readers who believe that science is objective and who believe that the diversity of results obtained from factor analyses is *prima facie* evidence that the technique is unscientific will find the tone of this contribution decidedly unsympathetic to that view.

What Personality Variables are to be Included in a Factor Analysis?

The first decision in any scientific study is what to study. This is an inherently subjective decision and, at its broadest level, is the reason that some of us become, say, chemists and others of us become, say, psychologists. In the more specific case of studying the structure of human personality, we must also begin with a decision of which types of variables are relevant to personality. Factor analysis, just as any other statistical technique, can only operate on the data that are presented. In the case of personality structure for example, a factor representing Extraversion will only be found if items that indicate Extraversion are present in the data: No Extroversion items; no Extroversion factor. An historical example of the influence of this decision on the study of personality structure was Cattell’s elimination of a measure of intelligence from early versions of the domains he factored. This marked the point at which a powerful individual difference variable, intelligence, disappeared from the study of personality. More recently, the decision on the part of Big Five theorists to exclude terms that are purely evaluative such as ‘nice’, or ‘evil’, from the personality domain meant that no factors representing general evaluation were

2 Factor Analysis of Personality Measures

Table 1 Illustration of the major structural models of personality based on factor analysis

Number of factors	Representative labels and structure	Associated theorists
2	Love-Hate; Dominance-Submission (interpersonal circle)	Leary, Wiggins
2	Alpha (A, C, N) Beta (E, O) (Higher order factors of the Big Five)	Digman
3	Extroversion, Neuroticism, Psychoticism	Eysenck
5	E, A, C, N, O (Big Five; Five-Factor Model)	Digman, Goldberg, Costa & McCrae
7	E, A, C, N, O + Positive and Negative Evaluation	Tellegen, Waller, Benet
16	16- PF; 16 Primary factors further grouped into five more global factors ^a	Cattell

Note: E = Extroversion, A = Agreeableness, C = Conscientiousness, N = Neuroticism, O = Openness.

^aA complete list of the labels for the 16 PF can be found in [3].

included in the structure of personality. Adding such items to the domain to be factored resulted, not surprisingly, in a model called the *Big Seven* as shown in Table 1.

Cattell's decision to exclude intelligence items or Big Five theorists' decisions to exclude purely evaluative items represent different views of what is meant by personality. It would be difficult to identify those views as correct or incorrect in any objective sense, but recognizing these different views can help clarify the differences in Table 1 and in other factor analyses of personality domains. The point is that understanding the results of a factor analysis of personality measures must begin with a careful evaluation of the measures that are included (and excluded) and the rationale behind such inclusion or exclusion. Probably the most prominent rationale for selecting variables for a factor analysis in personality has been the 'lexical hypothesis'. This hypothesis roughly states that all of the most important ways that people differ from each other in personality will become encoded in the natural language as single word person descriptive terms such as 'friendly' or 'dependable'. On the basis of this hypothesis, one selects words from a list of all possible terms that describe people culled from a dictionary and then uses those words as stimuli for which people are asked to describe themselves or others on those terms. Cattell used such a list that was compiled by Allport and Odbert [1] in his analyses, and more recently, the Big Five was based on a similar and more recent list compiled by Warren Norman [11].

How (and Why) Should Personality Factors be Extracted? The basic data used in a factor analysis

of personality items are responses (typically ratings of descriptiveness of the item about one's self or possibly another person) from N subjects to k personality items; for example, 'talks to strangers', 'is punctual', or 'relaxed'. These $N \times k$ responses are then converted into a $k \times k$ correlation (or less often a covariance) matrix, and the $k \times k$ matrix is then factor analyzed to yield a factor matrix showing the 'loadings' of the k variables on the m factors. Specifically, factor analysis operates on the common (shared) variance of the variables as measured by their intercorrelations. The amount of variance a variable shares with the other variables is called the *variables communality*. Factor analysis proceeds by extracting factors iteratively such that the first factor accounts for as much of the total common variance across the items (called the *factor's eigenvalue*) as possible, the second factor accounts for as much of the remaining common variance as possible and so on. Figure 1 shows a heuristic factor matrix. The elements of the matrix are the estimated correlations between each variable and each factor. These correlations are called 'loadings'. To the right of the matrix is a column containing the final communality estimates (usually symbolized as h^2). These are simply the sum of the squared loadings for each variable across the m factors and thus represent the total common variance in each variable that is accounted for by the factors. At the bottom of the matrix are the eigenvalues of the factors. These are the sum of the squared loadings for each factor across the k variables and thus represent the total amount of variance accounted for by each factor.

The point at which the correlation matrix is converted to a factor matrix represents the next

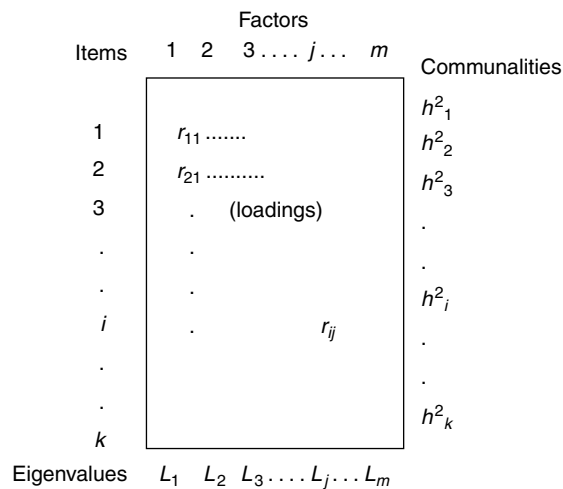


Figure 1 Heuristic representation of a factor matrix

crucial subjective decision point in the factor analysis. Although the communalities of the variables can be calculated from the final factor matrix, these communalities must be initially estimated for the factor analysis to proceed and the investigator must decide how those initial communality values are to be estimated. The vast majority of factor analyses are based on one of two possible decisions about these estimates. In principle, these decisions reflect the investigators belief about the nature of factors and the goal of the factor analysis. One reason for extracting factors from a matrix of correlations is simply as an aid to interpreting the complex patterns implied by those correlations. The importance of this aid can be readily appreciated by anyone who has tried to discern how groups of variables are similar and different on the basis of the 4950 unique correlations available from a set of 100 items, or, less ambitiously, among the 435 unique correlations available from a set of 30 items or measures. This 'orderly simplification' of a correlation matrix as a goal of factor analysis has been attributed to Cyril Burt, among others.

A second reason for extracting factors is perhaps more profound. This reason is to discover the underlying factors that 'cause' individuals to respond to the items in certain ways. This view of factors as 'causes', is, of course, more controversial because of the 'correlations do not imply causation' rule. However, this rule should not blind us to the fact that relations among variables are caused by something;

it is just that we do not necessarily know what that cause is on the basis of correlations alone.

The difference between this descriptive and explanatory view of factor analysis is the foundation of the two major approaches to factor extraction; **principal component analysis (PC)** and principle axis factor analysis (PF) (*see Factor Analysis: Exploratory*). Figure 2 illustrates the difference between these two approaches using structural model diagrams. As can be seen in Figures 2(a) and 2(b) the difference between PC and PF is the direction of the arrows in the diagram. Conceptually, the direction of the arrows indicates the descriptive emphasis of PC analysis and the causal emphasis of PF analysis. In PC analysis the items together serve to 'define' the component and it serves to summarize the items that define it. In PF analysis the underlying factor serves as a cause of why people respond consistently to a set of items. The similarity of the items, the responses to which are caused by the factor, is used to label the cause, which could be biological, conditioned, or cognitive.

Because correlations are bidirectional, the direction of the arrows in a path diagram is statistically arbitrary and both diagrams will be equally supported by the same correlation matrix. However, there is a crucial additional difference between Figures 2(a) and 2(b) that does lead to different results between PC and PF. This difference is shown in Figure 1(c), which adds error terms to the item responses when they are viewed as 'caused' by the factor in PF analysis. The addition of error in PF analysis recognizes that the response to items is not perfectly predicted by the underlying factor. That is, there is some 'uniqueness' or 'error' in individuals' responses in addition to the common influence of the factor. In PC analysis, no error is assigned to the item responses as they are not viewed as caused by the factor. It is at this point that the statistical consequence of these views becomes apparent. In PC analysis, the initial communality estimates for the item are all fixed at 1.0 as all of the variance is assumed to be common. In PF analysis, the initial communality estimates are generally less than 1.0 (see next paragraph) to reflect that some of the item variance is unique. The consequence of recognizing that some of the variability in people's responses to items is unique to that item is to reduce the amount of variance that can be 'explained' or attributed to the factors. Thus, PF analysis typically

4 Factor Analysis of Personality Measures

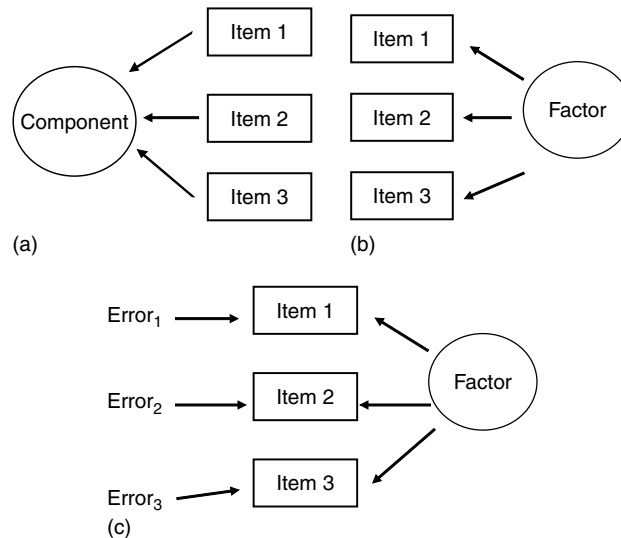


Figure 2 Path diagram Illustrating the difference between PC and PF

results in factors that account for less variance than PC analysis.

There is also a computational consequence of choosing PF over PC analysis. PF analysis is much more difficult from a computational standpoint than PC because one needs to estimate the error or uniqueness of the items before the analysis can proceed. This is typically done by regressing each item on all the others in the set to be factored, and using the resulting R^2 as the estimate of the items common variance (communality) and $1 - R^2$ as the items unique variance. **Multiple linear regression** requires inverting a correlation matrix, a time consuming, tedious, and error-prone task. If one were to factor, say, 100 items one would have to invert 100 matrices. This task would simply be beyond the skills and temperament of most investigators and as a consequence the vast majority of historical factor analyses used the PC approach, which requires no matrix inversion. Today we have computers, which, among other things are designed for time consuming, tedious, and error-prone tasks, so the computational advantage of PC is no longer of much relevance. However, the conservative nature of science, which tends to foster continuity of methods and measures, has resulted in the vast majority of factor analyses of personality items to continue to be based on PC, regardless of the (often unstated) view of investigator about the nature of factors or the goal of the analysis.

Within the domain of personality it is often the case that similar factor structures emerge from the same data regardless of whether PC or PF is employed, probably because the initial regression-based communality estimates for personality variables in PF tend to approach the 1.0 estimates used by PC analysis. Thus, the decision to use PC or PF on personality data may be of little practical consequence. However, the implied view of factors as descriptive or causal by PC or PF respectively still has important implications for the study of personality. The causal view of factors must be a disciplined view to avoid circularity. For example, it is easy to 'explain' that a person has responded in an agreeable manner because they are high on the agreeableness factor. Without further specifying, and independently testing, the source of that factor (e.g., genetic, cognitive, environmental), the causal assertion is circular ('He is agreeable because he is agreeable') and untestable. The PC view avoids this problem by simply using the factor descriptively without implying a cause.

However, the 'merely' descriptive view of factors is scientifically less powerful and two of the earliest and most influential factor analytic models of personality of Cattell [3] and Eysenck [5] both viewed factors as causal. Eysenck based his three factors on a strong biological theory that included the role of individual differences in brain structure and systems of biological activation and inhibition as the basis

of personality, and Eysenck used factor analysis to evaluate his theory by seeing if factors consistent with his theory could be derived from personality ratings. Cattell, on the other hand, did not base his 16 factors on an explicit theory but instead viewed factor analysis as a tool for empirically discovering the important and replicable factors that caused personality. The widely accepted contemporary model of five factors also has both descriptive and causal interpretations. The term 'Five-Factor Model' used by Costa and McCrae among others emphasizes a causal interpretation, whereas the term 'Big Five' used by Goldberg among others emphasizes the descriptive view.

How Many Factors are There? Probably the most difficult issue in factor analysis is deciding on the number of factors. Within the domain of personality, we have seen that the number of factors extracted is influenced crucially by the decision of how many and what type of items to factor. However, another reason that different investigators may report different numbers of factors is that there is no single criterion for deciding how many factors are needed or useful to account for the common variance among a set of items. The problem is that as one extracts more factors one necessarily accounts for more common variance. Indeed in PC analysis one can extract as many factors as there are items in the data set and in doing so one can account for all the variance. Thus, the decision about the number of factors to extract is ultimately based on the balance between the statistical goal of accounting for variance and the substantive goal of simplifying a set of data into a smaller number of *meaningful* descriptive components or underlying causal factors. The term 'meaningful' is the source of the inherent subjectivity in this decision

The most common objective criteria that has been used to decide on the number of factors is Kaiser's 'eigenvalues greater than 1.0' rule. The logic of this rule is that, at a minimum, a factor should account for more common variance than any single item. On the basis of this logic, it is clear that this rule only applies to PC analysis where the common variance of an item is set at 1.0 and indeed Kaiser proposed this rule for PC analysis. Nonetheless, one often sees this rule misapplied in PF analyses. Although there is a statistical objectivity about this rule, in practice its application often results in factors

that are specific to only one or two items and/or factors that are substantively difficult to interpret or name.

One recent development that addresses the number of factors problem is the use of factor analyses based on maximum-likelihood criteria. In principle, this provides a statistical test of the 'significance' of the amount of additional variance accounted for by each additional factor. One then keeps extracting factors until the additional variance accounted for by each factor does not significantly increase over the variance accounted for by the previous factor. However, it is still often the case that factors that account for 'significantly' more variance do not include large numbers of items and/or are not particularly meaningful. Thus, the tension between statistical significance and substantive significance remains, and ultimately the number of factors reported reflects a subjective balance between these two criteria.

How Should the Factors be Arranged (Rotated)?

Yet another source of subjectivity in factor analysis results because the initial extraction of factors does not provide a statistically unique set of factors. Statistically, factors are extracted to account for as much variance as possible. However, once a set of factors is extracted, it turns out that there are many different combinations of factors and item loadings that will account for exactly the same amount of total variance of each item. From a statistical standpoint, as long as a group of factors accounts for the same amount of total variance, there is no basis for choosing one group over another. Thus, investigators are free to select whatever arrangements of factors and item loadings they wish. The term that is used to describe the rearrangement of factors among a set of personality items is 'rotation', which comes from the geometric view of factors as vectors moving (rotating) through a space defined by items.

There is a generally accepted criterion, called *simple structure*, that is used to decide how to rotate factors. An ideal simple structure is one where each item correlates 1.0 with one factor and 0.00 with the other factors. In actual data this ideal will not be realized, but the goal is to come as close to this ideal as possible for as many items as possible. The rationale for simple structure is simplicity and this rationale holds for both PC and PF analyses. For PC analysis, simple

structure results in a description of the relations among the variables that is easy to interpret because there is little item overlap between factors. For PF analysis the rationale is that scientific explanations should be as simple as possible. However, there are several different statistical strategies that can be used to approximate simple structure and the decision about which strategy to use is again a subjective one.

The major distinction between strategies to achieve simple structure is oblique versus orthogonal rotation of factors. As the labels imply, oblique rotation allows the factors to be correlated with each other whereas orthogonal rotation constrains the factors to be uncorrelated. Most factor analyses use an orthogonal rotation based on a specific strategy called 'Varimax'. Although other orthogonal strategies exist (e.g., Equimax, Quartimax) the differences among these in terms of rotational results are usually slight and one seldom encounters these alternative orthogonal approaches. Orthogonal approaches to the rotation of personality factors probably dominate in the literature because of their computational simplicity relative to oblique rotations. However, the issue of computational simplicity is no longer of much concern with the computer power available today so the continued preference for orthogonal rotations may, as with the preference for PC over PF, be historically rather than scientifically based.

Oblique rotations of personality factors have some distinct advantages over orthogonal rotations. In general these advantages result because oblique rotations are less constrained than orthogonal ones. That is, oblique rotations *allow* the factors to be correlated with each other, whereas orthogonal rotations *force* the factors to be uncorrelated. Thus, in the pursuit of simple structure, oblique rotations will be more successful than orthogonal ones because oblique rotations have more flexibility. Oblique rotations can, in some sense, transfer the complexity of items that are not simple (load on more than one factor) to the factors by making the relations among the factors more complex. Perhaps the best way to appreciate the advantage of oblique rotations over orthogonal ones is to note that if the simple structure factors are orthogonal or nearly so, oblique rotations will leave the factors essentially uncorrelated and oblique rotations will become identical (or nearly so) to orthogonal ones. A second advantage of oblique rotations of personality factors is that it allows the investigator

to explore higher order factor models—that is factors of factors. Two of the systems shown in Table 1, Digman's Alpha and Beta factors and Cattell's five Global Factors for the 16 PF represent such higher order factor solutions.

Simple structure is generally accepted as a goal of factor rotation and is the basis for all the specific rotational strategies available in standard factor analytic software. However, within the field of personality there has been some theoretical recognition that simple structure may not be the most appropriate way to conceptualize personality. The best historical example of this view is the interpersonal circle of Leary, Wiggins, and others [14]. A circular arrangement of items around two orthogonal axes means that some items must load equally highly on both factors, which is not simple structure. In the interpersonal circle, for example, an item such as 'trusting' has both loving and submissive aspects, and so would load complexly on both the Love-Hate and Dominance-Submission factors. Likewise, 'cruel' has both Dominant and hateful aspects. More recently, a complex version of the Big Five called the *AB5C structure* that explicitly recognizes that many personality items are blends of more than one factor was introduced by [8]. In using factor analysis to identify or evaluate circumplex models or any personality models that explicitly view personality items as blends of factors, simple structure will not be an appropriate criterion for arranging the factors.

What Should the Factors be Called? In previous sections the importance of the meaningfulness and interpretation of personality factors as a basis for evaluating the acceptability of a factor solution has been emphasized. But, of course, the interpretation and naming of factors is another source of the inherent subjectivity in the process. This subjectivity is no different than the subjectivity of all of science when it comes to interpreting the results – but the fact that different, but reasonable, scientists will often disagree about the meaning or implications of the same data certainly applies to the results of a factor analysis.

The interpretation problem in factor analysis is perhaps particularly pronounced because factors, personality or otherwise, have no objective reality. Indeed, factors do not result from a factor analysis, rather the result is a matrix of factor loadings such as

the one shown in Figure 1. On the basis of the content of the items and their loadings in the matrix, the 'nature' of the factor is inferred. That is, we know a factor through the variables with which it is correlated. It is because factors do not exist and are not, therefore, directly observed that we often call them 'latent' factors. Latent factors have the same properties as other latent variables such as 'depression,' 'intelligence', or 'time'. None of these variables is observed or measured directly, but rather they are measured via observations that are correlated with them such as loss of appetite, vocabulary knowledge, or the movement of the hand on a watch. A second complication is that in the factor analyses described here there are no statistical tests of whether a particular loading is significant; instead different crude standards such as loadings over 0.50 or over 0.30 have been used to decide if an item is 'on' a factor. Different investigators can, of course, decide on different standards, with the result that factors are identified by different items, even in the same analysis.

Thus, it should come as no surprise that different investigators will call the 'same' factor by a different name. Within the domain of the interpersonal circle, for example, the factors have been called *Love and Hate*, or *Affiliation and Dominance*. Within the Big Five, various labels have been applied to each factor, as shown in Table 2. Although there is a degree of similarity among the labels in each column, there are clear interpretive differences as well. The implication of this point is that one should not simply look at the name or interpretation an investigator applies to a factor, but also at the factor-loading matrix so that the basis for the interpretation can be evaluated. It is not uncommon to see the same label applied to somewhat different patterns of loadings, or for different labels to be applied to the same pattern of loadings.

Some investigators, perhaps out of recognition of the difficulty and subjectivity of factor naming, have eschewed applying labels at all and instead refer to factors by number. Thus, in the literature on the Big Five, one may see reference to Factors I, II, III, IV, V. Of course, those investigators know the 'names' that are typically applied to the numbered factors and these are shown in Table 2. Another approach has been to name the factors with uncommon labels to try to separate the abstract scientific meaning of a factor from its everyday interpretation. In particular, Cattell used this approach with the 16PF, where he applied labels to his factors such as 'Parmia', 'Premsia', 'Autia', and so on. Of course, translations of these labels into their everyday equivalents soon appeared (Parmia is 'Social Boldness', Premsia is 'Sensitivity', and Autia is 'Abstractedness'), but the point can be appreciated, even if not generally followed.

A Note on Confirmatory Factor Analysis. This presentation of factor analysis of personality measures has focused almost exclusively on approaches to factor analysis that are often referred to as 'exploratory' (see **Factor Analysis: Exploratory**). This label is somewhat misleading as it implies that investigators use factor analysis just to 'see what happens'. Most investigators are not quite so clueless and the factor analysis of personality items usually takes place under circumstances where the investigator has some specific ideas about what items should be included in the set to be factored, and hypotheses about how many factors there are, what items will be located on the same factor, and even what the factors will be called. In this sense, the analysis has some 'confirmatory' components.

In fact the term 'exploratory' refers to the fact that in these analyses a correlation matrix is submitted for analyses and the analyses generates the optimal

Table 2 Some of the different names applied to the Big Five personality factors in different systems

Factor I	Factor II	Factor III	Factor IV	Factor V
Extroversion	Agreeableness	Conscientiousness	Emotional Stability	Openness to Experience
Surgency	Femininity	High Ego Control	Neuroticism (r)	Intellect
Power	Love	Prudence	Adjustment	Imagination
Low Ego Control	Likeability	Work Orientation	Anxiety (r)	Rebelliousness
Sociability		Impulsivity (r)		

Note: r = label is reversed relative to the other labels.

factors and loadings empirically for that sample of data and without regard to the investigators ideas and expectations. Thus the investigator's beliefs do not guide the analyses and so they are not directly tested. Indeed, there is no hypothesis testing framework within exploratory factor analysis and this is why most decisions associated with this approach to factor analysis are subjective.

The term confirmatory factor analysis (CFA) (*see* **Factor Analysis: Confirmatory**) is generally reserved for a particular approach that is based on **structural equation modeling** as represented in programs such as LISREL, EQS, or AMOS (*see* **Structural Equation Modeling: Software**). CFA is explicitly guided by the investigators beliefs and hypotheses. Specifically, the investigator indicates the number of factors, designates the variables that load on each factor, and indicates if the factors are correlated (oblique) or uncorrelated (orthogonal). The analyses then proceed to generate a hypothetical correlation matrix based on the investigator's specifications and this matrix is compared to the empirical correlation matrix based on the items. Chi-square goodness-of-fit tests and various modifications of these as 'fit indices' are available for evaluating how close the hypothesized matrix is to the observed matrix. In addition, the individual components of the model such as the loadings of individual variables on specific factors and proposed correlations among the factors can be statistically tested. Finally, the increment in the goodness-of-fit of more complex models relative to simpler ones can be tested to see if the greater complexity is warranted.

Clearly, when investigators have some idea about what type of factor structure should emerge from their analysis, and investigators nearly always have such an idea, CFA would seem to be the method of choice. However, the application of CFA to personality data has been slow to develop and is not widely used. The primary reason for this is that CFA does not often work well with personality data. Specifically, even when item sets that seem to have a well-established structure such as those contained in the Big Five Inventory (BFI-44 [2]) or the Eysenck Personality Questionnaire (EPQ [5]) are subjected to CFA based on that structure, the fit of the established structure to the observed correlations is generally below the minimum standards of acceptable fit.

The obvious interpretation of this finding is that factor analyses of personality measures do not lead

to structures that adequately summarize the complex relations among those measures. This interpretation is undoubtedly correct. What is not correct is the further conclusion that structures such as those represented by five or three or seven factors, or circumplexes, or whatever are therefore useless or misleading characterizations of personality.

Factor analyses of personality measures are intended to simplify the complex observed relations among personality measures. Thus, it is not surprising that factor analytic solutions do not summarize all the variation and covariation among personality measures. The results of CFA are indicating that factor analytic models of personality simply do not capture all the complexity in human personality, but this is not their purpose. To adequately represent this complexity items would need to load on a number of factors (no simple structure); factors would need to correlate with each other (oblique rotations), and many small factors representing only one or two items might be required. Moreover, such structures might well be specific to a given sample and would not generalize. The cost of 'correctly' modeling personality would be the loss of the simplicity that the factor analysis was initially designed to provide. Certainly the factor analysis of personality measures is an undertaking where Whitehead's dictum, 'Seek simplicity but distrust it', applies.

CFA can still be a powerful tool for evaluating the relations among personality measures. The point of this discussion is simply that CFA should not be used to decide if a particular factor analytic model is 'correct'; as the model almost certainly is not correct because it is too simple. Rather, CFA should be used to compare models of personality by asking if adding more factors or correlations among factors significantly improves the fit of a model. That is, when the question is changed from, 'Is the model correct'?, to 'Which model is significantly better'? CFA can be a most appropriate tool. Finally, it is important to note that CFA also does not address the decision in factor analysis of personality measures that probably has the most crucial impact on the results. This is the initial decision about what variables are to be included in the analysis.

Summary

Factor analysis of personality measures has resulted in a wide variety of possible structures of human

personality. This variety results because personality psychologists have different theories about what constitutes the domain of personality and they have different views about the goals of factor analysis. In addition, different reasonable criteria exist for determining the number of factors and for rotating and naming those factors. Thus, the evaluation of any factor analysis must include not simply the end result, but all the decisions that were made on the way to achieving that result. The existence of many reasonable factor models of human personality suggests that people are diverse not only in their personality, but in how they perceive personality.

References

- [1] Allport, G.W. & Odbert, H.S. (1936). Trait names: a psycho-lexical study, *Psychological Monographs* **47**, 211.
- [2] Benet-Martinez, V. & John, O.P. (1998). Los Cincos Grandes across cultures and ethnic groups: multitrait multimethod analyses of the Big Five in Spanish and English, *Journal of Personality and Social Psychology* **75**, 729–750.
- [3] Cattell, R.B. (1995). Personality structure and the new fifth edition of the 16PF, *Educational & Psychological Measurement* **55**(6), 926–937.
- [4] Digman, J.M. (1997). Higher-order factors of the Big Five, *Journal of Personality and Social Psychology* **73**, 1246–1256.
- [5] Eysenck, H.J. & Eysenck, S.B.G. (1991). *Manual of the Eysenck Personality Scales (EPQ Adults)*, Hodder and Stoughton, London.
- [6] Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research, *Psychological Methods* **3**, 272–299.
- [7] Goldberg, L.R. & Velicer, W.F. (in press). Principles of exploratory factor analysis, in *Differentiating Normal and Abnormal Personality*, 2nd Edition, S. Strack, ed., Springer, New York.
- [8] Hofstee, W.K.B., de Raad, B. & Goldberg, L.R. (1992). Integration of the Big Five and circumplex approaches to trait structure, *Journal of Personality and Social Psychology* **63**, 146–163.
- [9] John, O.P. (1990). The “Big Five” factor taxonomy: dimensions of personality in the natural language and in questionnaires, in *Handbook of Personality: Theory and Research*, L.A. Pervin, ed., Guilford Press, New York, pp. 66–100.
- [10] McCrae, R.R. & Costa Jr, P.T. (1996). Toward a new generation of personality theories: theoretical contexts for the five-factor model, in J.S. Wiggins, ed., *The Five-Factor Model of Personality: Theoretical Perspectives*, Guilford Press, New York, pp. 51–87.
- [11] Norman, W. (1963). Toward an adequate taxonomy of personality attributes, *Journal of Abnormal and Social Psychology* **66**(6), 574–583.
- [12] Spearman, C. (1904). “General intelligence” objectively determined and measured, *American Journal of Psychology* **15**, 201–293.
- [13] Thurstone, L.L. (1934). The vectors of mind, *Psychological Review* **41**, 1–32.
- [14] Wiggins, J.S. (1982). Circumplex models of interpersonal behavior in clinical psychology, in *Handbook of Research Methods in Clinical Psychology*, P.C. Kendall & J.N. Butcher, eds, Wiley, New York, pp. 183–221.

WILLIAM F. CHAPLIN

Factor Score Estimation

SCOTT L. HERSHBERGER

Volume 2, pp. 636–644

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factor Score Estimation

Introduction

Factor analysis is concerned with two problems. The first problem is concerned with determining a factor pattern matrix based on either the **principal components analysis** or the common factor model. Factor loadings in the pattern matrix indicate how highly the observed variables are related to the principal components or the common factors, both of which can be thought of as **latent variables**. The second problem is concerned with estimating latent variable scores for each case. Latent variable scores, commonly referred to as *factor scores*, are useful and often necessary. Consider that the number of observed variables may be large; obtaining the (typically fewer) factor scores facilitates subsequent analyses. To cite another example, factor scores – at least when derived under the common factor model – are likely to be more reliable than observed scores. Related to the idea of higher reliability is the belief that a factor score is a pure, *univocal* (homogenous) measure of a latent variable, while an observed score may be ambiguous because we do not know what combination of latent variables may be represented by that observed score.

A number of methods have been proposed for obtaining factor scores. When these methods are applied to factors derived under the principal components model, the scores are ‘exact’, exact in the sense that a unique set of factor scores can be found for the principal components that are supposed to denote their true population values. It does not matter whether scores are derived for all n components, or only for some m ($m \leq n$) of them. In contrast, factor scores are not uniquely determinable for the factors of the common factor model: An infinite number of sets of factor scores are possible for any one set of common factors and thus, their true values must be estimated. Factor score indeterminacy arises from the indeterminacy of the common factor model itself.

Principal Component Scores

Factor scores computed for a set of principal components – henceforth to be referred to as *principal component scores* – are straightforwardly calculated.

As noted above, this is true no matter how many of the n possible principal components are retained.

In order to describe principal component scores, we begin with a matrix equation for a single case in which only m of the principal components have been retained:

$$\mathbf{Z}_{n \times 1} = \mathbf{A}_{n \times m} \mathbf{F}_{m \times 1}, \quad (1)$$

where $\mathbf{Z}$ is an $n \times 1$ column vector of n standardized observed variables, $\mathbf{A}$ is an $n \times m$ pattern matrix of the loadings of n observed variables on the m principal components, and $\mathbf{F}$ is an $m \times 1$ column vector of m principal component scores. The principal component scores are given by

$$\begin{aligned} \mathbf{F}_{m \times 1} &= \mathbf{A}_{n \times m}^{-1} \mathbf{Z}_{n \times 1} \\ \mathbf{A}' \mathbf{Z} &= \mathbf{A}' \mathbf{A} \mathbf{F} \\ &= (\mathbf{A}' \mathbf{A})^{-1} \mathbf{A}' \mathbf{Z} \\ &= \Lambda'_{D_{n \times n}} \mathbf{A}_{m \times n}^{-1} \mathbf{Z}_{n \times 1}, \end{aligned} \quad (2)$$

where λ_D is an $m \times m$ diagonal matrix of m eigenvalues. Equation (2) implies that a principal component score is constructed in the following way. First, each of the n loadings (symbolized by a) from the principal component's column in pattern matrix $\mathbf{A}$ is divided by the eigenvalue (λ) of the principal component (i.e., a/λ). Second, a/λ is multiplied by the score of the observed variable z associated with the loading (i.e., $a/\lambda \times z$). And then third, the n $a/\lambda \times z$ terms are summed, constructing the principal component f from their linear combination:

$$f_k = \sum_{j=1}^n \frac{a_{jk}}{\lambda_k} \times z_j, \quad (3)$$

where a_{jk} is the loading for the j th observed variable ($j = 1, 2, \dots, n$) on the k th principal component ($k = 1, 2, \dots, m, m \leq n$), and λ_k is the eigenvalue of the k th principal component.

For example, assume we have retained three principal components from eight observed variables:

$$\mathbf{A} = \begin{bmatrix} .71 & .11 & .16 \\ .82 & .15 & .20 \\ .93 & .19 & .24 \\ .10 & .77 & .28 \\ .22 & .88 & .32 \\ .24 & .21 & .36 \\ .28 & .23 & .71 \\ .39 & .32 & .77 \end{bmatrix} \quad (4)$$

2 Factor Score Estimation

and the eight observed scores for a person are

$$\mathbf{z} = \begin{bmatrix} .10 \\ .22 \\ -.19 \\ -.25 \\ .09 \\ .23 \\ .15 \\ -.19 \end{bmatrix}. \quad (5)$$

The three eigenvalues are, respectively, 2.39, 1.64, and 1.53. The first, second, and third principle component scores are calculated as

$$\begin{aligned} f_1 &= .04 = \left(\frac{.71}{2.39} \times .10 \right) + \left(\frac{.82}{2.39} \times .22 \right) \\ &\quad + \left(\frac{.93}{2.39} \times -.19 \right) + \left(\frac{.10}{2.39} \times -.25 \right) \\ &\quad + \left(\frac{.22}{2.39} \times .09 \right) + \left(\frac{.24}{2.39} \times .23 \right) \\ &\quad + \left(\frac{.28}{2.39} \times .15 \right) + \left(\frac{.39}{2.39} \times -.19 \right) \\ f_2 &= -.05 = \left(\frac{.11}{1.64} \times .10 \right) + \left(\frac{.15}{1.64} \times .22 \right) \\ &\quad + \left(\frac{.19}{1.64} \times -.19 \right) + \left(\frac{.77}{1.64} \times -.25 \right) \\ &\quad + \left(\frac{.88}{1.64} \times .09 \right) + \left(\frac{.21}{1.64} \times .23 \right) \\ &\quad + \left(\frac{.23}{1.64} \times .15 \right) + \left(\frac{.32}{1.64} \times -.19 \right) \\ f_3 &= .01 = \left(\frac{.16}{1.53} \times .10 \right) + \left(\frac{.20}{1.53} \times .22 \right) \\ &\quad + \left(\frac{.24}{1.53} \times -.19 \right) + \left(\frac{.28}{1.53} \times -.25 \right) \\ &\quad + \left(\frac{.32}{1.53} \times .09 \right) + \left(\frac{.36}{1.53} \times .23 \right) \\ &\quad + \left(\frac{.71}{1.53} \times .15 \right) + \left(\frac{.77}{1.53} \times -.19 \right). \end{aligned} \quad (6)$$

Component scores can be computed using either the unrotated pattern matrix or the rotated pattern matrix; both are of equivalent statistical validity. The scores obtained using the rotated matrix are simply

rescaled transformations of scores obtained using the unrotated matrix.

Common Factor Scores

Why are Common Factor Scores Indeterminate?

Scores from the common factor model are estimated because it is mathematically impossible to determine a unique set of them – an infinite number of such sets exist. This results from the *underidentification* of the common factor model (*see Factor Analysis: Exploratory; Identification*). An underidentified model is a model for which not enough information in the data is present to estimate all of the model's unknown parameters. In the principal components model, identification is achieved by imposing two restrictions: (a) the first component accounts for the maximum amount of variance possible, the second the next, and so on and so forth, and (b) the components are uncorrelated with each other. Imposing these two restrictions, the unknown parameters in the principal components model – the $n \times m$ factor loadings – can be uniquely estimated. Thus, the principal components model is identified: The $n \times m$ factor loadings to be estimated are $\leq$ in number to the $n(n+1)/2$ correlations available to estimate them.

In contrast, even with the imposition of the two restrictions, the common factor model remains underidentified for the following reason. The model postulates not only the existence of m common factors underlying n variables, requiring the specification of $n \times m$ factor loadings (as in the principal components model), it also postulates the existence of n specific factors, resulting in a model with $(n \times m) + n$ parameters to be estimated, greater in number than the $n(n+1)/2$ available to estimate them. As a result, the $n \times m$ factor loadings have an infinite number of possible values. Logically then, the factor scores would be expected to have an infinite number of possible values.

Methods for Estimating Common Factor Scores

Estimation by Regression

Thomson [9] was the first to suggest that ordinary least-squares regression methods (*see Least Squares Estimation*) can be used to obtain estimates of factor

scores. The information required to find the regression weights for the factors on the observed variables – the correlations among the observed variables and the correlation between the factors and observed variables – is available from the factor analysis. The least-squares criterion is to minimize the sum of the squared differences between predicted and true factor scores, which is analogous to the generic least-squares criterion of minimizing the sum of the squared differences between predicted and observed scores.

We express the linear regression of any factor f on the observed variables z in matrix form for one case as

$$\hat{\mathbf{F}}_{1 \times m} = \mathbf{Z}_{1 \times n} \mathbf{B}_{n \times m}, \quad (7)$$

where $\mathbf{B}$ is a matrix of weights for the regression of the m factors on the n observed variables, and $\hat{\mathbf{F}}$ is a row vector of m estimated factor scores.

When the common factors are orthogonal, we use the following matrix equation to obtain $\mathbf{B}$:

$$\mathbf{B}_{n \times m} = \mathbf{R}_{n \times n}^{-1} \mathbf{A}_{n \times m}, \quad (8)$$

where $\mathbf{R}$ is the matrix of correlations between the n observed variables. If the common factors are nonorthogonal, we also require Φ , the correlation matrix among the m factors:

$$\mathbf{B}_{n \times m} = \mathbf{R}_{n \times n}^{-1} \mathbf{A}_{n \times m} \Phi_{m \times m}. \quad (9)$$

To illustrate the regression method, we perform a common factor analysis on a set of six observed variables, retaining nonorthogonal two factors. We use data from three cases and define

$$\mathbf{Z} = \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix},$$

$$\mathbf{R} = \begin{bmatrix} 1.00 & & & & & \\ 0.31 & 1.00 & & & & \\ 0.48 & 0.54 & 1.00 & & & \\ 0.69 & 0.31 & 0.45 & 1.00 & & \\ 0.34 & 0.30 & 0.26 & 0.41 & 1.00 & \\ 0.37 & 0.41 & 0.57 & 0.39 & 0.38 & 1.00 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix},$$

$$\Phi = \begin{bmatrix} 1.00 & \\ 0.45 & 1.00 \end{bmatrix}. \quad (10)$$

On the basis of (9), the regression weights are

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} 1.00 & & & & & \\ 0.31 & 1.00 & & & & \\ 0.48 & 0.54 & 1.00 & & & \\ 0.69 & 0.31 & 0.45 & 1.00 & & \\ 0.34 & 0.30 & 0.26 & 0.41 & 1.00 & \\ 0.37 & 0.41 & 0.57 & 0.39 & 0.38 & 1.00 \end{bmatrix}^{-1} \\ &\quad \times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 1.00 & \\ 0.45 & 1.00 \end{bmatrix} \\ &= \begin{bmatrix} 2.05 & & & & & \\ -0.01 & 0.48 & & & & \\ -0.41 & -0.64 & 1.99 & & & \\ -1.18 & -0.01 & -0.20 & 2.11 & & \\ -0.09 & -0.20 & 0.16 & -0.35 & 1.32 & \\ -0.02 & -0.14 & -0.70 & -0.12 & -0.33 & 1.64 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 1.00 & \\ 0.45 & 1.00 \end{bmatrix} \\ &= \begin{bmatrix} 0.65 & -0.19 \\ 0.33 & -0.01 \\ 0.33 & 0.10 \\ -0.29 & 0.42 \\ 0.22 & 0.54 \\ -0.14 & 0.01 \end{bmatrix} \quad (11) \end{aligned}$$

Then, based on (7), the two-factor scores for the three cases are

$$\begin{aligned} \hat{\mathbf{F}} &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 0.65 & -0.19 \\ 0.33 & -0.01 \\ 0.33 & 0.10 \\ -0.29 & 0.42 \\ 0.22 & 0.54 \\ -0.14 & 0.01 \end{bmatrix} \end{aligned}$$

4 Factor Score Estimation

$$= \begin{bmatrix} -0.69 & -0.72 \\ -0.44 & 0.68 \\ 1.14 & 0.04 \end{bmatrix} \quad (12)$$

The regression estimates have the following properties:

1. The multiple correlation between each factor score and the common factors is maximized.
2. Each factor score estimate $\hat{f}$ is uncorrelated with its own residual $f - \hat{f}$ and the residual of every other estimate.
3. Even when the common factors are orthogonal, the estimates $\hat{f}$ are mutually correlated.
4. Even when the common factors are orthogonal, the estimate $\hat{f}$ of one factor can be correlated with any of the other $m - 1$ common factors.
5. Factor scores obtained through regression are biased estimates of their population values.

Depending on one's point of view as to what properties factor scores should have, properties 3 and 4 may or may not be problematic. If one believes that the univocality of a factor is diminished when it is correlated with another factor, then estimating factor scores by regression is considered a significantly flawed procedure. According to this view, univocality is compromised when variance in the factor is in part due to the influence of other factors. According to an alternative view, if in the population factors are correlated, then their estimated scores should be as well.

Minimizing Unique Factors

Bartlett [2] proposed a method of factor score estimation in which the least-squares criterion is to minimize the difference between the predicted and unique factor scores instead of minimizing the difference between the predicted and 'true' factor scores that is used in regression estimation. Unlike the regression method, Bartlett's method produces a factor score estimate that only correlates with its own factor and not with any other factor. However, correlations among the estimated scores of different factors still remain. In addition, Bartlett estimates, again unlike regression estimates, are unbiased. This is because they are **maximum likelihood estimates** of the population factor scores: It is assumed that the unique factor scores are multivariate normally

distributed (*see Catalogue of Probability Density Functions*).

Bartlett's method specifies that for one case

$$\mathbf{Z}_{1 \times n} = \hat{\mathbf{F}}_{1 \times m} \mathbf{A}_{m \times n} + \hat{\mathbf{V}}_{1 \times n} \mathbf{U}_{n \times n}, \quad (13)$$

where $\mathbf{Z}$ is a row vector of n observed variables scores, $\hat{\mathbf{F}}$ is a column vector of m estimated factor scores, $\mathbf{A}$ is the factor pattern matrix of loadings for the n observed variables on the m factors, $\hat{\mathbf{V}}$ is a row vector of n estimated unique scores, and $\mathbf{U}$ is a diagonal matrix of the standard deviations of the n unique factors. The common factor analysis provides both $\mathbf{A}$ and $\mathbf{U}$.

Recalling that $\hat{\mathbf{F}}_{1 \times m} = \mathbf{Z}_{1 \times n} \mathbf{B}_{n \times m}$ (1), we obtain the factor score weight matrix $\mathbf{B}$ to estimate the factor scores in $\hat{\mathbf{F}}$:

$$\begin{aligned} \mathbf{B}_{n \times m} &= \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m} (\mathbf{A}'_{m \times n} \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m})^{-1} \cdot \\ \hat{\mathbf{F}}_{1 \times m} &= \mathbf{Z}_{1 \times n} \mathbf{B}_{n \times m} \\ &= \mathbf{Z}_{1 \times n} \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m} (\mathbf{A}'_{m \times n} \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m})^{-1}. \end{aligned} \quad (14)$$

$\mathbf{U}^{-2}$ is the inverse of a diagonal matrix of the variances of the n unique factor scores. Using the results from (14), we can obtain the unique factor scores with

$$\hat{\mathbf{V}}_{1 \times n} = \mathbf{Z}_{1 \times n} \mathbf{U}_{n \times n}^{-1} - \hat{\mathbf{F}}_{1 \times m} \mathbf{A}_{m \times n} \mathbf{U}_{n \times n}^{-1}. \quad (15)$$

For our example of Bartlett's method, we define $\mathbf{U}^2$ as a diagonal matrix of the n unique factor variances,

$$\mathbf{U}^2 = \begin{bmatrix} 0.41 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.61 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.43 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.39 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.54 \end{bmatrix}, \quad (16)$$

and $\mathbf{U}$ as a diagonal matrix of the n unique factor standard deviations,

$$\mathbf{U} = \begin{bmatrix} 0.64 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.78 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.66 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.62 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.87 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \end{bmatrix}. \quad (17)$$

Following (14), the factor score weight matrix is obtained by

$$\begin{aligned}
 \mathbf{B} &= \begin{bmatrix} 0.41 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.61 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.43 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.39 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.54 \end{bmatrix}^{-2} \\
 &\times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \left(\begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \right)' \\
 &\times \begin{bmatrix} 0.41 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.61 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.43 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.39 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.54 \end{bmatrix}^{-2} \\
 &\times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix}^{-1} \\
 &= \begin{bmatrix} 2.44 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.64 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 2.38 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 2.63 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.33 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.85 \end{bmatrix} \\
 &\times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix}^{-1} \\
 &= \begin{bmatrix} 1.59 & 0.46 \\ 0.97 & 0.13 \\ 1.40 & 0.40 \\ 0.31 & 1.89 \\ 0.19 & 0.93 \\ 0.44 & 0.63 \end{bmatrix} \\
 &\times \begin{bmatrix} 0.48 & -0.23 \\ -0.23 & 0.52 \end{bmatrix} \\
 &= \begin{bmatrix} 0.66 & -0.17 \\ 0.44 & -0.15 \\ 0.59 & -0.11 \\ -0.28 & 0.92 \\ -0.12 & 0.45 \\ 0.07 & 0.23 \end{bmatrix}. \tag{18}
 \end{aligned}$$

With $\mathbf{B}$ defined as above, the estimated common factor scores for the three cases are

$$\begin{aligned}
 \hat{\mathbf{F}} &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \\
 &\times \begin{bmatrix} 0.66 & -0.17 \\ 0.44 & -0.15 \\ 0.59 & -0.11 \\ -0.28 & 0.92 \\ -0.12 & 0.45 \\ 0.07 & 0.23 \end{bmatrix} \\
 &= \begin{bmatrix} -0.85 & -0.49 \\ -0.69 & 0.45 \\ 1.54 & 0.04 \end{bmatrix}, \tag{19}
 \end{aligned}$$

and from (15), the unique factor scores for the three cases are

6 Factor Score Estimation

$$\begin{aligned}
\hat{\mathbf{V}} &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \begin{bmatrix} 0.64 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.78 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.66 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.62 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.87 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \end{bmatrix}^{-1} \\
&\quad - \begin{bmatrix} -0.85 & -0.49 \\ -0.69 & 0.45 \\ 1.54 & 0.04 \end{bmatrix} \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 0.64 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.78 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.66 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.62 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.87 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.73 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \begin{bmatrix} 1.56 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.28 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.51 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.61 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.50 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.37 \end{bmatrix} \\
&\quad - \begin{bmatrix} -0.85 & -0.49 \\ -0.69 & 0.45 \\ 1.54 & 0.04 \end{bmatrix} \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 1.56 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.28 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.51 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.61 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.50 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.37 \end{bmatrix} \\
&= \begin{bmatrix} 0.00 & -0.73 & -1.74 & 0.00 & -1.29 & -1.19 \\ -1.56 & -0.73 & 0.86 & 0.00 & 0.92 & -0.29 \\ 1.56 & 1.47 & 0.86 & 0.00 & 0.36 & 1.49 \end{bmatrix} - \begin{bmatrix} -0.41 & -0.42 & -0.39 & -0.28 & -0.40 & -0.27 \\ -0.23 & -0.28 & -0.22 & 0.15 & 0.19 & -0.01 \\ 0.64 & 0.71 & 0.60 & 0.13 & 0.21 & 0.28 \end{bmatrix} \\
&= \begin{bmatrix} 0.41 & -0.31 & -1.39 & 0.28 & -0.89 & -0.92 \\ -1.33 & -0.44 & 1.09 & -0.15 & 0.73 & -0.27 \\ 0.92 & 0.76 & 0.28 & -0.13 & 0.16 & 1.21 \end{bmatrix} \tag{20}
\end{aligned}$$

Bartlett factor score estimates can always be distinguished from regression factor score estimates by examining the variance of the factor scores. While regression estimates have variances ≤ 1 , Bartlett estimates have variances ≥ 1 [7]. This can be explained as follows: The regression estimation procedure divides the factor score f into two uncorrelated parts, the regression part $\hat{f}$ and the residual part $f - \hat{f}$.

Thus,

$$f = \hat{f} + e, \tag{21}$$

where

$$e = f - \hat{f}. \tag{22}$$

Since the e are assumed multivariate normally distributed, the $\hat{f}$ can further be written as the sum

of the factor score f and a residual ε :

$$\hat{f} = f + \varepsilon, \tag{23}$$

where

$$\varepsilon = \hat{f} - f. \tag{24}$$

The result is that the variance of $\hat{f}$ is the sum of the unit variance of f and the variance of ε , the error about the true value.

Uncorrelated Scores Minimizing Unique Factors

Anderson and Rubin [1] revised Bartlett's method so that factor score estimates are both uncorrelated the $m - 1$ with the other factors and are not correlated with each other. These two properties result

from the following matrix equation for the factor score estimates:

$$\hat{\mathbf{F}} = \mathbf{Z}_{1 \times n} \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m} (\mathbf{A}_{m \times n}' \mathbf{U}_{n \times n}^{-2} \Phi_{n \times n} \mathbf{U}_{n \times n} \mathbf{A}_{n \times m})^{-1/2}, \quad (25)$$

where Φ is a matrix of factor correlations.

While resembling (14), (25) is substantially more complex to solve: The term, $\mathbf{A}_{m \times n}' \mathbf{U}_{n \times n}^{-2} \Phi_{n \times n} \mathbf{U}_{n \times n} \mathbf{A}_{n \times m}$, is raised to a power of $-1/2$. This power indicates that the inversion of the symmetric square root of the matrix product is required. The symmetric square root of a matrix can be found for any positive definite symmetric matrix. To illustrate, we define $\mathbf{G}$ as an $n \times n$ positive semidefinite symmetric matrix. The symmetric square root of $\mathbf{G}$, $\mathbf{G}^{1/2}$, must meet the following condition:

$$\mathbf{G}_{n \times n} = \mathbf{G}_{n \times n}^{1/2} \mathbf{G}_{n \times n}^{1/2}. \quad (26)$$

Perhaps the most straightforward method of obtaining $\mathbf{G}^{1/2}$ is to obtain the spectral decomposition of $\mathbf{G}$, such that $\mathbf{G}$ can be reproduced by a function of its eigenvalues (λ) and eigenvectors (x):

$$\mathbf{G}_{n \times n} = \mathbf{X}_{n \times n} \Lambda_{D_{n \times n}} \mathbf{X}_{n \times n}', \quad (27)$$

where $\mathbf{X}$ is an $n \times n$ matrix of eigenvectors and Λ_D is an $n \times n$ diagonal matrix of eigenvalues. It follows then that

$$\mathbf{G}_{n \times n}^{1/2} = \mathbf{X}_{n \times n} \Lambda_{D_{n \times n}}^{1/2} \mathbf{X}_{n \times n}'. \quad (28)$$

If we set $\mathbf{G}^{1/2} = \mathbf{A}_{m \times n}' \mathbf{U}_{n \times n}^{-2} \Phi_{n \times n} \mathbf{U}_{n \times n} \mathbf{A}_{n \times m}$, (25) can now be rewritten as

$$\hat{\mathbf{F}} = \mathbf{Z}_{1 \times n} \mathbf{U}_{n \times n}^{-2} \mathbf{A}_{n \times m} \mathbf{G}^{-1/2}. \quad (29)$$

To illustrate the Anderson and Rubin method, we specify

$$\mathbf{X} = \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix} \quad (30)$$

and

$$\Lambda = \begin{bmatrix} 23.73 & 0.00 \\ 0.00 & 1.64 \end{bmatrix}. \quad (31)$$

Then, for $\mathbf{A}_{m \times n}' \mathbf{U}_{n \times n}^{-2} \Phi_{n \times n} \mathbf{U}_{n \times n} \mathbf{A}_{n \times m}$, the spectral decomposition is

$$\mathbf{G} = \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix} \begin{bmatrix} 23.73 & 0.00 \\ 0.00 & 1.64 \end{bmatrix} \times \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix}' \quad (32)$$

and therefore

$$\begin{aligned} \mathbf{G}^{1/2} &= \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix} \begin{bmatrix} 23.73 & 0.00 \\ 0.00 & 1.64 \end{bmatrix}^{1/2} \\ &\quad \times \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix}' \\ &= \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix} \begin{bmatrix} 4.87 & 0.00 \\ 0.00 & 1.28 \end{bmatrix} \begin{bmatrix} 0.74 & -0.67 \\ 0.67 & 0.74 \end{bmatrix}' \\ &= \begin{bmatrix} 3.26 & 1.79 \\ 1.79 & 2.89 \end{bmatrix}. \end{aligned} \quad (33)$$

Then, from Equation

$$\begin{aligned} \hat{\mathbf{F}} &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 0.41 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.61 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.43 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.39 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.54 \end{bmatrix}^{-2} \\ &\quad \times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 3.26 & 1.79 \\ 1.79 & 2.89 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 2.44 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.64 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 2.38 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 2.63 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.33 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.85 \end{bmatrix} \\ &\quad \times \begin{bmatrix} 0.65 & 0.19 \\ 0.59 & 0.08 \\ 0.59 & 0.17 \\ 0.12 & 0.72 \\ 0.14 & 0.70 \\ 0.24 & 0.34 \end{bmatrix} \begin{bmatrix} 0.46 & -0.29 \\ -0.29 & 0.52 \end{bmatrix} \\ &= \begin{bmatrix} 0.00 & -0.57 & -1.15 & 0.00 & -1.12 & -0.87 \\ -1.00 & -0.57 & 0.57 & 0.00 & 0.80 & -0.21 \\ 1.00 & 1.15 & 0.57 & 0.00 & 0.32 & 1.09 \end{bmatrix} \end{aligned}$$

$$\times \begin{bmatrix} 0.60 & -0.21 \\ 0.41 & -0.21 \\ 0.53 & -0.19 \\ -0.40 & 0.90 \\ -0.18 & 0.43 \\ 0.03 & 0.20 \end{bmatrix} = \begin{bmatrix} -0.67 & -0.32 \\ -0.68 & 0.53 \\ 1.35 & -0.20 \end{bmatrix}. \quad (34)$$

The unique factor scores are computed as in the Bartlett method (15). Substituting the results of the Anderson and Rubin method into (15) yields

$$\hat{\mathbf{V}} = \begin{bmatrix} 0.32 & -0.40 & -1.44 & 0.19 & -1.01 & -0.99 \\ -1.34 & -0.45 & 1.07 & -0.18 & 0.68 & -0.30 \\ 1.03 & 0.86 & 0.36 & -0.01 & 0.33 & 1.31 \end{bmatrix} \quad (35)$$

Conclusion

For convenience, we reproduce the factor scores estimated by the regression, Bartlett, and Anderson and Rubin methods:

Regression	Bartlett
$\begin{bmatrix} -0.69 & -0.72 \\ -0.44 & 0.68 \\ 1.14 & 0.04 \end{bmatrix}$	$\begin{bmatrix} -0.85 & -0.49 \\ -0.69 & 0.45 \\ 1.54 & 0.04 \end{bmatrix}$
Anderson-Rubin	
$\begin{bmatrix} -0.67 & -0.32 \\ -0.68 & 0.53 \\ 1.35 & -0.20 \end{bmatrix}$	

The similarity of the factor score estimates computed by the three methods is striking.

This is in part surprising. Empirical studies have found that although the factor score estimates obtained from different methods correlate substantially, they often have very different values [8]. So

it would seem that the important issue is not which of the three estimation methods should be used, but whether any of them should be used at all due to factor score indeterminacy, implying that only principal component scores should be obtained. Readers seeking additional information on this area of controversy specifically, and factor scores generally, should consult, in addition to those references already cited [3–6, 10].

References

- [1] Anderson, T.W. & Rubin, H. (1956). Statistical inference in factor analysis, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 5, 111–150.
- [2] Bartlett, M.S. (1937). The statistical conception of mental factors, *British Journal of Psychology* 28, 97–104.
- [3] Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in the Behavioral and Life Sciences*, Plenum Press, New York.
- [4] Gorsuch, R.L. (1983). *Factor Analysis*, 2nd Edition, Lawrence Erlbaum Associates, Hillsdale.
- [5] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Edition revised, The University of Chicago Press, Chicago.
- [6] Lawley, D.N. & Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, American Elsevier Publishing, New York.
- [7] McDonald, R.P. (1985). *Factor Analysis and Related Methods*, Lawrence Erlbaum Associates, Hillsdale.
- [8] Mulaik, S.A. (1972). *The Foundations of Factor Analysis*, McGraw-Hill, New York.
- [9] Thomson, G.H. (1939). *The Factorial Analysis of Human Ability*, Houghton Mifflin, Boston.
- [10] Yates, A. (1987). *Multivariate Exploratory Data Analysis: A Perspective on Exploratory Factor Analysis*, State University of New York Press, Albany.

SCOTT L. HERSHBERGER

Factorial Designs

PHILIP H. RAMSEY

Volume 2, pp. 644–645

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Factorial Designs

A factorial design is one in which two or more treatments (or classifications for variables such as sex) are investigated simultaneously and, in the ideal case, all possible combinations of each treatment (or classification) occur together in the design. In a one-way design, we might ask whether two different drugs lead to a significant difference in average adjustment scores. In a factorial design, we might ask whether the two drugs differ in effectiveness and whether the effectiveness of the drugs changes when each drug is applied under the administration of four different dosage levels. The first independent variable is the type of drug (with two levels) and the second independent variable is the dosage (with four levels). This design would be a 2×4 factorial ANOVA (see **Analysis of Variance**) design.

A major advantage of a factorial design is that we can evaluate two independent variables in a single experiment, as illustrated in Table 1. Another advantage is that we can evaluate the interaction of the two independent variables.

An Example of a 2×4 ANOVA Design

Table 1 presents hypothetical data in which the mean is presented for adjustment scores under two drugs and four dosage levels. Each of the eight means is based on six observations and the error term is $MS_E = 4.0$. Higher adjustment scores indicate better adjustment. The level of significance is set at an *a priori* level of .01 for each effect in the design. The means in the margins of Table 1 show the overall effects of the two independent variables. Testing the significance of the differences among the overall means can be done with *F* tests, as described by various texts such as [2], [3], or [4].

Taking the two drugs as Factor A, it can be shown that the overall *F* test is $F(1, 40) = 84.92$,

$p < 0.0001$, which is significant at the .01 level. Thus, the adjustment mean of 6.375 for Drug 1 is significantly greater than the adjustment mean of 2.5 for Drug 2. The subjects in this experiment are better adjusted when treated with Drug 1 than Drug 2. Taking the four dosages as Factor B, it can be shown that the overall *F* test is $F(3, 40) = 23.67$, $p < 0.0001$, which is also significant at the 0.01 level. One relatively simple method of evaluating the four overall means of Factor B would be with orthogonal contrasts such as orthogonal polynomials (see **Multiple Comparison Procedures**). Orthogonal contrasts are described by [2], [3], or [4], and can be used in testing main effects in a factorial design just as in a one-way ANOVA. For the four overall means of Factor B, only the linear trend is significant. There is a significant tendency for adjustment to increase linearly with increases in dosage.

The test of the interaction between A and B in Table 1 can be shown ([2], [3], or [4]) to be $F(3, 40) = 6.92$, $p = 0.0007$, which is also significant at a 0.01 level.

Following a significant interaction, many researchers prefer to test simple main effects. That is, they would test the four means separately for Drug 1 and Drug 2. Applying orthogonal polynomials to the four means under Drug 1 would show a significant linear trend with $F(1, 40) = 84.17$, $p < 0.0001$, but no other significant effect. Applying orthogonal polynomials to the four means under Drug 2 would show a linear trend also significant at the .01 level with $F(1, 40) = 7.5$, $p = 0.0092$. With significant linear increases in adjustment for both drugs, such testing of simple main effects does not explain the significant interaction.

It is usually better to evaluate significant interactions with interaction contrasts as described by Boik [1]. In Table 1, it can be shown that the contrast identified by the linear trend for the four dosages is significantly greater under Drug 1 than under Drug 2. That is, the tendency for adjustment to increase linearly from 10 mg to 40 mg is significantly greater under Drug 1 than under Drug 2.

The present example illustrates the problem of testing simple main effects to explain significant interactions. The opposite problem of possible misleading interpretations of overall main effects in the presence of significant interactions is illustrated in Kirk [2, p. 370].

Table 1 Hypothetical data of means for a 2×4 design

	10 mg	20 mg	30 mg	40 mg	Overall
Drug 1	3.0	6.5	7.0	9.0	6.375
Drug 2	1.0	2.0	3.0	4.0	2.5
Overall	2.0	4.25	5.0	6.5	4.4375

2 Factorial Designs

References

- [1] Boik, R.J. (1979). Interactions, partial interactions, and interaction contrasts in the analysis of variance, *Psychological Bulletin* **86**, 1084–1089.
- [2] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole, Pacific Grove.
- [3] Maxwell, S.E. & Delaney, H.D. (1990). *Designing Experiments and Analyzing Data*, Wadsworth, Belmont.
- [4] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Designs*, McGraw-Hill, New York.

PHILIP H. RAMSEY

Family History Versus Family Study Methods in Genetics

AILBHE BURKE AND PETER MCGUFFIN

Volume 2, pp. 646–647

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Family History Versus Family Study Methods in Genetics

In family research, a sample of index cases, or probands, with the disorder of interest is ascertained and the pattern of morbidity is then investigated in first degree relatives (parents, offspring, siblings) second degree relatives (e.g., uncle, aunts, grandparents), third degree relatives (e.g., cousins) and perhaps more distant family members. Familial aggregation is indicated by a significantly higher rate of the disorder in relatives of the probands than in relatives of a selected control population, or than in the general population [2]. It is important to note that a key issue in this kind of research is making sure individuals are correctly categorized with regard to their affected status. Therefore, the sensitivity (proportion of affected relatives correctly identified as affected) and specificity (proportion of unaffected relatives correctly identified as unaffected) of the methods are important considerations.

The two methods for this kind of research are the family history method and the family study method. In the family history method, one or more family members, usually including the proband, are used as informants to provide information on the pedigree structure and which members are affected. This method offers good specificity, and requires relatively few resources. It provides information on all family members, even those who refuse to participate, are unavailable or are deceased, thus reducing sample bias. Further, it may be more accurate for certain types of socially deviant information that people might be unwilling to admit to in interview for example, substance abuse.

The primary concern in using the family history method relates to poor sensitivity. That is, there may be substantial underreporting of morbidity in family members. This may be due to some disorders being masked or unrecognized, such that the informant is unaware of illness in affected family members. Also, family members may vary in the quality of information they provide, for example, mothers of probands have been shown to be may be better informants than fathers [4]. There is also evidence that using more than one informant improves sensitivity.

Another important drawback is that, even if a relative is identified as having a disorder, it may not be possible to make a definite diagnosis on the basis of hearsay. Access to medical case notes (charts) may help overcome this.

In the family study method, affected status is determined by direct interview and examination of all available consenting relatives. Again, this can be supplemented by examining medical case notes. The major positive advantage of this kind of study is its generally superior sensitivity [1], although the specificity is much the same. It is however, more expensive and time consuming, and less convenient. Additionally, sample ascertainment will be incomplete due to absent, deceased, or refusing family members, thus introducing possible sampling bias because only the most cooperative family members are seen. The family study method may also have inferior sensitivity for 'socially undesirable' traits including substance abuse, antisocial behaviour, or some psychiatric disorders, as individuals may be unwilling to admit present or past symptoms in a direct interview.

Whether to use family history or family study method is an important decision in family research, due to the implications in terms of resources required and extent and accuracy of information obtained. The family history method may be best suited to exploratory studies and identifying interesting pedigrees, while the family study method would be better for studying interesting pedigrees in depth. In practice, family studies often include an element of the family history too in order to complete the information gaps and avoid a selection bias [3].

A final caution is that although both of these methods provide information on the familial clustering of a disorder, they cannot distinguish between genetic and shared environmental effects. These issues can only be addressed by performing twin and/or adoption studies [2].

References

- [1] Andreasen, N.C., Endicott, J., Spitzer, R.L. & Winokur, G. (1977). The family history method using diagnostic criteria, *Archives of General Psychiatry* **34**, 1229–1235.
- [2] Cardno, A. & McGuffin, P. (2002). Quantitative genetics, in *Psychiatric Genetics and Genomics*, P. McGuffin, M.J. Owen & I.I. Gottesman, eds, Oxford University Press, New York, 31–54.

2 Family History Versus Family Study Methods in Genetics

- [3] Heun R., Hardt J., Burkart, M. & Maier, W. (1996). (See also **Family Study and Relative Risk**)
Validity of the family history method in relatives of gerontopsychiatric patients, *Psychiatry Research* **62**, 227–238. AILBHE BURKE AND PETER MCGUFFIN
- [4] Szatmari P. & Jones M.B. (1999). The effects of misclassification on estimates of relative risk in family history studies, *Genetic Epidemiology* **16**, 368–381.

Family Study and Relative Risk

PETER MCGUFFIN AND AILBHE BURKE

Volume 2, pp. 647–648

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Family Study and Relative Risk

Our goal in the family study is to investigate the extent to which disorders aggregate in families. We can take as the starting point ascertainment of a series of index cases or probands. Preferably, this should be done in a systematic way, for example, on the basis of a consecutive series of patients referred to a particular clinical service or by linking a register of hospital admissions to a population register. Alternatively, ascertainment strategies may be aimed at detecting all the cases of a particular disorder in a defined population. This can have an advantage over clinical ascertainment, which may be biased in favor of probands with more severe forms of disorder and high levels of comorbidity (multiple coexisting disorders). However, other biases may operate in population-based studies. In particular, there are may be volunteer effects. Individuals who agreed to participate in studies tend to be better educated than nonvolunteers and are less likely to come from socially disadvantaged backgrounds. Furthermore, probands having severe disorders may actually be underrepresented because more severely ill individuals refuse to participate.

Relative risk is often used as a measure of the extent to which common disorders, which do not show obvious Mendelian patterns of inheritance, are familial. In general, the relative risk of a disorder, disease, death, or other outcome is the proportion of individuals exposed to the risk factor who are affected divided by the proportion of affecteds among those not exposed. More specifically, therefore, the relative risk for developing a disorder for individuals related to a proband with that disorder is given by:

Relative risk

$$\begin{aligned}
 & \frac{N \text{ of affected relatives of probands} \div}{N \text{ of all relatives of probands}} \\
 &= \frac{N \text{ of affected relatives of controls} \div}{N \text{ of all relatives of controls}} \\
 &= \frac{\text{Proportion of probands' relatives affected}}{\text{Proportion of controls' relatives affected}} \quad (1)
 \end{aligned}$$

Here, the controls will typically be individuals screened for absence of the disease being studied. However, an alternative which may in practice be

less troublesome to obtain, is a sample of unrelated subjects drawn from the general population who are not screened for the disorder, so that relative risk is estimated as:

$$\frac{\text{Proportion of probands' relatives affected}}{\text{Proportion of population sample affected}} \quad (2)$$

This will tend to give a lower estimate of relative risk because the relatives of healthy controls will tend to have a lower proportion of affecteds than that found in the population as a whole. The size of the relative risk in familial disorders tends to increase according to the degree of relatedness to the proband, such that first-degree relatives (e.g., offspring, siblings) who have half their genes shared with the proband tend to have a higher relative risk than second-degree relatives (e.g., nieces/nephews/grandchildren) who share a quarter of their genes, who in turn have a higher risk than third-degree relatives (e.g., cousins). The extent to which relative risk reduces as a function of decreasing genetic relatedness has been proposed as an indicator of whether multiple genes contributing to a disease interact as opposed to behaving additively [3]. However, it must be emphasized that we cannot infer that a disorder or trait is necessarily genetically influenced simply because it is familial. Traits that are almost certainly influenced by multiple factors, for example, career choice, may show a very high 'relative risk' and can even mimic Mendelian inheritance [1]. To tease apart the effects of genes and shared family environmental influences, we need to perform twin or adoption studies.

An important fundamental issue in family studies is what measure do we use for the proportion of relatives affected? The simplest answer is to use the lifetime **prevalence** of the disorder. That is, the number of family members who have ever been affected divided by the total number. However, a complication arises because complex disorders tend to have their first onset over a range of ages, a so-called period of risk. For adult onset disorders, some family members under investigation may not yet have entered the period of risk; others may have lived through the entire period without becoming affected, while others who are unaffected will still be within the period of risk. Clearly, only those who remain unaffected after having passed through the entire risk period can be classed as definitely unaffected. Therefore, lifetime prevalence underestimates the true lifetime risk

2 Family Study and Relative Risk

of the disorder. This problem can be addressed using an age correction, the most straightforward of which, originally proposed by Weinberg (the same Weinberg after whom the Hardy Weinberg equilibrium in population genetics is named) is to calculate a corrected denominator or *bezugsziffer* (BZ). The lifetime risk or morbidity risk (MR) of the disorder can be estimated as the number of affecteds (A) divided by the BZ, where the BZ is calculated as:

$$\sum_i w_i + A$$

and where w_i is the weight given to the i th unaffected individual on the basis of their current age. The simplest system of assigning weights, the shorter Weinberg method, is to give the weight of zero to those younger than the age of risk, a weight of a half to those within the age of risk, and weight of one to those beyond the age of risk. A more accurate modification devised by Erik Strömberg is to use an empirical age of onset distribution from a large separate sample, for example, a national registry of psychiatric disorders, to obtain the cumulative frequency of disorder over a range of age bands from which weights can be derived [4]. Unfortunately, national registry data are often unavailable and an alternative method is to take the age of onset distribution in the probands and transform it to a normal distribution, for example, using a log transform [2]. The log age for each unaffected relative can be converted to a standard score and a weight, the proportion of the period a risk that has been lived through, can be assigned by reference to the standard normal integral.

Another approach is to carry out life table analysis. The method most often used in family studies is called the *Weinberg morbidity table*, but essentially the method is the same as in the life table analysis performed in other spheres. The distribution of

survival times (or times to becoming ill) is divided into a number of intervals. For each of these, we can calculate the number and proportion of subjects who entered the interval unaffected and the number and proportion of cases that became affected during that interval, as well as the number of cases that were lost to follow-up (because they had died or had otherwise 'disappeared from view'). On the basis of these numbers and proportions, we can calculate the proportion 'failing' or becoming ill over a certain time interval that is usually taken as the entire period of risk. A further alternative is to use a Kaplan Meier product limit estimator. This allows us to estimate the survival function (*see Survival Analysis*) directly from continuous survival or failure times instead of classifying observed survival times into a life table. Effectively, this means creating a life table in which each time interval contains exactly one case. It therefore has an advantage over a life table method in that the results do not depend on grouping of the data.

References

- [1] McGuffin, P. & Huckle, P. (1990). Simulation of Mendelism revisited: the recessive gene for attending medical school, *American Journal of Human Genetics* **46**, 994–999.
- [2] McGuffin, P., Katz, R., Aldrich, J. & Bebbington, P. (1988). The Camberwell Collaborative Depression Study. II. Investigation of family members, *British Journal of Psychiatry* **152**, 766–774.
- [3] Risch, N. (1990). Linkage strategies for genetically complex traits. III: the effect of marker polymorphism analysis on affected relative pairs, *American Journal of Human Genetics* **46**, 242–253.
- [4] Slater, E. & Cowie, V. (1971). *The Genetics of Mental Disorders*, Oxford University Press, London.

PETER MCGUFFIN AND AILBHE BURKE

Fechner, Gustav T

HELEN ROSS

Volume 2, pp. 649–650

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fechner, Gustav T

Born: April 19, 1801, in Gross Särchen, Germany.

Died: November 18, 1887, in Leipzig, Germany.

October 22 is celebrated as Fechner Day – the day in 1850 on which he claimed to have had the insight now known as Fechner’s Law: that is, that sensation increases as the logarithm of the stimulus intensity. However, his main contributions to statistics lie elsewhere.

Gustav Theodor Fechner was the second of five children of a Lutheran pastor, who died when he was 5 years old. He went to live with his uncle, also a pastor, and attended several schools. In 1817, he enrolled on a medical course at the University of Leipzig, obtaining his masters degree and title of doctor in 1823. One of his lecturers was E. H. Weber, after whom Fechner named Weber’s Law. Fechner stayed on in Leipzig to study physics and mathematics, and earned some money translating French science textbooks into German. He was appointed a lecturer in physics in 1824, professor in 1831, and given the professorial chair in 1834. He published many important physics papers. Fechner married A. W. Volkmann’s sister, Clara Maria, in 1833. He resigned his chair in 1840, because he had become ill, perhaps through overwork. He had also damaged his eyes, through staring at the sun for his research on subjective colors. He became reclusive, and turned his mind to philosophical and religious problems. In 1851, he published the *Zend-Avesta*, a philosophical work about the relation between mind and matter, which he regarded as two aspects of the same thing. The appendix to this book contained his program of psychophysics and a statement of his law (see [2]). He then undertook measurements of the Weber fraction to support his own law. He used his training in physics to develop three psychophysical test methods, and to give mathematical formulae for his theories. In 1860, he published his main psychophysical work, the *Elemente der Psychophysik* [3], only the first volume of which has

been translated [1]. The book influenced contemporary physiologists, who then investigated Weber’s Law in several senses. Fechner himself turned his attention to experimental aesthetics, which he studied from 1865 to 1876. He returned to psychophysics to answer some criticisms, publishing his *Revision* in 1882. He was a founder member of the Royal Saxon Academy of Sciences in 1846.

Fechner was a wide-ranging author, even writing satirical material under the pseudonym of Dr. Mises. From the point of view of statisticians, his main contributions lie in the development of probabilistic test procedures for measuring thresholds and in the associated mathematics. He used probability theory in the ‘Method of Right and Wrong Cases’, applying Gauss’s normal distribution of errors. He took h , the measure of precision, as an inverse measure of the differential threshold [$h = 1/(\sigma\sqrt{2})$]. He wrote several papers on statistics, but his important unfinished book on the topic, *Kollektivmasslehre* [*Measuring Collectives*] [4], was not published till 10 years after his death. Fechner believed in indeterminism, but reconciled this with the laws of nature through the statistics of probability distributions: distributions are lawful, while individual phenomena are not (see [5]). He also argued that some distributions were asymmetrical, a novel idea at the time. However, Fechner’s statistics never achieved the lasting fame of his psychophysical law.

References

- [1] Adler, H.T. (1966). *Gustav Fechner: Elements of Psychophysics*, Vol.1, Translated by, Howes, D.H. & Borning E.G., eds Holt, Rinehart & Winston, New York.
- [2] Fechner, G.T. & Scheerer, E. (1851/1987). Outline of a new principle of mathematical psychology (1851), *Psychological Research* **49**, 203–207. Transl.
- [3] Fechner, G.T. (1860). *Elemente der Psychophysik*, Breitkopf & Härtel, Leipzig. 2 vols.
- [4] Fechner, G.T. (1897). *Kollektivmasslehre*, in *Auftrage der Königlich Sächsischen Gesellschaft der Wissenschaften*, G.F. Lipps, ed., Engelmann, Leipzig.
- [5] Heidelberger, M. (2003). *Nature from within: Gustav Theodor Fechner’s Psychophysical Worldview*, Translated by Klohr, C. University of Pittsburgh Press, Pittsburgh.

HELEN ROSS

Field Experiment

PATRICIA L. BUSK

Volume 2, pp. 650–652

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Field Experiment

Experimental research studies are used when the researcher wishes to make causal inferences about the variables under investigation. Such studies can be designed to have high **internal validity**. Because of the controlled nature of the research, they may have weak **external validity**, that is, the results do not generalize to other individuals or settings. Sometimes, the experimental studies are criticized for being laboratory studies where the setting is artificial, because the subjects know that they are being studied, the researcher does not behave in a natural manner, and subjects may react to the researcher in an unnatural manner when they are not able to forget that they are participating in an experiment. Such studies may lack ecological validity, because participants are observed for what they can do and not what they would do normally. Other threats to external validity include volunteer bias, mortality effects, and a limited population. Finally, the results of laboratory experiments may be so limited that they can be found only by exact replication of the design.

If the above limitations to the laboratory study are of concern to the researcher and external validity is desired, then the study may be performed in the real world or a natural setting and includes the procedures of an experiment, that is, random assignment of participants to treatments or conditions, manipulated levels of the independent variable, control of extraneous variables, and reliable and valid measurement of the dependent variable. Such research investigations are called (randomized) field experiments.

As an example of a field study, consider the randomized experiment of Comer's School Development Program conducted in 23 middle schools in Prince George's County, Maryland [4]. The study was conducted over 4 years with 23 middle schools of which 21 were matched on the basis of racial composition and achievement test scores for 2 years prior to the study. These schools were assigned randomly to program or comparison status using a coin toss. The two additional schools were pilot program sites for a year before the study began. They were included as program schools, because no difference was found when these two schools were included or excluded from the analyses with the other 21 schools, which resulted in 13 experimental schools and 10 control schools. Repeated measurements were made with more than

12,300 students and 2000 staff, 1000 parents were surveyed, and schools records were accessed in the study. The middle-school program in Prince George's County is a 2-year one, which resulted in three cohorts of students. Even with the initial matching of schools on racial composition and achievement test scores, data were adjusted on the basis of three school-level covariates: the average socioeconomic status of students, their average elementary school California Achievement Test, and school size. The covariate adjustments served to reduce the error and not to correct initial bias, given the random assignment employed at the beginning of the study.

This study incorporated the elements of an experiment that resulted in internal validity, with its random assignment of Comer's School Development Program to schools, controlled extraneous variables by the matching of schools and the adjustments based on covariates, and measured the dependent variables in a reliable, sensitive, and powerful manner. Data were analyzed using **multiple linear regression**, school-level **multivariate analysis of variance**, school-level **analysis of variance**, and correlational analyses (*see Correlation Studies*). When results were statistically significant, magnitude of the effect was obtained. Because Cook et al. [4] conducted a field experiment, they were able to conclude that their findings apply to common, realistic situations and actually reflect natural behaviors.

Two examples of field experiments that employed selection are Tennessee's Project STAR, which involved an investigation into reduction in class size [8], and the Milwaukee Parental Choice Program that tested the use of school vouchers with a random selection of participants when there were more applicants to a particular school and grade than could be accommodated [9]. In the 1985–1986 school year, Project STAR included 128 small classes (of approximately 1900 students), 101 regular classes (of approximately 2300 students), and 99 regular classes with teacher aides (of approximately 2200 students). Details regarding the political and institutional origins of the randomized controlled trial on elementary-school class-size design can be found in Ritter and Boruch [8].

Not all field experiments involve such large numbers. Fuchs, Fuchs, Karns, Hamlett, and Katzaroff [5]

2 Field Experiment

examined the effects of classroom-based performance-assessment-driven instruction using 16 teachers who were assigned randomly to performance-assessment and nonperformance-assessment conditions. All of the teachers had volunteered to participate. Neither the teachers nor their classes were matched, but Fuchs et al. used inferential statistics to indicate that the teachers in the two conditions were comparable on demographic variables of years of teaching, class size, ethnicity, and educational level. Teachers completed a demographic information form reporting on information about each student in the class. Results of statistical tests revealed that the groups were comparable. Although the researchers were not able to control extraneous variables, they tested to assess whether extraneous variables could affect the outcome of the research.

Field experiments can be conducted with the general public and with selected groups. Each of these two types of studies with the general public has certain limitations. Field experiments that are conducted in an unrestricted public area in order to generalize to the typical citizen generally are studying social behaviors. Such investigations if conducted as laboratory experiments would reduce the reliability and validity of the results. The experiments with the general public generally are carried out in one of two ways: individuals can be targeted and their responses observed to a condition of an environmental independent variable or the researcher or a confederate creates a condition by approaching the public and exhibiting a behavior to elicit a response. Some of the limitations of field studies with the public are that the situations are contrived, external validity is limited to situations similar to those in the study, and random selection is not possible in that the sample is a convenient one depending on the individuals who are present in the location at the time of the study. Albas and Albas [1] studied personal-space behavior while conducting a fictitious poll by measuring how far the participant would stop from the pollster. They manipulated the factors of the meeting occurring in the safety of a shopping mall versus a less safe city park and of whether the pollster made eye contact or did not because of wearing dark glasses.

The other approach to field experiments involves studying a specific group of participants that exists already. Some examples would be studying young children at a day-care center or elderly individuals at a senior center. Sometimes the researcher is not

able to disguise the study, especially if the researcher must provide instructions; other times the researcher is able to act unobtrusively. For example, if there is a one-way mirror at the day-care center, the researcher may be able to view the behavior of the children without their knowing that they are being observed. If teachers or supervisors will be serving as the researcher, they should be given explicit training, and the researcher should use a double-blind procedure. A double-blind procedure was used in the 1954 field trial of the Salk poliomyelitis vaccine. Both the child getting the treatment and the physician who gave the vaccine and evaluated the outcome were kept in ignorance of the treatment given [7].

Several references that can be consulted for additional details regarding field experiments are [2], [3], and [10]. Kerlinger [6], in his second edition, has a detailed discussion with examples of field experiments and field studies (see **Quasi-experimental Designs**).

References

- [1] Albas, D.C. & Albas, C.A. (1989). Meaning in context: the impact of eye contact and perception of threat on proximity, *The Journal of Social Psychology* **129**, 525–531.
- [2] Boruch, R.F. (1997). *Randomized Experiments for Planning & Evaluation: A Practical Guide*, Sage Publications, Thousand Oaks.
- [3] Boruch, R.F. & Wothke, W., eds (1985). *Randomized Field Experimentation*, Jossey-Bass, San Francisco.
- [4] Cook, T.D., Habib, F.-N., Phillips, M., Settersten, R.A., Shagle, S.C. & Degirmencioglu, S.M. (1999). Comer's school development program in Prince George's County, Maryland: a theory-based evaluation, *American Educational Research Journal* **36**, 543–597.
- [5] Fuchs, L.S., Fuchs, D., Karns, K., Hamlett, C.L. & Katzaroff, M. (1999). Mathematics performance assessment in the classroom: effects on teacher planning and student problem solving, *American Educational Research Journal* **36**, 609–646.
- [6] Kerlinger, F.N. (1973). *Foundations of Behavioral Research*, 2nd Edition, Holt, Rinehart and Winston, New York.
- [7] Meier, P. (1972). The biggest public health experiment ever: the 1954 field trial of 6th Salk poliomyelitis vaccine, in *Statistics: A Guide to the Unknown*, J.M. Tanur, F. Mosteller, W.H. Kruskal, R.F. Link, R.S. Pieters & G.R. Rissing, eds, Holden-Day, San Francisco, pp. 2–13.
- [8] Ritter, G.W. & Boruch, R.F. (1999). The political and institutional origins of a randomized controlled trial on

- elementary class size: Tennessee's project STAR, *Educational Evaluation and Policy Analysis* **21**, 111–125.
- [9] Rouse, C.E. (1998). Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program, *Quarterly Journal of Economics* **113**, 553–602.
- [10] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin Co., Boston.

PATRICIA L. BUSK

Finite Mixture Distributions

BRIAN S. EVERITT

Volume 2, pp. 652–658

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Finite Mixture Distributions

Early in the 1890s, Professor W.R. Weldon consulted statistician **Karl Pearson** about a set of measurements on the ratio of forehead to body length for 1000 crabs. A plot of the data (see Figure 1) showed that they were skewed to the right. Weldon suggested that the reason for this skewness might be that the sample contained representatives of two types of crabs, but, when the data were collected, they had not been labeled as such. This led Pearson to propose that the distribution of the measurements might be modeled by a weighted sum of two *normal distributions* (see **Catalogue of Probability Density Functions**), with the two weights being the proportions of the crabs of each type. This appears to be the first application of what is now generally termed a *finite mixture distribution*.

In mathematical terms, Pearson's suggestion that the distribution for the measurements on the crabs was of the form

$$f(x) = pN(x, \mu_1, \sigma_1) + (1 - p)N(x; \mu_2, \sigma_2) \quad (1)$$

where p is the proportion of a type of crab for which the ratio of forehead to body length has mean μ_1 and standard deviation σ_1 , and $(1 - p)$ is the proportion of a type of crab for which the corresponding values are μ_2 and σ_2 . In equation (1)

$$N(x; \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2\sigma_i^2}(x - \mu_i)^2\right] \quad (2)$$

The distribution in (1) will be bimodal if the two component distributions are widely separated or will simply display a degree of skewness when the separation of the components is not so great.

To fit the distribution in (1) to a set of data, its five parameters, $p, \mu_1, \sigma_1, \mu_2, \sigma_2$ have to be estimated from the data. Pearson, in his classic 1894 paper [10], devised a method that required the solution of a ninth-degree polynomial, a task (at the time) so computationally demanding that it lead Pearson to state

'the analytic difficulties, even for a mixture of two components are so considerable that it may be

questioned whether the general theory could ever be applied in practice to any numerical case'.

But Pearson did manage the heroic task of finding a solution, and his fitted two-component normal distribution is shown in Figure 1, along with the solution given by **maximum likelihood estimation** (see later), and also the fit given by a single normal distribution.

Estimating the Parameters in Finite Mixture Distributions

Pearson's original estimation procedure for the five parameters in (1) was based on the *method of moments*, but is now only really of historical interest (see **Estimation**). Nowadays, the parameters of a simple finite mixture model such as (1) or more complex examples with more than two components, or other than univariate normal components, would generally be estimated using a maximum likelihood approach, often involving the *EM algorithm*. Details are given in [4, 9, and 12].

In some applications of finite mixture distributions, the number of components distributions in the mixture is known *a priori* (this was the case for the crab data where two types of crabs were known to exist in the region from which the data were collected). But, finite mixture distributions can also be used as the basis of a *cluster analysis* of data (see **Hierarchical Clustering; k-means Analysis**), with each component of the mixture assumed to describe the distribution of the measurement (or measurements) in a particular cluster, and the maximum value of the estimated *posterior probabilities* of an observation being in a particular cluster being used to determine cluster membership (see [4]). In such applications, the number of components of the mixture (i.e., the number of clusters in the data) will be unknown and therefore will also need to be estimated in some way. This difficult problem is considered in [4 and 9] and see **Number of Clusters**.

Some Examples of the Application of Finite Mixture Distributions

A sex difference in the age of onset of schizophrenia was noted in [6]. Subsequently, it has been found to

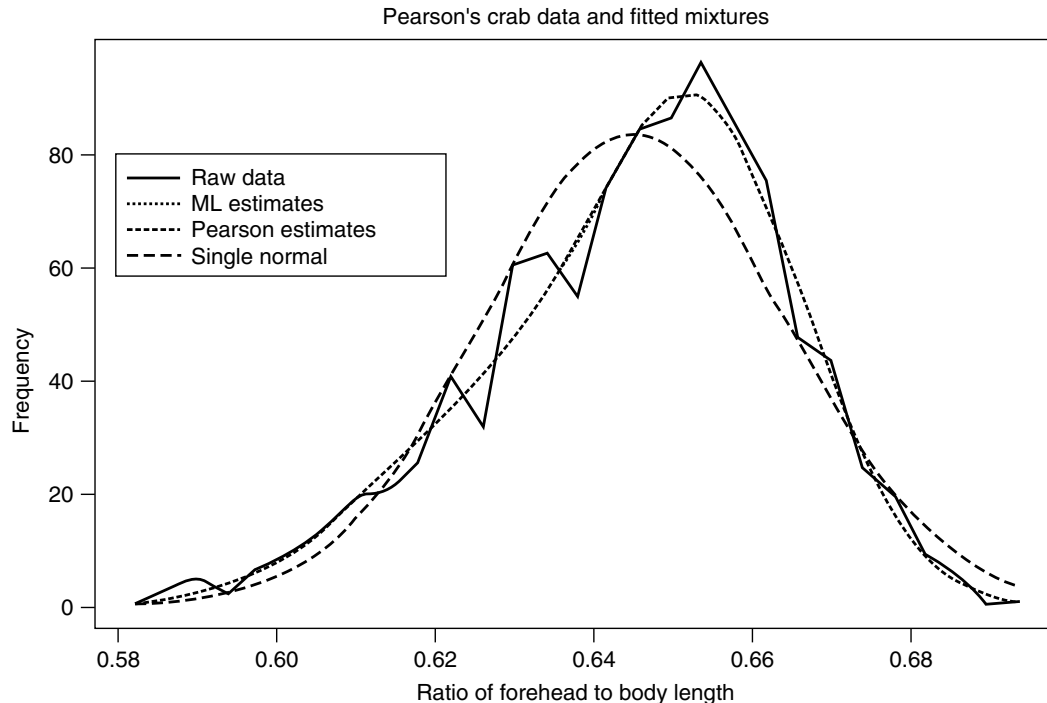


Figure 1 Frequency polygon of ratio of forehead to body length in 1000 crabs and fitted mixture and single normal distributions

be one of the most consistent findings in the epidemiology of the disorder. Levine [7], for example, collated the results of 7 studies on the age of onset of the illness, and 13 studies on age at first admissions, and showed that all these studies were consistent in reporting an earlier onset of schizophrenia in men than in women. Levine suggested two competing models to explain these data:

The timing model states that schizophrenia is essentially the same disorder in the two sexes, but has an early onset in men and a late onset in women ... In contrast with the timing model, the subtype model posits two types of schizophrenia. One is characterized by early onset, typical symptoms, and poor premorbid competence, and the other by late onset, atypical symptoms, and good premorbid competence ... the early onset typical schizophrenia is largely a disorder of men, and late onset, atypical schizophrenia is largely a disorder in women.

The subtype model implies that the age of onset distribution for both male and female schizophrenics

will be a mixture, with the mixing proportion for early onset schizophrenia being larger for men than for women. To investigate this model, finite mixture distributions with normal components were fitted to age of onset (determined as age on first admission) of 99 female and 152 male schizophrenics using maximum likelihood. The data are shown in Table 1 and the results in Table 2. **Confidence intervals** were obtained by using the *bootstrap* (see [2] and **Bootstrap Inference**). The bootstrap distributions for each parameter for the data on women are shown in Figure 2.

Histograms of the data showing both the fitted two-component mixture distribution and a single normal fit are shown in Figure 3.

For both sets of data, the *likelihood ratio test* for number of groups (see [8, 9], and **Maximum Likelihood Estimation**) provides strong evidence that a two-component mixture provides a better fit than a single normal, although it is difficult to draw convincing conclusions about the proposed subtype model of schizophrenia because of the very wide

Table 1 Age of onset of schizophrenia (years)**(1) Women**

20	30	21	23	30	25	13	19	16	25	20	25	27	43	6	21	15	26	23	21	23	23
34	14	17	18	21	16	35	32	48	53	51	48	29	25	44	23	36	58	28	51	40	43
21	48	17	23	28	44	28	21	31	22	56	60	15	21	30	26	28	23	21	20	43	39
40	26	50	17	17	23	44	30	35	20	41	18	39	27	28	30	34	33	30	29	46	36
58	28	30	28	37	31	29	32	48	49	30											

(2) Men

21	18	23	21	27	24	20	12	15	19	21	22	19	24	9	19	18	17	23	17	23	19
37	26	22	24	19	22	19	16	16	18	16	33	22	23	10	14	15	20	11	25	9	22
25	20	19	22	23	24	29	24	22	26	20	25	17	25	28	22	22	23	35	16	29	33
15	29	20	29	24	39	10	20	23	15	18	20	21	30	21	18	19	15	19	18	25	17
15	42	27	18	43	20	17	21	5	27	25	18	24	33	32	29	34	20	21	31	22	15
27	26	23	47	17	21	16	21	19	31	34	23	23	20	21	18	26	30	17	21	19	22
52	19	24	19	19	33	32	29	58	39	42	32	32	46	38	44	35	45	41	31		

Table 2 Age of onset of schizophrenia results of fitting finite mixture densities**(1) Women**

Parameter	Initial value	Final value	Bootstrap 95% CI*
p	0.5	0.74	(0.19, 0.83)
μ_1	25	24.80	(21.72, 27.51)
σ_1^2	10	42.75	(27.92, 85.31)
μ_2	50	46.45	(34.70, 50.50)
σ_2^2	10	49.90	(18.45, 132.40)

(2) Men

Parameter	Initial value	Final value	Bootstrap 95% CI ^a
p	0.5	0.51	(0.24, 0.77)
μ_1	25	20.25	(19.05, 22.06)
σ_1^2	10	9.42	(3.43, 36.70)
μ_2	50	27.76	(23.48, 34.67)
σ_2^2	10	112.24	(46.00, 176.39)

^aNumber of bootstrap samples used was 250.

confidence intervals for the parameters. Far larger sample sizes are required to get accurate estimates than those used here.

Identifying Activated Brain Regions

In [3], an experiment is reported in which functional magnetic resonance imaging (fMRI) data were collected from a healthy male volunteer during a visual stimulation procedure. A measure of the experimentally determined signal at each voxel in the image was calculated, as described in [1]. Under the null hypothesis of no experimentally determined signal change (no activation), the derived statistic has a chi-square distribution with two degrees of freedom (see **Catalogue of Probability Density Functions**). Under the presence of an experimental effect (activation), however, the statistic has a noncentral chi-squared distribution (see [3]). Consequently, it follows that the distribution of the statistic over all voxels in an image, both activated and nonactivated, can be modeled by a mixture of those two component densities (for details, again see [3]). Once the parameters of the assumed mixture distribution have been estimated, so can the probability of each voxel in the image being activated or nonactivated. For the visual simulation data, voxels were classified as ‘activated’ if their posterior probability of activation was greater than 0.5 and ‘nonactivated’ otherwise. Figure 4 shows the ‘mixture model activation map’

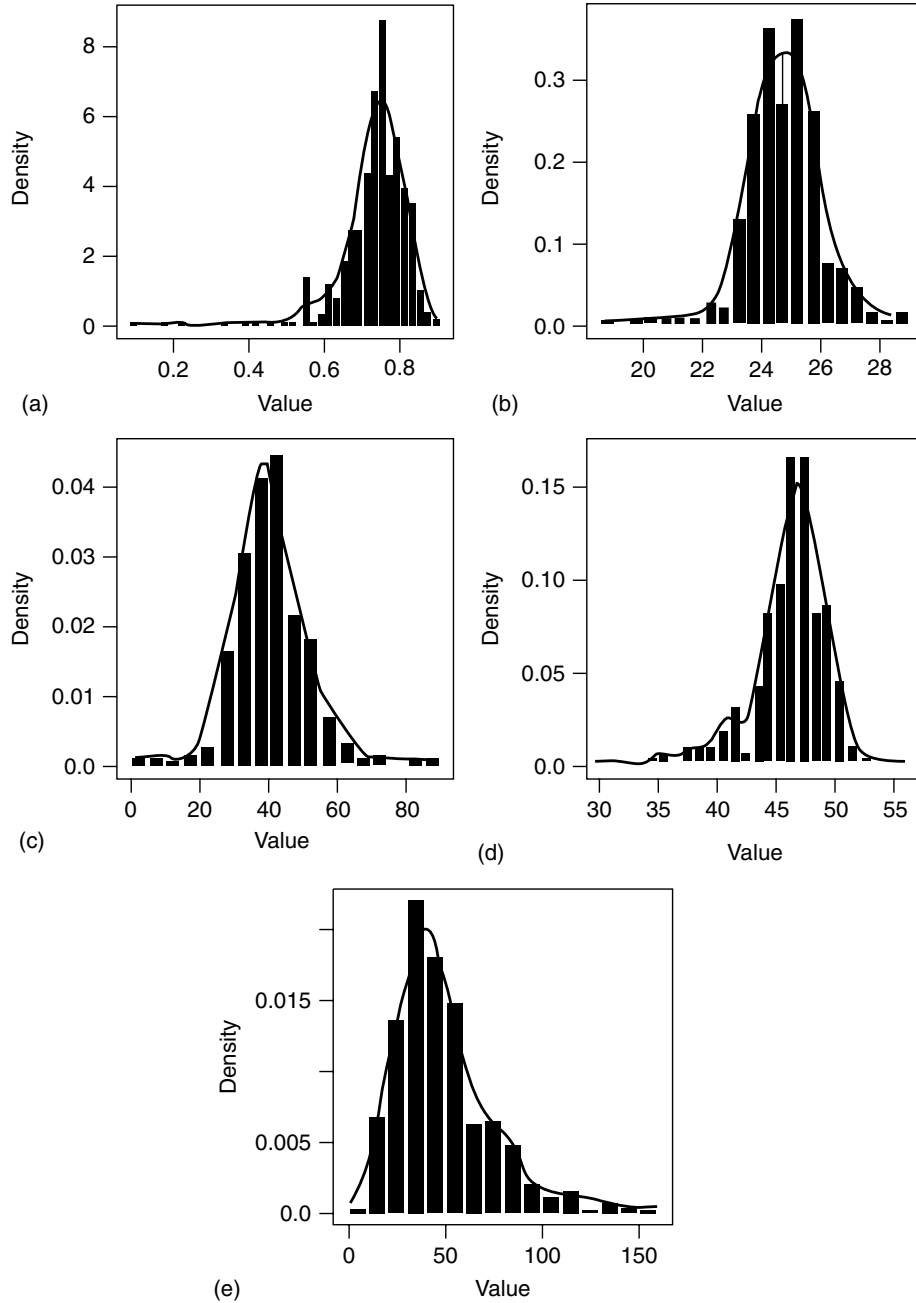


Figure 2 Bootstrap distributions for five parameters of a two-component normal mixture fitted to the age of onset data for women: (a) mixing proportion; (b) mean of first distribution; (c) standard deviation of first distribution; (d) mean of second distribution; (e) standard deviation of second distribution

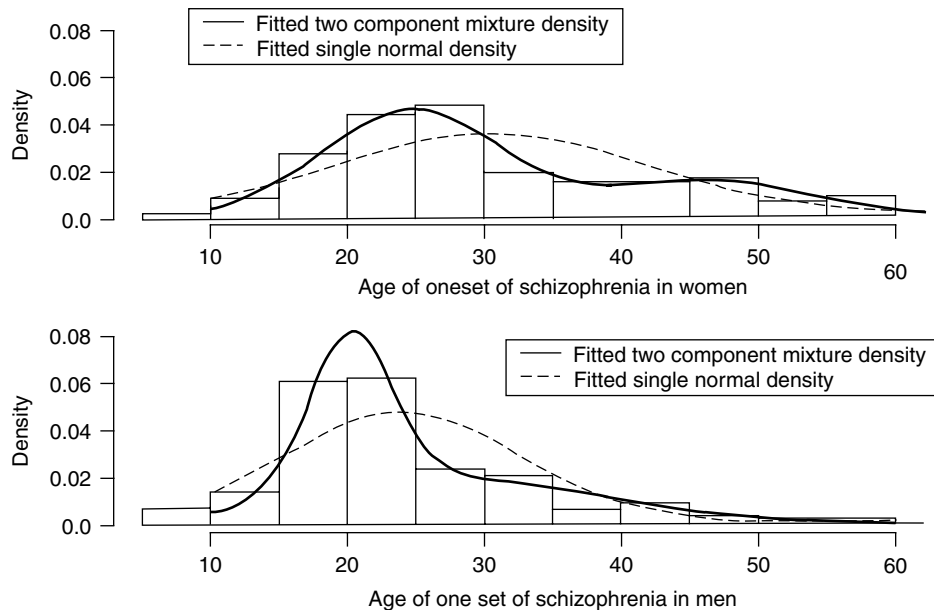


Figure 3 Histograms and fitted mixture distributions for age of onset data for women and men

of the visual simulation data for selected slices of the brain (activated voxels indicated).

Finite mixture models are now widely used in many disciplines. In psychology, for example, latent class analysis, essentially a finite mixture with multivariate Bernoulli components (*see Catalogue of Probability Density Functions*), is often used as a categorical data analogue of **factor analysis**. And, in medicine, mixture models have been successful in analyzing *survival times* (*see* [9] and **Survival Analysis**). Many other applications of finite mixture distributions are described in the comprehensive account of finite mixture distributions given in [9].

Software for Fitting Finite Mixture Distributions

The following sites provide information on software for mixture modeling:

NORMIX was the first program for clustering data that consists of mixtures of multivariate Normal distributions. The program was originally written John H. Wolfe in the 1960s [13]. A version that runs under MSDOS-Windows is available as freeware at <http://alumni.caltech.edu/~wolfe/normix.htm>.

Geoff McLachlan and colleagues have developed the EMMIX algorithm for the automatic fitting and testing of normal mixtures for multivariate data (*see* <http://www.maths.uq.edu.au/~gjm/>).

A further program is that developed by Jorgensen and Hunt [5] and the source code is available from Murray Jorgensen's website (<http://www.stats.waikato.ac.nz/Staff/maj.html>).

References

- [1] Bullmore, E.T., Brammer, M.J., Rouleau, G., Everitt, B.S., Simmons, A., Sharma, T., Frangou, S., Murray, R. & Dunn, G. (1995). Computerized brain tissue classification of magnetic resonance images: a new approach to the problem of partial volume artifact, *Neuroimage* **2**, 133–147.
- [2] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York.
- [3] Everitt, B.S. & Bullmore, E.T. (1999). Mixture model mapping of brain activation in functional magnetic resonance images, *Human Brain Mapping* **7**, 1–14.
- [4] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman & Hall/CRC, London.
- [5] Jorgensen, M. & Hunt, L.A. (1999). Mixture model clustering using the MULTIMIX program, *Australian and New Zealand Journal of Statistics* **41**, 153–171.

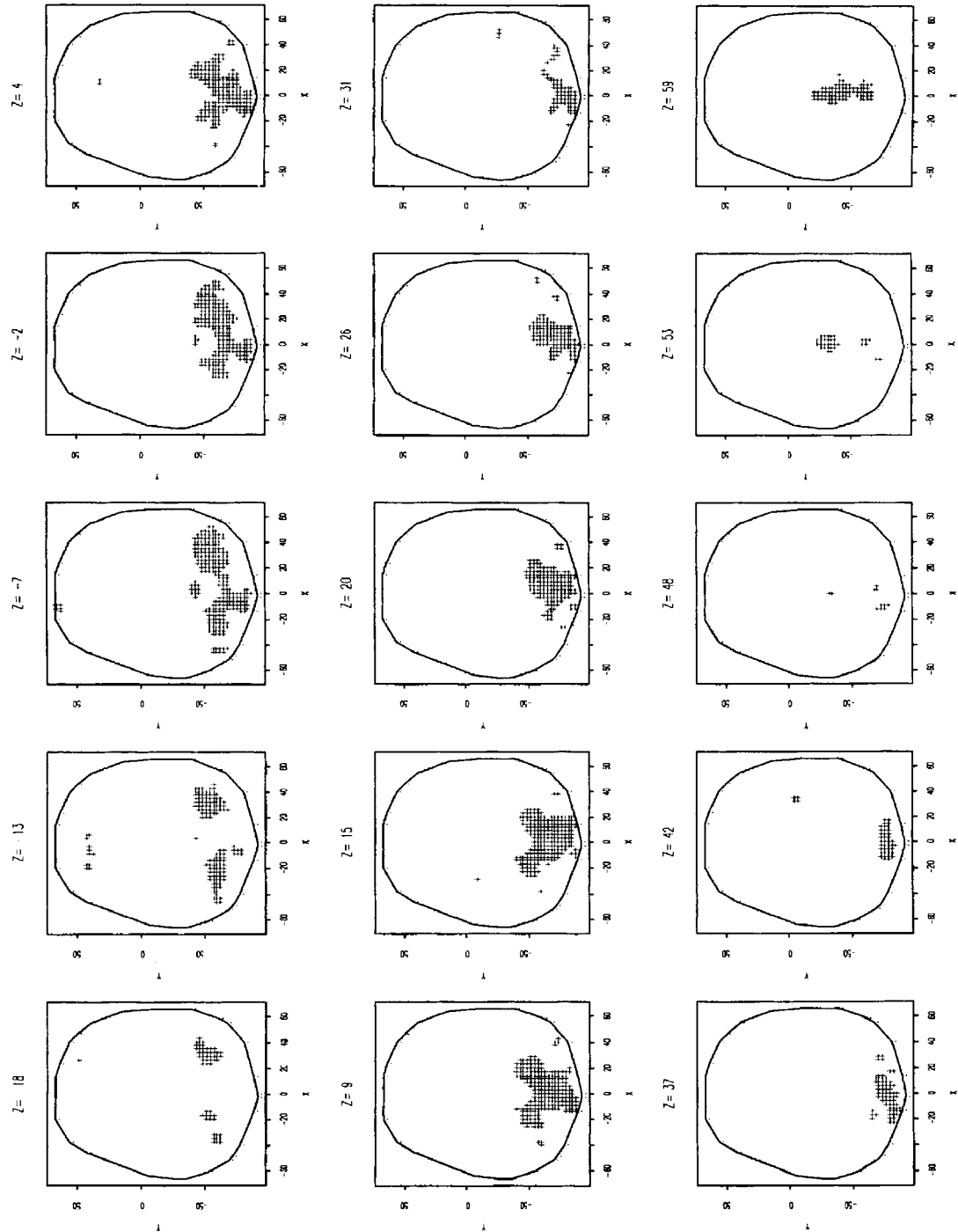


Figure 4 Mixture model activation map of visual simulation data. Each slice of data (Z) is displayed in the standard anatomical space described in [11]

-
- [6] Kraepelin, E. (1919). *Dementia Praecox and Paraphrenia*, Churchill Livingstone, Edinburgh.
- [7] Levine, R.R.J. (1981). Sex differences in schizophrenia: timing or subtypes? *Psychological Bulletin* **90**, 432–444.
- [8] McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Applied Statistics* **36**, 318–324.
- [9] McLachlan, G.J. & Peel, D. (2000). *Finite Mixture Distributions*, Wiley, New York.
- [10] Pearson, K. (1984). Contributions to the mathematical theory of evolution, *Philosophical Transactions A* **185**, 71–110.
- [11] Talairach, J. & Tournoux, P. (1988). *A Coplanar Stereotaxic Atlas of the Human Brain*, Thieme-Verlag, Stuttgart.
- [12] Titterton, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.
- [13] Wolf, J.H. (1970). Pattern clustering by multivariate mixture analysis, *Multivariate Behavioral Research* **5**, 329–350.

BRIAN S. EVERITT

Fisher, Sir Ronald Aylmer

ANDY P. FIELD

Volume 2, pp. 658–659

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fisher, Sir Ronald Aylmer

Born: February 17, 1890, in London, England.

Died: July 29, 1962, in Adelaide, Australia.

In Cambridge, in the late 1920s, a group of University dons and their wives and guests sat for afternoon tea. One woman asserted that the taste of tea depended on whether the milk was added to the cup before or after the tea. While the academics put the proposition down as sheer nonsense, a short, bearded man by the name of Ronald Fisher set about putting it to the test. This tale and its implications for statistics and experimentation are well documented (see [2, 3, 8]); however, it unveils much about Fisher as a person. He was a man fascinated by discovery and science. As well as being ‘perhaps the most original mathematical scientist of the (last) century’ (Efron, in [9, p. 483], he published widely on genetics, biology (notably inbreeding), and medicine (the links between smoking and cancer). Even within the domain of statistics, his interests were wide: Barnard [1, p. 162] concluded that ‘apart from the work done on multivariate problems . . . Fisher left little here for others to do’. For many, Fisher laid the foundations of modern statistics [1, 11].

Fisher was born in London in 1890, and after an education at Harrow public school, studied mathematics at Caius College, Cambridge, from where he graduated with the prestigious title of a Wrangler (the highest honor in mathematics) in 1912. He stayed at Cambridge to study physics following which he had a succession of jobs (the Mercantile and General investment Company, a farm in Canada, and a school teacher) before taking up a post at the Rothamsted Experimental Station in 1919. Although Fisher published several key papers during his time as a school teacher (including that on the distributional properties of the correlation coefficient), it was at Rothamsted that he produced his famous series of ‘Studies in Crop Variation’ in which he developed the **analysis of variance** and **analysis of covariance** and explored the importance of **randomization** in experimentation. In 1933, he became the Galton Professor of Eugenics at University College, London following **Karl Pearson’s** retirement; then in 1943 he became the Arthur Balfour Professor of Genetics at Cambridge University.

He retired in 1957, and in 1959 settled in Adelaide, Australia. His honors include being elected Fellow of the Royal Society (1929), the Royal Medal (1938), the Darwin Medal (1948), the Copley Medal (1955), and a knighthood (1952) (see [4] for an extensive biography).

Savage [9] suggests that it would be easier to list the areas of statistics to which Fisher took no interest, and this is evidenced by the breadth of his contributions. Savage [9] and Barnard [1] provide thorough reviews of Fisher’s contributions to statistics, but his most well known are in the development of (1) sampling theory and significance testing; (2) estimation theory (he developed **maximum likelihood estimation**); and (3) multivariate techniques such as Fisher’s **discriminant function analysis**. In addition, through his development of experimental designs and their analysis, and his understanding of the importance of randomization, he provided a scientific community dominated by idiosyncratic methodological practices with arguably the first blueprint of how to conduct research [8].

Although some have noted the similarities between aspects of Fisher’s work and his predecessor’s (for example, Stigler, [10], notes similarities with Francis Edgeworth’s work on analysis of two-way classifications and maximum likelihood estimators), Fisher’s contributions invariably went well beyond those earlier sources (see [9]). His real talent was in his intuition for statistics and geometry that allowed him to ‘see’ answers, which he regarded as obvious, without the need for mathematical proofs (although invariably the proofs, when others did them, supported his claims). This is not to say that Fisher was not a prodigious mathematician (he was); however, it did mean that, even when connected to other’s earlier work, his contributions had greater clarity and theoretical and applied insight [5].

Much is made of Fisher’s prickly temperament and the public disagreements he had with Karl Pearson and **Jerzy Neyman**. His interactions with Pearson got off on the wrong foot when Pearson published Fisher’s paper on the distribution of **Galton’s** correlation coefficient but added an editorial, which, to the casual reader, belittled Fisher’s work. The footing became more unstable when two years later Pearson’s group published extensive tables of this distribution without consulting Fisher [5]. Their relationship remained antagonistic with Fisher publishing ‘improvements’ on Pearson’s ideas, and Pearson

referring in his own journal to apparent errors made by Fisher [8]. With Neyman, the rift developed from a now infamous paper delivered by Neyman to the Royal Statistical Society that openly criticized Fisher's work [7]. Such was their antagonism that Neyman openly attacked Fisher's factorial designs and ideas on randomization in lectures while they both worked at University College, London. The two feuding groups even took afternoon tea (a common practice in the British academic community of the time) in the same room but at different times [6]. However, these accounts portray a one-sided view of all concerned, and it seems fitting to end with Irwin's (who worked with both K. Pearson and Fisher) observation that 'Fisher was always a wonderful conversationalist and a good companion. His manners were informal . . . and he was friendliness itself to his staff and any visitors' ([5, p. 160]).

References

- [1] Barnard, G.A. (1963). Ronald Aylmer Fisher, 1890–1962: Fisher's contributions to mathematical statistics, *Journal of the Royal Statistical Society, Series A (General)* **126**, 162–166.
- [2] Field, A.P. & Hole, G. (2003). *How to Design and Report Experiments*, Sage Publications, London.
- [3] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [4] Fisher Box, J. (1978). *R.A. Fisher: The Life of a Scientist*, Wiley, New York.
- [5] Irwin, J.O. (1963). Sir Ronald Aylmer Fisher, 1890–1962: Introduction, *Journal of the Royal Statistical Society, Series A (General)* **126**, 159–162.
- [6] Olkin, I. (1992). A conversation with Churchill Eisenhart, *Statistical Science* **7**, 512–530.
- [7] Reid, C. (1982). *Neyman-From Life*, Springer-Verlag, New York.
- [8] Salsburg, D. (2001). *The Lady Tasting Tea: How statistics Revolutionized Science in the Twentieth Century*, W.H. Freeman, New York.
- [9] Savage, L.J. (1976). On Re-reading R. A. Fisher, *The Annals of Statistics* **4**, 441–500.
- [10] Stigler, S.M. (1999). *Statistics on the Table: A history of statistical Concepts and Methods*, Harvard University Press, Cambridge.
- [11] Yates, F. (1984). Review of 'Neyman-From Life' by Constance Reid, *Journal of the Royal Statistical Society, Series A (General)* **147**, 116–118.

ANDY P. FIELD

Fisherian Tradition in Behavioral Genetics

DAVID DICKINS

Volume 2, pp. 660–664

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fisherian Tradition in Behavioral Genetics

R.A. Fisher was a major figure in both mathematics and biology: besides his major contribution to statistics he was greatly respected by other mathematicians. Through his applications of statistical theory to biology, Fisher achieved the integration of genetics¹ with evolution, which was the keystone of the evolutionary synthesis [15]. For three decades after the rediscovery in 1901 of the work of Mendel, enthusiastic supporters had felt that ‘Mendelism’ invalidated Darwin’s adoption of natural selection as the main mechanism of evolution. This was overthrown by Fisher’s demonstration that, on the contrary, Mendelian genetics provides both the source of appropriate variation on which selection can act, and the means for conserving and accumulating these favorable variations in the population.

Since the synthesis, most biologists, behavior scientists, in particular, have been more Darwinian than Darwin. Though modern geneticists are busy with postgenomic opportunities² to discover how specific genes actually exert their effects upon development and metabolism, and often take evolution for granted as a purely historical explanation, the logic of selection at the level of the gene offers important insights into such phenomena as genomic imprinting [19] and aging [4]. However, it is in the field of behavior that the pioneering work of Fisher has facilitated major advances in understanding. It has provided the heady theoretical underpinning of sociobiology [21] and its more politically correct descendant, evolutionary psychology [3]. Fisher’s contribution to the sister science of behavior genetics per se³ [18] was in quantitative genetics, particularly in the part played by estimating degrees of relatedness [12].

Fisher [13] powerfully analyzed the logical difficulties inherent for Darwin in his acceptance of blending inheritance. This assumed that the offspring of any novel variant would be more nearly a fusion of the variant and the other parent of normal type, with the problem that after a very few generations all trace of that variation would be lost. In such a situation, natural selection could not take hold. Darwin noted [6] that variation was greater in domesticated species, and thought this must

be due to novel environmental effects in captivity, though the undiminished variance of ancient domestic breeds implied the implausible requirement of a long history of continuously new environmental factors. He knew, of course, through his study of artificial selection, that dramatic transformation analogous to natural evolution could be produced in these varying domestic animals. Wild populations would only match this, and be especially subject to natural selection, in situations where they had entered markedly novel environments, or big changes had come about having similar effects to those of the environments of domestication. This rather went against Darwin’s conviction that evolution occurred through the very gradual and continuous accumulation of tiny changes.

The particulate theory of Mendelism explains how novelties, however they may arise initially, are not lost in the population, but appear perennially (unless natural selection takes a hand). The Hardy–Weinberg equation was used (but not referenced) by Fisher [13] to clarify this point. If two alternative alleles exist, one with a frequency of p in the gene pool, and the other with the complementary frequency of q (where $p + q = 1$), then the three genotypes resultant from random matings of individuals carrying these alleles in a large population will inexorably be in the ratios corresponding to the three terms of the equation:

$$p^2 + 2pq + q^2$$

with p^2 as the frequency of the homozygote for p , pq for the heterozygote, and q^2 for the homozygote for q .

Against this understood background, even slight selective advantages will produce gradual directional shifts, the ‘tiny wedges’ beloved of Darwin. The flamboyant variance of domestic species can be accounted for by the higher chances of survival of mutations that would be much less readily weeded out than in a natural population, together with their often positive promotion by artificial selection (like the ‘appealing’ features of many rather monstrous breeds of dogs, for example).

As for the notion, supported by De Vries, Bateson, and the mutationist Mendelians, that mutation could somehow drive evolution, Fisher lists the possible hypothetical mechanisms through which this might operate, for which no evidence exists, and most of which are inherently implausible. Even Weissman’s

notion that once mutations in a certain direction occurred, it was more likely that further ones in the same direction would follow, is shown to be redundant (and incorrect).

Fisher was fond of comments such as ‘Natural selection is a mechanism for generating an exceedingly high degree of improbability’ [11]. Now the idea of evolution being due to natural selection acting on mutations occurring by chance was repellent to many critics from the publication of Darwin’s *Origin of Species* [5]. Fisher explained, however, that this did not resemble an extraordinary run of good luck such as every client of a casino wishes he might enjoy, but the inexorable workings of the laws of chance over a longer sample upon which the profitability of such establishments depends. Just as Mendel had been influenced by the physicist von Ettinghausen to apply combinatorial algebra to his breeding experiments, so a postdoctoral year at Cambridge with James Jeans⁴ (after his Mathematics degree there) enthused Fisher with Maxwell’s statistical theory of gases, Boltzmann’s statistical thermodynamics, and quantum mechanics. He looked for an integration in the new physics of the novelty generation of biological evolution and the converse principle of entropy in nonliving systems [10].

Alongside these adjustments to post-Newtonian science, transcending the Newtonian scientific world view of Darwin, Fisher was in a sense the grandson of Darwin (see [10]), for he was much influenced by two of Charles’s sons, Horace, and particularly Leonard. This must have been very exciting to a young middle-class man who might just as well have taken a scholarship in biology at Cambridge, and took as the last of his many school prizes the complete set of Darwin’s books which he read and reread throughout his life, and with Leonard in particular he interchanged many ideas and was strongly encouraged by him. Now both these sons of such a renowned father and from such a prominent family were involved in the eugenics movement to which Fisher heartily subscribed⁵. There is a lot of this in the second half of his *The Genetical Theory of Natural Selection*. It is not clear how the logical Fisher reconciled his attachment to Nietzsche’s ideas of superior people who should endeavor to generate lots of offspring and be placed ‘beyond good or evil,’ with his stalwart adherence to Anglicanism⁶, but he certainly practiced what he preached by in

due course fathering eight children on his 17-year-old bride. There is inflammable material here for those whose opposition to sociobiology and evolutionary psychology is colored by their political sensitivity to the dangerous beasts of Nazism and fascism forever lurking in human society.

In his population genetics, Fisher shifted the emphasis from the enhanced chances of survival of favored individuals to the study, for each of the many alleles in a large population, of the comparative success of being duplicated by reproduction. In this calculus, alleles conferring very slight benefits (to their possessor) in terms of Darwinian fitness⁷ would spread in the gene pool at the expense of alternative alleles. This gave rise to important theoretical advances that were particularly influential in the study of behavior. The adaptive significance of the minutiae of behavior could in principle be assessed in this way, and in some cases measured in the field with some confidence about its validity. When Fisher was back in Cambridge as Professor of Genetics, one of his most enthusiastic students was William Hamilton, who from this notion of the gene as the unit of selection, derived inclusive fitness theory [14], popularly expounded by Richard Dawkins in [9] and a series of similar books. Expressing Fisher’s [12] degrees of genetic resemblance between closer and more distant relatives in terms of the probability, r , of a rare gene in one individual also occurring in a relative, Hamilton propounded that for an apparently altruistic social action, in which an actor appears to incur a cost C (in terms of Darwinian fitness) in the course of conferring a benefit B to a relative, the following equation applies:

$$rB - C > 0$$

This has become known as Hamilton’s rule, and it means that an allele predisposing an animal to help a relative will tend to spread if the cost to the ‘donor’ is less than the benefit to the recipient, downgraded by the degree of relatedness between the two.

This is because r is an estimate of the chances that the allele will indirectly make copies of itself via the progeny of the beneficiary, another way of spreading in the gene pool. This is also known as kin selection.

Robert Trivers [20] soon followed this up with the notion of reciprocal altruism in which similar altruistic acts could be performed by a donor for an unrelated recipient, provided that the cost was

small in relation to a substantial benefit, if the social conditions made it likely that roles would at some future time be liable to be reversed, and that the former recipient could now do a favor for the former donor. This would entail discrimination by the individuals concerned between others who were and others who were not likely to reciprocate in this way (probably on the basis of memory of past encounters). There are fascinating sequelae to these ideas, as when the idea of ‘cheating’ is considered, which Trivers does in his benchmark paper [20]. Such a cozy mutual set-up is always open to exploitation, either by ‘gross’ cheaters, who are happy to act as beneficiaries but simply do not deliver when circumstances would require them to act as donors, or by more subtle strategists who give substandard altruism, spinning the balance of costs and benefits in their own favor. These provide fertile ground for cheater-detection counter measures to evolve, an escalating story intuitively generative of many of the social psychological features of our own species.

The pages of high quality journals in animal behavior (such as *Animal Behaviour*) are today packed with meticulous studies conducted in the field to test the ideas of Hamilton and Trivers, and a corresponding flow of theoretical papers fine-tuning the implications.

Another key idea attributed to Fisher is the ‘run-away theory’ of sexual selection. This concerns male adornment, and Darwin’s complementary notion of female choice, long treated with skepticism by many, but now demonstrated across the animal kingdom. If it comes about that a particular feature of a male bird, for example, such as a longer than usual tail, attracts females more than tails of more standard length (which in some species can be demonstrated by artificial manipulation [1]), and both the anatomical feature and the female preference are under genetic control, then the offspring of resultant unions will produce sexy sons and size-sensitive females who in turn will tend to corner the mating market. This is likely to lead, according to Fisher [13], to further lengthening of the tail and strengthening of the preference, at an exponential rate, since the greater the change, the greater the reproductive advantage, so long as this is not outweighed by other selective disadvantages, such as dangerous conspicuousness of the males. Miller [17] has inverted this inference, and from fossil data supporting such a geometric increase

in brain size in hominids, has speculated that a major influence on human evolution has been because of the greater attractiveness to females of males with larger brains enabling them to generate a more alluring diversity of courtship behaviors. Other explanations, either alternative or complementary, have also been forwarded for the evolution of flamboyant attraction devices in male animals [22].

The evolutionary stabilization of the sex ratio – that except under special circumstances the proportion of males to females in a population will always approximate to 1 : 1 – is another fecund idea that has traditionally been attributed to Fisher. Actually the idea (like many another) goes back to Darwin, and to *The Descent of Man*. Like many of us Fisher possessed (and read and reread) the Second Edition of this book [8] in which Darwin backtracks in a quote I would have been critical of were it to occur in a student’s essay today⁸:

‘I formerly thought that when a tendency to produce the two sexes in equal numbers was advantageous to the species, it would follow from natural selection, but I now see that the whole problem is so intricate that it is safer to leave its solution for the future’ [8] (pp. 199–200).

Fisher [13] also quotes this, but gives an incorrect citation (there are no references in his book) as if it were from the first edition. In the first edition [7], Darwin does indeed essay ways in which the sex ratio might come under the influence of natural selection. He does not rate these effects of selection as a major force compared with ‘unknown forces’:

‘Nevertheless we may conclude that natural selection will always tend, though sometimes inefficiently, to equalise the relative numbers of the two sexes.’ [ibid. Vol. I, p. 318]

Then Darwin acknowledges Herbert Spencer, not for the above, but for the idea of what we would now call the balance between *r* and *K* selection. Darwin is unclear how fertility might be reduced by natural selection once it has been strongly selected, for direct selection, by chance, would always favor parents with more offspring in overpopulation situations, but the cost of producing more, to the parents, and the likely lower quality of more numerous offspring, would be ‘indirect’ selective influences reducing fertility in severely competitive conditions.

There is an anticipation here too of **Game Theory** which was developed by the late, great successor to Fisher as the (on one occasion self-styled!) ‘Voice of Neodarwinism’, John Maynard Smith [16]. The relevance of game theory is to any situation in which the adaptive consequences (Darwinian fitness) of some (behavioral or other) characteristic of an individual depend, not only on the environment in general, but upon what variants of this other members of the same species possess. Put succinctly, your best strategy depends on others’ strategies⁹. In the case of the sex ratio, if other parents produce lots of daughters, it is to your advantage to produce sons. In the case of sexual selection, if males with long tails are cornering the mating market, because females with a preference for long tails are predominant, it is to your advantage to produce sons with even longer tails. The combination of some of the theories here mentioned, such as game theory and the principle of reciprocal altruism [2] is an index of the potential of the original insights of Fisher. Iterative computer programs have made such subtle interactions easier to predict, and fruitfully theorize about than the unaided though brilliant mathematics of Fisher.

Notes

1. ‘Genetics, the study of the hereditary mechanism, and of the rules by which heritable qualities are transmitted from one generation to the next...’ Fisher, R.A. (1953) Population genetics (The Croonian Lecture). *Proceedings of the Royal Society, B*, 141, 510–523.
2. Work made possible by the delineation of the entire genome of mice and men and an increasing number of other species.
3. For sociobiology and evolutionary psychology, some degree of a hereditary basis for behavior is axiomatic. Behavior genetics seeks to demonstrate and analyze specific examples of this, both in animal breeding experiments and human familial studies, for practical as well as theoretical purposes.
4. It was Jeans who was once asked whether it was true that he was one of the only three people to understand relativity theory. ‘Who’s the third’? he allegedly asked.
5. I cherish the conviction that Charles was entirely egalitarian.
6. While Nietzsche clearly recognized Christian values as in direct opposition to his own.
7. Measured as the number of fertile offspring an individual produces which survive to sexual maturity.
8. The troublesome phrase here is ‘advantageous to the species.’ The point about the action of selection here is that it is the advantage to the genes of the individual that lead it to produce more male or more female offspring.
9. There can be interactions between species as well, as for example in arms races, for example, selection for ever increasing speed both of predator and prey in say cheetahs hunting antelope.

References

- [1] Andersson, M. (1982). Female choice selects for extreme tail length in a widowbird, *Nature* **299**, 818–820.
- [2] Axelrod, R. (1990). *The Evolution of Co-operation*, Penguin, London.
- [3] Barrett, L., Lycett, J. & Dunbar, R.I.M. (2002). *Human Evolutionary Psychology*, Palgrave, Basingstoke & New York.
- [4] Charlesworth, B. (2000). Fisher, Medawar, Hamilton and the evolution of aging, *Genetics* **156**(3), 927–931.
- [5] Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, 1st Edition, John Murray, London.
- [6] Darwin, C. (1868). *The Variation of Plants and Animals Under Domestication*, Vol. 1–2, John Murray, London.
- [7] Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*, John Murray, London.
- [8] Darwin, C. (1888). *The Descent of Man, and Selection in Relation to Sex*, 2nd Edition, John Murray, London.
- [9] Dawkins, R. (1989). *The Selfish Gene*, 2nd Edition, Oxford University Press, Oxford.
- [10] Depew, D.J. & Weber, B.H. (1996). *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*, The MIT Press, Cambridge & London.
- [11] Edwards, A.W.F. (2000). The genetical theory of natural selection, *Genetics* **154**(4), 1419–1426.
- [12] Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh* **222**, 309–368.
- [13] Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*, Oxford University Press, Oxford.
- [14] Hamilton, W.D. (1964). The genetical evolution of social behaviour I & II, *Journal of Theoretical Biology* **7**, 1–52.
- [15] Huxley, J. (1942). *Evolution, the Modern Synthesis*, Allen & Unwin, London.
- [16] Maynard Smith, J. (1982). *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.
- [17] Miller, G. (2000). *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*, Heinemann, London.
- [18] Plomin, R., DeFries, J.C., McGuffin, P. & McClearn, G.E. (2000). *Behavioral Genetics*, 4th Edition, Worth, New York.
- [19] Spencer, H.G., Clark, A.G. & Feldman, M.W. (1999). Genetic conflicts and the evolutionary origin of genomic

- imprinting, *Trends in Ecology & Evolution* **14**(5), 197–201.
- [20] Trivers, R.L. (1971). The evolution of reciprocal altruism, *Quarterly Review of Biology* **46**, 35–57.
- [21] Wilson, E.O. (1975). *Sociobiology: The New Synthesis*, Harvard University Press, Cambridge.
- [22] Zahavi, A. & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*, Oxford University Press, New York.

DAVID DICKINS

Fixed and Random Effects

TOM A.B. SNIJDERS

Volume 2, pp. 664–665

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fixed and Random Effects

In the specification of multilevel models (*see Linear Multilevel Models*), as discussed in [1] and [3], an important question is which explanatory variables (also called *independent variables* or *covariates*) to give random effects. A quantity being random means that it fluctuates over units in some population; and which particular unit is being observed depends on chance. When some effect in a statistical model is modeled as being random, we mean that we wish to draw conclusions about the population from which the observed units were drawn rather than about these particular units themselves.

The first decision concerning random effects in specifying a multilevel model is the choice of the levels of analysis. These levels can be, for example, individuals, classrooms, schools, organizations, neighborhoods, and so on. Formulated generally, a level is a set of units, or equivalently a system of categories, or a classification factor in a statistical design. In statistical terminology, a level in a multilevel analysis is a design factor with random effects. What does this mean?

The main point of view that qualifies a set of units, for example, organizations, as a level in the multilevel sense is that the researcher is not primarily interested in the particular organizations (units) in the sample but in the population (the wider set of organizations) for which the observed units are deemed to be representative. Statistical theory uses the word ‘exchangeability’, meaning that from the researcher’s point of view any unit in the population could have taken the place of each unit in the observed sample. (Whether the sample was drawn randomly according to some probability mechanism is a different issue – sometimes it can be argued that convenience samples or full population inventories can reasonably be studied as if they constituted a random sample from some hypothetical larger population.) It should be noted that what is assumed to be exchangeable are the residuals (sometimes called *error terms*) associated with these units, which means that any fixed effects of explanatory variables are already partialled out.

In addition, to qualify as a nontrivial level in a multilevel analysis, the dependent variable has to show some amount of residual, or unexplained, variation, associated with these units: for example, if the

study is about the work satisfaction (dependent variable) of employees (level-one units) in organizations (level-two units), this means that employees in some organizations tend to have a higher satisfaction than in some other organizations, and the researcher cannot totally pin this down to the effect of particular measured variables. The researcher being interested in the population means here that the researcher wants to know the amount of residual variability, that is, the residual variance, in average work satisfaction within the population of organizations (and perhaps also in the more complicated types of residual variability discussed below). If the residual variance is zero, then it is superfluous to use this set of units as a level in the multilevel analysis.

A next decision in specifying a multilevel model is whether the explanatory variables considered in a particular analysis have fixed or random effects. In the example, such a variable could be the employee’s job status: a level-one variable, since it varies over employees, the level-one units. The vantage point of multilevel analysis is that the effect of job status on work satisfaction (i.e., the regression coefficient of job level) could well be different across organizations. The fixed effect of this variable is the average effect in the entire population of organizations, expressed by the regression coefficient. Since mostly it is not assumed that the average effect of an interesting explanatory variable is exactly zero, almost always the model will include the fixed effect of all explanatory variables under consideration. When the researcher wishes to investigate differences between organizations in the effect of job level on work satisfaction, it will be necessary to specify also a random effect of this variable, meaning that it is assumed that the effect varies randomly within the population of organizations, and the researcher is interested in testing and estimating the variance of these random effects across this population. Such an effect is also called a *random slope*.

When there are no theoretical or other prior guidelines about which variables should have a random effect, the researcher can be led by the substantive focus of the investigation, the empirical findings, and parsimony of modeling. This implies that those explanatory variables that are especially important or have especially strong effects could be modeled with random effects, if the variances of these effects are important enough as evidenced by their significance and size, but one should take care that

2 Fixed and Random Effects

the number of variables with random effects should not be so large that the model becomes unwieldy.

Modeling an effect as random usually – although not necessarily – goes with the assumption of a normal distribution for the random effects. Sometimes, this is not in accordance with reality, which then can lead to biased results. The alternative, entertaining models with nonnormally distributed residuals, can be complicated, but methods were developed, see [2]. In addition, the assumption is made that the random effects are uncorrelated with the explanatory variables. If there are doubts about normality or independence for a so-called nuisance effect, that is, an effect the researcher is interested in not for its own sake but only because it must be statistically controlled for, then there is an easy way out. If the doubts concern the main effect of a categorical variable, which also would be a candidate for being modeled as a level as discussed above, then the easy solution (at least for linear models) is to model this categorical control variable by fixed effects, that is, using **dummy variables** for the units in the sample. If it is a random slope for which such a statistical control is required without making the assumption of residuals being normally distributed and independent

of the other explanatory variables, then the analog is to use an interaction variable obtained by multiplying the explanatory variable in question by the dummy variables for the units. The consequence of this easy way out, however, is that the statistical generalizability to the population of these units is lost (see **Generalizability**).

References

- [1] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods*, 2nd Edition, Sage Publications, Newbury Park.
- [2] Seltzer, M. & Choi, K. (2002). Model checking and sensitivity analysis for multilevel models, in *Multilevel Modeling: Methodological Advances, Issues, and Applications*, N., Duan & S., Reise, eds, Lawrence Erlbaum, Hillsdale.
- [3] Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publishers, London.

(See also **Random Effects and Fixed Effects Fallacy**)

TOM A.B. SNIJDERS

Fixed Effect Models

ROBERT J. VANDENBERG

Volume 2, pp. 665–666

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fixed Effect Models

Typically, a term associated with the application of **analysis of variance in experimental designs**, fixed-effect refers to a type of independent variable or factor. Most researchers employing experimental designs use fixed-effect models in that the levels of the independent variables are finite, known, and in most cases, manipulated by the researcher to be reasonably different to detect a difference on the dependent variable if there is any internal validity to the independent variable. Examples include gender (two-level factor), comparing the impact of the only three drugs known to address a specific condition (three-level factor), or creating different experimental conditions such as receiving negative performance feedback from a same race versus a different race confederate rating source (three-level factor). Fixed-effect independent variables differ from random factors in experimental designs. Random factors are independent variables that consist of a pool of potential levels, and thus, the levels are not manipulated by the researcher but rather sampled. For example, assume that a researcher is interested in examining the impact of a particular management practice on nurse retention in hospitals of a given state. The fixed factor would be the presence and absence of the management practice, but since it is highly improbable that the researcher can undertake this experiment in all hospitals, a random sample (usually conducted systematically such as employing stratified schemes) of hospitals would be selected. Hospitals would be the random factor.

A common criticism of purely fixed-effect designs, particularly those where the levels of the independent variable are the design of the researcher, is the fact that the results may only be generalized to comparable levels of that independent variable

within the population. In contrast, results from purely random factor models presumably are generalizable because the levels included in the experiment were randomly selected, and hence, will be representative of the population at large. As indicated in the hospital example above, if a random factor is included, it is typically crossed with a fixed-effect factor, which constitutes a mixed model design. While the technical explanation for the following is provided in the attached references, the inclusion of random factors complicates the analysis in that the error term for evaluating effects (main or interaction) is different. For example, assume we have two fixed-effect independent variables, A and B. The significance of their main and interaction effects are evaluated against the same source of error variance; the within-group mean square error. If A were a random factor, however, the main effect of B (the fixed factor) would need to be evaluated against a different source of error variance because the interaction effect intrudes upon the main effect of the fixed-effect factor. Not making the adjustment could result in detecting a main effect for B when none exists in reality.

Further Reading

- Glass, G.V. & Stanley, J.C. (1970). *Statistical Methods in Education and Psychology*, Prentice Hall, Englewood Cliffs.
- Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook*, Prentice Hall, Englewood Cliffs.
- Winer, B.J. (1971). *Statistical Principles in Experimental Designs*, McGraw Hill, New York.

(See also **Random Effects and Fixed Effects Fallacy**)

ROBERT J. VANDENBERG

Focus Group Techniques

TILLMAN RODABOUGH AND LACEY WIGGINS

Volume 2, pp. 666–668

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Focus Group Techniques

Introduction

Given the training most scientists have received in quantitative techniques, focus group research is probably the most underrated of all research techniques. The term is inappropriately applied to activities ranging from broad discussions to town hall meetings. Some think of focus groups as the easiest technique to use; it does not require a tight questionnaire – just getting people to talk, and any person with reasonably good social skills can do that. Unlike that stereotype, however, focus groups require professional design, execution, and interpretation. Focus group research is an interview style designed for small groups. It can be seen as directed group brainstorming where a facilitator initiates a discussion and keeps it on topic as the responses of one participant serve as stimuli for other participants. Members' opinions about, or emotional response to, a particular topic, procedure, or product are used in market, political, and social science research.

Techniques

Several good handbooks on focus group techniques are available, for example, Greenbaum [2] and Client Guide to Focus Groups [1]. A brief overview of the focus group technique follows.

Because focus groups 'focus' on a specific concern and a specific category of participants, most focus group research is client driven. When contacted by the client, the researcher must determine whether focus group research is the best technique to use. Focus group research is a qualitative method by which the researcher gains in-depth, sensitive data. However, it is not a reliable technique for collecting quantitative information. It cannot produce frequencies or percentages because the participants are not necessarily representative – the number of participants is too small and they do not have the same proportions of subgroups as the population in question. Therefore, focus groups should not be used to uncover complex relationships that require sophisticated statistical techniques. Also, future projections based on focus groups are inferior to those based on the past experiences of large numbers of people.

If focus group research is the appropriate method, then the researcher must establish what the client wants to know and from whom the information can be gained – the targeted customers or clients. One of the most time consuming, and expensive, aspects of focus group research is locating and gaining the cooperation of members of the target population. Focus groups, because they are small, are rarely representative samples of a given population. The usual technique is to conduct focus groups until the responses of the targeted population are exhausted – continuing until responses are mostly repetitive. Sometimes this can be accomplished with two focus groups if the respondents are not divided by categories such as age, gender, level of income, and previous use of the product, and so on, in an effort to fine-tune responses. Respondents in each focus group should be similar to enhance the comfort level and the honesty of responses and thereby avoid attempts to protect the feelings of dissimilar others. For example, a study conducted by the authors on attitudes toward birth control among teenagers divided respondents by gender (2 categories), age (3 categories), ethnicity (3 categories), and level of sexual activity (3 categories). This gave 54 different categories with at least two focus groups conducted in each, for a minimum of 108 focus group discussions. By using this procedure, 13–14 year-old, nonsexually active, white males were not in the same group with 17–18 year-old, pregnant, black females.

Eight to ten participants in a focus group is about average, depending on the topic and the category of the respondents. In the study mentioned above, six teenagers worked better than the usual 8 to 10 participants because they all wanted to talk and had information to share. Usually, more than eight to ten in a group is undesirable. If the group is too large, the discussants will not have an opportunity to express their attitudes on all facets of the topic.

Besides the discussants, a facilitator is needed who knows how to make the participants feel comfortable enough to disclose the needed information. This person knows how to get reticent members to contribute and how to limit the participation of those who would monopolize the conversation. The facilitator keeps the discussion on topic without appearing to overly structure it. These tasks are difficult to accomplish without alienating some participants unless the facilitator is well trained. The techniques for achieving

2 Focus Group Techniques

good facilitation, verbal and otherwise, are discussed in published research.

The facilitator and discussants are only the most obvious members of focus group research. Someone, usually the primary researcher, must meet with clients to determine the focus of the topic and the sample. Then, someone must locate the sample, frequently using a sampling screen, and contact them with a prepared script that will encourage their participation. Depending on the topic, the client, and the respondents' involvement with the issue, the amount that discussants are paid to participate varies. Participants are usually paid \$25 to \$50 dollars each. Professionals can usually only be persuaded to participate by the offer of a noon meeting, a catered lunch, and the promise that they will be out by 1:15. For professionals, money is rarely an effective incentive to participate. The script inviting them to participate generally states that although the money offered will not completely reimburse them for their time, it will help pay for their gas. Some participants are willing to attend at 5:30 so they can drop by on their way home from work. Older or retired participants prefer meetings in late morning, at noon, or at three in the afternoon. Other participants can only attend at around 7:30 or 9:00 in the evening. Groups are scheduled according to such demographic characteristics as age, location and vocation.

The site for the meeting must be selected for ease of accessibility for the participants. Although our facilities have adequate parking and can be easily accessed from the interstate and from downtown, focus groups for some of our less affluent participants have been held in different parts of the city, sometimes in agency facilities with which they are familiar and comfortable. Focus group facilities in most research centers have a conference room that is wired for audio and video and a one-way mirror for unobtrusive viewing. Participants are informed if they are being taped and if anyone is behind the mirror. They are told that their contributions are so important that the researchers want to carefully record them. In addition, our center also employs two note takers who are in the room, but not at the table, with the discussants. They are introduced when the procedure and reasons for the research are explained, but before the discussion begins. Note takers have the interview questions on a single sheet and record responses on a notepad with answers following the number of each question. They are instructed to type

up their answers the day of the interviews while the discussion is fresh in their minds and before they take notes on another group.

When all focus groups have been conducted, the primary investigator must organize the data so that it can be easily understood, interpret the data in terms of the client's goals, and make recommendations.

Focus groups are fairly expensive if done correctly. The facilitator's fee and the discussants' payments alone can run almost \$1000 per focus group. This amount does not include the cost of locating and contacting the participants, the note takers' payment, the facilities rental, data analysis, and report preparation.

Uses of Focus Groups

Focus group research has become a sophisticated tool for researching a wide variety of topics. Focus groups are used for planning programs, uncovering background information prior to quantitative surveys, testing new program ideas, discovering what customers consider when making decisions, evaluating current programs, understanding an organization's image, assessing a product, and providing feedback to administrators. Frequently linked with other research techniques, it can precede the major data-gathering technique and be used for general exploration and questionnaire design. Researchers can learn what content is necessary to include in questionnaires that are administered to representative samples of the target populations.

The focus group technique is used for a variety of purposes. One of the most widespread uses of focus groups involves general exploration into unfamiliar territory or into the area of new product development. Also, habits and usage studies utilize the focus group technique in obtaining basic information from participants about their use of different products or services and for identifying new opportunities to fill the shifting needs in the market. The dynamic of the focus group allows information to flow easily and allows market researchers to find the deep motivations behind people's actions. Focus groups can lead to new ideas, products, services, themes, explanations, thoughts, images, and metaphors.

Focus groups are commonly used to provide information about or predict the effectiveness of advertising campaigns. For example, participants can be

shown promotional material, or even the sales presentation in a series of focus groups. Focus groups also provide an excellent way for researchers to listen as people deliberate a purchase. The flexibility of focus groups makes them an excellent technique for developing the best positioning for products. They also are used to determine consumer reactions to new packaging, consumer attitudes towards products, services, programs, and for public relations purposes.

The authors have used focus groups to learn how to recruit Campfire Girls leaders, to uncover attitudes toward a river-walk development, to determine the strengths and needed changes for a citywide library system, to discover how to involve university alumni in the alumni association, to determine which magazine supplement to include in the Sunday newspaper; to learn ways to get young single nonsubscribers who read the newspaper to subscribe to it; to determine what signs and slogans worked best for political candidates; to determine which arguments worked best in specific cases for trial lawyers; to decide changes needed in a university continuing education program, to establish the packages and pricing for a new telecommunications company, to

uncover non-homeowner's opinions toward different loan programs that might put them in their own home, to determine which student recruitment techniques work best for a local community college, to discover how the birthing facilities at a local hospital could be made more user friendly, and to uncover attitudes toward different agencies supported by the United Way and ways to encourage better utilization of their services.

Focus groups are flexible, useful, and widely used.

References

- [1] Client Guide to the Focus Group. (2000). Retrieved November 24, 2003, from <http://www.mnav.com/cligd.htm>.
- [2] Greenbaum, T.L. (1993). *The Handbook for Focus Group Research*, Lexington Books, New York.

(See also **Qualitative Research; Survey Questionnaire Design**)

TILLMAN RODABOUGH AND LACEY WIGGINS

Free Response Data Scoring

BRIAN E. CLAUSER AND MELISSA J. MARGOLIS

Volume 2, pp. 669–673

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Free Response Data Scoring

Introduction

The use of computers in assessment affords the opportunity for use of complex, interactive item formats. The availability of multimedia stimuli along with free-text, touch-screen, and voice-activated response formats offers a means of creating assessments that more closely approximate the 'real world' criterion behaviors of interest. However, for this expansion of assessment to be practically useful, it is necessary for the computer both to deliver and score assessments.

This entry presents the state of the art for computerized scoring of complex item formats. It begins by describing the types of procedures currently in use. The entry will then discuss the strengths and weaknesses of differing approaches. The discussion will conclude with comments of the future of automated scoring.

The State of the Art

It is common for technical and logistical issues to limit the extent to which a proficiency of interest can directly be assessed. Real-world tasks are often modified or abstracted to make assessment feasible (e.g., multiple-choice items may assess recognition rather than recall). This type of compromise has also been invoked in order to simplify the task of creating computerized scoring procedures. For example, the complex problem of scoring a task in which the examinee was required to write a computer program could be simplified substantially if the examinee were required to identify and correct errors in an existing program. Two papers by researchers from the Educational Testing Service (ETS) report on efforts to use artificial intelligence procedures to score computer programs. Bennett et al. [2] describe a complex algorithm designed to score constructed-response items in a computer science test. Evaluation of the procedure showed that it was of limited utility. Braun, Bennett, Frye, and Soloway [5] describe a procedure for scoring a simplified alternative task in which the examinee corrected errors in a computer

program rather than producing the program. With this modification, the examinee's changes to the program could be matched to a key and the complexity of the scoring task was therefore greatly constrained.

A similar approach to computerized scoring of a writing task was reported by Davey, Goodwin, and Mettelholtz [8]. They developed a computer-delivered writing task in which the examinee identifies and corrects errors in an essay. The examinee identifies sections of the essay that they believe contain an error. Possible corrections are presented, the examinee selects one, and that choice is then compared to a scoring key. As with the task from the computer science test, constraining the task makes computerized scoring more straightforward; it is a simple matter of matching to a key.

The previous paragraphs describe procedures that allow for scoring by direct matching of an examinee response to a specified key. Another general class of scoring procedures is represented by the approaches to scoring new item formats developed by researchers at ETS. These include mathematical reasoning [4], hypothesis formulation [10], and quantitative representation tasks [3]. With these formats, computerized scoring is accomplished by transforming examinee-constructed responses before matching them to a key. With one of the mathematical reasoning items, the examinee is given a verbal description and asked to produce a mathematical expression that represents the answer (e.g., 'During one week in Trenton in January, it snowed for s days and was fair on the other days. What is the probability that a randomly selected day from that week was fair?' (p. 164)). There are a nearly limitless number of equivalent ways that the mathematical expression could be formulated. For example:

$$1 - \frac{s}{7}$$

or,

$$\frac{7-s}{7}.$$

The format takes advantage of computer software that can reduce any mathematical expression to a base form. This allows for direct matching to a key and, unlike the examples given previously, the match to the key is made possible by the computer and not by restriction of the task posed to the examinee or to the response format.

The scoring procedures described to this point have in common that they allow for directly matching

2 Free Response Data Scoring

the response(s) to a simple key. In some sense, this is true of all scoring procedures. However, as the complexity of the key increases, so does the complexity of the required computer algorithm. Essay scoring has been a kind of Holy Grail of computerized scoring. Researchers have devoted decades to advancing the state of the art for essay scoring. Page's efforts [14, 15] have been joined by those of numerous others [9]. All of these efforts share in common the basic approach that quantifiable aspects of the performance are used to predict expert ratings for a sample of essays. Although the analytic procedures may vary, the regression procedures used by Page provide a conceptual basis for understanding the general approach used by these varying procedures. Early efforts in this arena used relatively simple variables. When the score was to be interpreted as a general measure of writing proficiency, this was a reasonable approach. More recently, the level of sophistication has increased as serious efforts were made to evaluate the content as well as the stylistic features of the essay. One obvious approach to assessing content is to scan the essay for the presence of 'key words'; an essay about the battle of Gettysburg might reasonably be expected to make reference to 'Pickett's charge'. This approach is less likely to be useful when the same concept can be expressed in many different ways. To respond to this problem, Landauer and Dumais [11] developed a procedure in which the relationship between words is represented mathematically. To establish these relationships, large quantities of related text are analyzed. The inferred relationships make it possible to define any essay in terms of a point in n -dimensional space. The similarity between a selected essay and other previously rated essays can then be defined as a function of the distance between the two essays in n -dimensional space.

Essays are not the only context in which complex constructed responses have been successfully scored by computer. Long-term projects by the National Council of Architectural Registration Boards [17] and the National Board of Medical Examiners (NBME) [12] have resulted in computerized scoring procedures for simulations used in certification of architects and licensure of physicians, respectively. In the case of the architectural simulations, the examinee uses a computer interface to complete a design problem. When the design is completed, the computer scores it by applying a branching rule-based algorithm that maps the presence or absence of

different design components into corresponding score categories.

With the computer-based case simulations used in medical licensure assessment, examinees manage patients in a simulated patient-care environment. The examinee uses free-text entry to order diagnostic tests and treatments and results become available after the passage of simulated time. As the examinee advances the case through simulated time, the patient's condition changes based both on the examinee's actions and the underlying problem. Boolean logic is applied to the actions ordered by the examinee to produce scorable items. For example, an examinee may receive credit for an ordered treatment if (a) it occurs after the results of an appropriate diagnostic test were seen, (b) if no other equivalent treatment had already been ordered, and (c) if the treatment were ordered within a specified time frame. After the logical statements are applied to the performance record to convert behaviors into scorable actions, regression-based weights are applied to the items to calculate a score on the case. These weights are derived using expert ratings as the dependent measure in a regression equation.

Empirical Results

Most of the empirical research presented to support the usefulness of the various procedures focuses on the correspondence between scores produced by these systems and those produced by experts. In general, the relationship between automated scores and those of experts is at least as strong as that between the same criterion and those produced by a single expert rater [7, 11, 12, 15]. In the case of the hypothesis generation and mathematical expression item types, the procedures have been assessed in terms of the proportion of examinee responses that could be interpreted successfully by the computer.

Several authors have presented conceptual discussions of the validity issues that arise with the use of computerized scoring procedures [1, 6]. There has, however, been relatively little in the way of sophisticated psychometric evaluation of the scores produced with these procedures. One exception is the evaluative work done on the NBME's computer-based case simulations. A series of papers summarized by Margolis and Clauser [12] compare not only the correspondence between ratings and automated scores

but (a) compare the generalizability of the resulting scores (they are similar), (b) examine the extent to which the results vary as a function of the group of experts used as the basis for modeling the scores (this was at most a minor source of error), (c) examine the extent to which the underlying proficiencies assessed by ratings and scores are identical (correlations were essentially unity), and (d) compare the performance of rule-based and regression-based procedures (the regression-based procedures were superior in this application).

Conceptual Issues

It may be too early in the evolution of automated scoring procedures to establish a useful taxonomy, but some conceptual distinctions between procedures likely will prove helpful. One such distinction relates to whether the criterion on which the scores are based is explicit or implicit. In some circumstances, the scoring rules can be made explicit. When experts can define scorable levels of performance in terms of variables that can directly be quantified by the computer, these rules can be programmed directly. Both the mathematical formulation items and the architectural problems belong to this category. These approaches have the advantage that the rules can be explicitly examined and openly critiqued. Such examination facilitates refinement of the rules; it also has the potential to strengthen the argument supporting the validity of the resulting score interpretations.

By contrast, in some circumstances it is difficult to define performance levels in terms of quantifiable variables. As a result, many of the currently used procedures rely on implicit, or inferred, criteria. Examples of these include essentially all currently available approaches for scoring essays. These procedures require expert review and rating of a sample of examinee performances. Scores are then modeled on the basis of the implicit relationship between the observed set of ratings and the quantifiable variables from the performances. The most common procedure for deriving this implicit relationship is **multiple linear regression**; Page's early work on computerized scoring of essays and the scoring of computer-based case simulations both took this approach. One important characteristic of the implicit nature of this relationship is that the quantified characteristics may

not actually represent the characteristics that experts consider when rating a performance; they may instead act as proxies for those characteristics.

The use of proxies has both advantages and disadvantages. One advantage is that it may be difficult to identify or quantify the actual characteristics of interest. Consider the problem of defining and quantifying the characteristics that make one essay better than another. However, the use of proxy variables may have associated risks. If examinees know that the essay is being judged on the basis of the number of words, and so on, they may be able to manipulate the system to increase their scores without improving the quality of their essays. The use of implicit criteria opens the possibility of using proxy measures as the basis for scoring, but it does not require the use of such measures.

Another significant issue in the use of computer-delivered assessments that require complex automated scoring procedures is the potential for the scores to be influenced by construct-irrelevant variance. To the extent that computer delivery and/or scoring of assessments results in modifying the assessment task so that it fails to correspond to the criterion 'real world' behavior, the modifications may result in construct-irrelevant variance. Limitations of the scoring system may also induce construct-irrelevant variance. Consider the writing task described by Davey, Goodwin, and Mettelholtz [8]. To the extent that competent writers may not be careful and critical readers, the potential exists for scores that are interpreted as representing writing skills to be influenced by an examinee's editorial skills. Similarly, in the mathematical expressions tasks, Bennett and colleagues [4] describe a problem with scoring resulting from the fact that, if examinees include labels in their expression (e.g., 'days'), the scoring algorithm may not correctly interpret expressions that would be scored correctly by expert review.

A final important issue with computerized scoring procedures is that the computer scores with mechanical consistency. Even the most highly trained human raters will fall short of this standard. This level of consistency is certainly a strength for these procedures. However, to the extent that the automated algorithm introduces error into the scores, this error will also be propagated with mechanical consistency. This has important implications because it has the potential to replace random errors (which will tend

to average out across tasks or raters) with systematic errors.

The Future of Automated Scoring

It does not require uncommon prescience to predict that the use of automated scoring procedures for complex computer-delivered assessments will increase both in use and complexity. The improvements in recognition of handwriting and vocal speech will broaden the range of the assessment format. The availability of low-cost computers and the construction of secure computerized test-delivery networks has opened the potential for large and small-scale computerized testing administrations in high and low-stakes contexts.

The increasing use of computers in assessment and the concomitant increasing use of automated scoring procedures seems all but inevitable. This increase will be facilitated to the extent that two branches of research and development are successful. First, there is the need to make routine what is now state of the art. The level of expertise required to develop the more sophisticated of the procedures described in this article puts their use beyond the resources available to most test developers. Secondly, new procedures are needed that will support development of task-specific keys for automated scoring. Issues of technical expertise aside, the human resources currently required to develop the scoring algorithms for individual tasks is well in excess of that required for testing on the basis of multiple-choice items. To the extent that computers can replace humans in this activity, the applications will become much more widely applicable.

Finally, at present, there are a limited number of specific formats that are scorable by computer; this entry has referenced many of them. New and increasingly innovative formats and scoring procedures are sure to be developed within the coming years. Technologies such as artificial **neural networks** are promising [16]. Similarly, advances in cognitive science may provide a framework for developing new approaches [13].

References

- [1] Bejar, I.I. & Bennett, R.E. (1997). Validity and automated scoring: It's not only the scoring, *Educational Measurement: Issues and Practice* 17(4), 9–16.
- [2] Bennett, R.E., Gong, B., Hershaw, R.C., Rock, D.A., Soloway, E. & Macalalad, A. (1990). Assessment of an expert systems ability to automatically grade and diagnose student's constructed responses to computer science problems, in *Artificial Intelligence and the Future of Testing*, R.O. Freedle, ed., Lawrence Erlbaum Associates, Hillsdale, 293–320.
- [3] Bennett, R.E. & Sebrechts, M.M. (1997). A computerized task for representing the representational component of quantitative proficiency, *Journal of Educational Measurement* 34, 64–77.
- [4] Bennett, R.E., Steffen, M., Singley, M.K., Morley, M. & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive testing, *Journal of Educational Measurement* 34, 162–176.
- [5] Braun, H.I., Bennett, R.E., Frye, D. & Soloway, E. (1990). Scoring constructed responses using expert systems, *Journal of Educational Measurement* 27, 93–108.
- [6] Clauser, B.E., Kane, M.T. & Swanson, D.B. (2002). Validity issues for performance based tests scored with computer-automated scoring systems, *Applied Measurement in Education* 15, 413–432.
- [7] Clauser, B.E., Margolis, M.J., Clyman, S.G. & Ross, L.P. (1997). Development of automated scoring algorithms for complex performance assessments: a comparison of two approaches, *Journal of Educational Measurement* 34, 141–161.
- [8] Davey, T., Goodwin, J. & Mettelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment, *Journal of Educational Measurement* 34, 21–41.
- [9] Deane, P. (in press). Strategies for evidence identification through linguistic assessment of textual responses, in *Automated Scoring of Complex Tasks in Computer Based Testing*, D. Williamson, I. Bejar & R. Mislevy, eds, Lawrence Erlbaum Associates, Hillsdale.
- [10] Kaplan, R.M. & Bennett, R.E. (1994). *Using a Free-response Scoring Tool to Automatically Score the Formulating-hypotheses Item (RR 94-08)*, Educational Testing Service, Princeton.
- [11] Landauer, T.K., Haham, D., Foltz, P.W. (2003). Automated scoring and annotation of essays within the Intelligent Essay Assessor, in *Automated essay scoring: A cross Disciplinary Perspective*, M.D. Shermis & J. Burstein, eds, Lawrence Erlbaum Associates, London, 87–112.
- [12] Margolis, M.J. & Clauser, B.E. (in press). A regression-based procedure for automated scoring of a complex medical performance assessment, in *Automated Scoring of Complex Tasks in Computer Based Testing*, D. Williamson, I. Bejar & R. Mislevy, eds, Lawrence Erlbaum Associates, Hillsdale.
- [13] Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G. & Johnson, L. (2002). Making sense of data from complex assessments, *Applied Measurement in Education* 15, 363–390.

- [14] Page, E.B. (1966). Grading essays by computer: progress report, *Proceedings of the 1966 Invitational Conference on Testing*, Educational Testing Service, Princeton, 87–100.
- [15] Page, E.B. & Petersen, N.S. (1995). The computer moves into essay grading, *Phi Delta Kappan* **76**, 561–565.
- [16] Stevens, R.H. & Casillas, A. (in press). Artificial neural networks, in *Automated Scoring of Complex Tasks in Computer Based Testing*, D. Williamson, I. Bejar & R. Mislevy, eds, Lawrence Erlbaum Associates, Hillsdale.
- [17] Williamson, D.M., Bejar, I.I. & Hone, A.S. (1999). “Mental model” comparison of automated and human scoring, *Journal of Educational Measurement* **36**, 158–184.

BRIAN E. CLAUSER AND MELISSA
J. MARGOLIS

Friedman's Test

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Volume 2, pp. 673–674

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Friedman's Test

The nonparametric Friedman [4] test expands the **sign (S) test** to k related samples. The null hypothesis is that the samples come from the same population, which is tested against the alternative that at least one of the samples comes from a different population. The data are arranged in k columns and n rows, where each row contains k related observations. It is frequently positioned as an alternative to the parametric repeated measures one-way analysis of variance. Note that the Agresti-Pendergast procedure [1] and the Neave and Worthington [6] Match Test are often more powerful competitors to the Friedman test.

Procedure

Rank the observations for each row from 1 to k . For each of the k columns, the ranks are added and averaged, and the mean is designated $\bar{R}_j$. The mean of the ranks is $\bar{R} = 1/2(k + 1)$. The sum of the squares of the deviations of mean of the ranks of the columns from the mean rank is computed. The test statistic is a multiple of this sum.

Assumptions

It is assumed that the rows are independent and there are no tied observations in a row. Because comparisons are made within rows, tied values may not pose a serious threat. Typically, average ranks are assigned to ties.

Test Statistic

The test statistic, M , is a multiple of S :

$$S = \sum_{j=1}^k (\bar{R}_j - \bar{R})^2$$

$$M = \frac{12n}{k(k+1)} S, \quad (1)$$

where n is the number of rows, and k is the number of columns. An alternate formula is:

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1), \quad (2)$$

where n is the number of rows, k is the number of columns, and R_j is the rank sum for the j th column, $j = 1, 2, 3, \dots, k$.

Large Sample Sizes

For large sample sizes, the critical values can be approximated by the Chi-square distribution with $k - 1$ degrees of freedom. Monte Carlo simulations conducted by Fahoome and Sawilowsky [3] and Fahoome [2] indicated that the large sample approximation requires a minimum sample size of 13 for $\alpha = 0.05$, and 23 for $\alpha = 0.01$. Hodges, Ramsey, and Shaffer [5] provided a competitive alternative in computing critical values.

Example

Friedman's test is calculated with Samples 1 to 5 in the table below in Table 1, $n_1 = n_2 = n_3 = n_4 = n_5 = 15$.

The rows are ranked, with average ranks assigned to tied ranks as in Table 2.

The column sums are: $R_1 = 48.5$, $R_2 = 47.0$, $R_3 = 33.0$, $R_4 = 52.5$, and $R_5 = 44.0$. The sum of the squared rank sums is 10 342.5. $M = (12/15 \times 5 \times 6)(10\,342.5) - 3 \times 15 \times 6 = 0.02667(10\,342.5) - 270 = 5.8$. The large sample approximation of the critical value is 9.488, chi-square with $5 - 1 = 4$ degrees of freedom, and $\alpha = 0.05$. Because $5.8 <$

Table 1 Sample data

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
20	11	9	34	10
33	34	14	10	2
4	23	33	38	32
34	37	5	41	4
13	11	8	4	33
6	24	14	26	19
29	5	20	10	11
17	9	18	21	21
39	11	8	13	9
26	33	22	15	31
13	32	11	35	12
9	18	33	43	20
33	27	20	13	33
16	21	7	20	15
36	8	7	13	15

2 Friedman's Test

Table 2 Computations for the Friedman test

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	
	4	3	1	5	2	
	4	5	3	2	1	
	1	2	4	5	3	
	3	4	2	5	1	
	4	3	2	1	5	
	1	4	2	5	3	
	5	1	4	2	3	
	2	1	3	4.5	4.5	
	5	3	1	4	2	
	3	5	2	1	4	
	3	4	1	5	2	
	1	2	4	5	3	
	4.5	3	2	1	4.5	
	3	5	1	4	2	
	5	2	1	3	4	
Total	48.5	47.0	33.0	52.5	44.0	
R ²	2352.25	2209.00	1089.00	2756.25	1936.00	10 342.50

9.488, the null hypothesis cannot be rejected on the basis of the evidence from these samples.

References

- [1] Agresti, A. & Pendergast, J. (1986). Comparing mean ranks for repeated measures data, *Communications in Statistics, Theory and Methods* **15**, 1417–1433.
- [2] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.
- [3] Fahoome, G. & Sawilowsky, S. (2000). Twenty nonparametric statistics, *Annual Meeting of the American Educational Research Association*, SIG/Educational Statisticians, New Orleans.
- [4] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.
- [5] Hodges, J.L., Ramsey, P.H. & Shaffer, J.P. (1993). Accurate probabilities for the sign test, *Communications in Statistics, Theory and Methods* **22**, 1235–1255.
- [6] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Functional Data Analysis

JAMES RAMSAY

Volume 2, pp. 675–678

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Functional Data Analysis

What is Functional Data Analysis?

Functional data analysis, or FDA, is the modeling of data using functional parameters. By a functional parameter, we mean a function whose shape and complexity is not known in advance of the analysis, and therefore the modeling process must provide as much flexibility in the estimated function as the data require. By contrast, more classical parametric approaches to function estimation assume a fixed form for the function defined by a small number of parameters, and focus on estimating these parameters as the goal of the modeling process. As a consequence, while FDA certainly estimates parameters, the attention is on the entire function rather than on the values of these parameters.

Some of the oldest problems in psychology and education are functional in nature. Psychophysics aims to estimate a curve relating a physical measurement to a subjective or perceived counterpart (*see Psychophysical Scaling*), and learning theory in its earlier periods attempted to model the relationship between either gain or loss of performance over time.

Many functional problems involve data where we can see the function in the data, but where we need to smooth the data in order to study the relation more closely (*see Kernel Smoothing; Scatterplot Smoothers*). Figure 1 shows the relation between measurements of the heights of 10 girls and their ages. The data are taken from the Berkeley Growth Study [5]. But, it is really Figure 2 that we need: The acceleration in height, or its second derivative, where we can see the pubertal growth spurt more clearly and how it varies, and a number of additional features as well (*see Growth Curve Modeling*).

However, other types of data require functional models even though the data themselves would not usually be seen as functional. A familiar example comes from test theory, where we represent for each test item the relation between the probability of an examinee getting an item correct to a corresponding position on a latent ability continuum (*see Item Response Theory (IRT) Models for Polytomous Response Data*). An example of three item response functions from an analysis reported in [4] is shown in Figure 3.

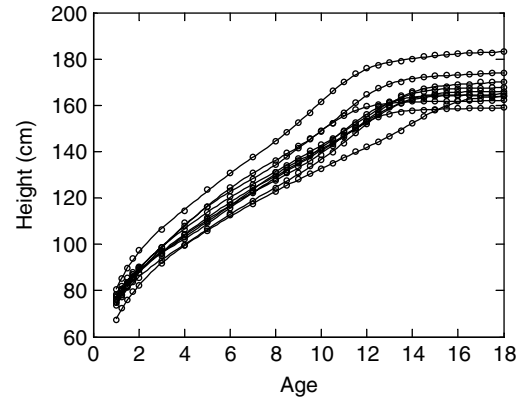


Figure 1 The heights of 10 girls. The height measurements are indicated by the circles. The smooth lines are height functions estimated using monotone smoothing methods (Data taken from [5])

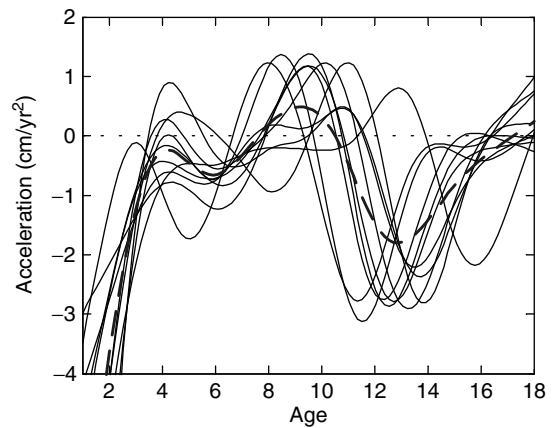


Figure 2 The height acceleration curves, that is, the second derivatives, of the curves shown in Figure 1. The pubertal growth spurt is the large positive peak followed by the large valley, and the midpoint of the spurt is the age at which the curve crosses zero. The heavy dashed line is the cross-sectional mean of these curves, and indicates the need for registration or alignment of the pubertal growth spurt (Reproduced from [2] with permission from Springer)

In what follows, we will, for simplicity, refer to the argument t of a function $x(t)$ as ‘time’, but, in fact, the argument t can be space, frequency, depression level, or virtually any continuum. Higher dimensional arguments are also possible; an image, for example, is a functional observation with values defined over a planar region.

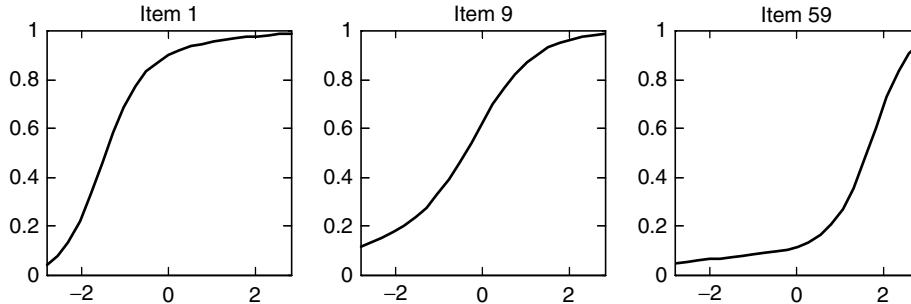


Figure 3 Three item response functions showing the relation between the probability of getting an item correct on a mathematics test as a function of ability (Reproduced from [4]. Copyright 2002 by the American Educational Research Association and the American Statistical Association; reproduced with permission from the publisher)

How are Functions Constructed in Functional Data Analysis?

Functions of arbitrary shape and complexity are constructed using a set of K functional building blocks called *basis functions*. These basis functions are combined linearly by multiplying each basis function $\phi_k(t)$, $k = 1, \dots, K$ by a coefficient c_k and summing results. That is,

$$x(t) = \sum_{k=1}^K c_k \phi_k(t). \quad (1)$$

A familiar example is the polynomial, constructed by taking linear combinations of powers of t . When $x(t)$ is constructed with a *Fourier series*, the basis functions are one of a series of sine and cosine pairs, each pair being a function of an integer multiple of a base frequency. This is appropriate when the data are periodic.

Where an unconstrained function is to be estimated, the preferred basis functions tend to be the splines, constructed by joining polynomial segments together at a series of values of t called knots. Splines have pretty much replaced polynomials for functional work because of their much greater flexibility and computational convenience (*see Scatterplot Smoothers*).

No matter what the basis system, the flexibility of the resulting curve is determined by the number K of basis functions, and a typical analysis involves determining how large K must be in order to capture the required features of the function being estimated.

Smoothness, Derivatives, and Functional Data Analysis

FDA assumes that the function being estimated is smooth. In practice, this means that the function has one or more derivatives that are themselves smooth or at least continuous. Derivatives play many roles in the technology of FDA. The growth curve analysis had the study of the second derivative as its immediate goal.

Derivatives are also used to quantify smoothness. A frequently used method is to define the total curvature of a function by the integral of the square of its second or higher-order derivative. This measure is called a *roughness penalty*, and a functional parameter is estimated by explicitly controlling its roughness.

In many situations, we need to study rates of change directly, that is, the *dynamics* of a process distributed over time, space, or some other continuum. In these situations, it can be natural to develop differential equations, which are functional relationships between a function and one or more of its derivatives. For example, sinusoidal oscillation in a function $x(t)$ can be expressed by the equation $D^2x(t) = -\omega^2x(t)$, where the notation $D^2x(t)$ refers to the second derivative of x , and $2\pi/\omega^2$ is the period. The use of FDA techniques to estimate differential equations has many applications in fields such as chemical engineering and control theory, but should also prove important in the emerging study of the dynamic aspects of human behavior.

We often need to fit functions to data that have special constraints. A familiar example is the probability

density function $p(t)$ that we estimate to summarize the distribution of a sample of N values t_i (see **Catalogue of Probability Density Functions**). A density function must be positive and must integrate to one. It is reasonable to assume that growth curves such as those shown in Figure 1 are strictly increasing. Item response functions must take values within the unit interval $[0, 1]$. Constrained functions like these can often be elegantly expressed in terms of differential equations. For example, any strictly increasing curve $x(t)$ can be expressed in terms of the equation $D^2x(t) = w(x)Dx(t)$, where the alternative functional parameter $w(t)$ has no constraints whatever. Estimating $w(t)$ rather than $x(t)$ is both easier and assures monotonicity.

Phase Variation and Registration in FDA

A FDA can confront new problems not encountered in multivariate and other older types of statistical procedures. One of these is the presence of phase variation, illustrated in Figure 2. We see there that the pubertal growth spurt varies in both amplitude and timing from girl to girl. This is because each child has a physiological age that does not evolve at the same rate as chronological age. We call this variation in timing of curve features *phase variation*.

The problem with phase variation is illustrated in the heavy dashed mean curve in Figure 2. Because girls are at different stages of their maturation at any particular clock age, the cross-sectional mean is a terrible estimate of the average child's growth pattern. The mean acceleration displays a pubertal growth spurt that lasts longer than that for any single girl, and also has less *amplitude variation* as well.

Before we can conduct even the simplest analyses, such as computing means and standard deviations, we must remove phase variation. This is done by computing a nonlinear, but strictly increasing, transformation of clock time called a *time warping function*, such that when a child's curve values are plotted against transformed time, features such as the pubertal growth spurt are aligned. This procedure is often called *curve registration*, and can be an essential first step in an FDA.

What are Some Functional Data Analyses?

Nearly all the analyses that are used in multivariate statistics have their functional counterparts. For example, estimating functional descriptive statistics such as mean curve, a standard deviation curve, and a bivariate *correlation function* are usual first steps in an FDA, after, of course, registering the curves, if required.

Then many investigators will turn to a functional version of **principal components analysis** (PCA) to study the dominant modes of variation among a sample of curves. Here, the principal component vectors in multivariate PCA become principal functional components of variation. As in ordinary PCA, a central issue is determining how many of these components are required to adequately account for the functional variation in the data, and rotating principal components can be helpful here, too. A functional analogue of **canonical correlation** analysis may also be useful.

Multiple regression analysis or the **linear model** has a wide range of functional counterparts. A functional **analysis of variance** involves dependent variables that are curves. We could, for example, compute a functional version of the t Test to see if the acceleration curves in Figure 2 differ between boys and girls. In such tests, it can be useful to identify regions on the t -axis where there are significant differences, rather than being content just to show that differences exist. This is the functional analogue of the multiple comparison problem (see **Multiple Comparison Procedures**).

What happens when an independent variable in a regression analysis is itself a function? Such situations often arise in medicine and engineering, where a patient or some industrial process produces a measurable response over time to a time-varying input of some sort, such as drug dose or raw material, respectively. In some situations, the effect of varying the input has an immediate effect on the output, and, in other situations, we need to compute causal effects over earlier times as well. Functional independent variables introduce a fascinating number of new technical challenges, as it is absolutely essential to impose smoothness on the estimated functional regression coefficients.

Differential equations are, in effect, functional linear models where the output is a derivative, and among the inputs are the function's value and

perhaps also those of a certain number of lower-order derivatives. Differential equations, as well as other functional models, can be linear or nonlinear. An important part of FDA is the emergence of new methods for estimating differential equations from raw noisy discrete data.

Resources for Functional Data Analysis

The field is described in detail in [3], a revised and expanded edition of [1]. A series of case studies involving a variety of types of FDA are described in [2].

Along with these monographs are a set of functions in the high-level computer languages R, S-PLUS, and Matlab that are downloadable from the web site www.functionaldata.org. This software comes with a number of sets of sample data, including the growth data used in this article, along with various analyses using these functions.

References

- [1] Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*, Springer, New York.
- [2] Ramsay, J.O. & Silverman, B.W. (2002). *Applied Functional Data Analysis*, Springer, New York.
- [3] Ramsay, J.O. & Silverman, B.W. (2005). *Functional Data Analysis*, 2nd Edition, Springer, New York.
- [4] Rossi, N., Wang, X. & Ramsay, J.O. (2002). Nonparametric item response function estimates with the EM algorithm, *Journal of the Behavioral and Educational Sciences* **27**, 291–317.
- [5] Tuddenham, R.D. & Snyder, M.M. (1954). Physical growth of California boys and girls from birth to eighteen years, *University of California Publications in Child Development* **1**, 183–364.

(See also **Structural Equation Modeling: Latent Growth Curve Analysis**)

JAMES RAMSAY

Fuzzy Cluster Analysis

KENNETH G. MANTON, GENE LOWRIMORE, ANATOLI YASHIN AND MIKHAIL KOVTUN

Volume 2, pp. 678–686

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Fuzzy Cluster Analysis

Cluster analysis is an exploratory method used to classify, or to ‘cluster’, objects under consideration (see **Cluster Analysis: Overview**). However, the crisp membership of objects in clusters, derived by classic cluster analysis, is not always satisfactory. Consider a simple example.

Let the weights of fish in a pond be measured. Assume that there are two generations of fish, one $\frac{1}{2}$ years old and another $1\frac{1}{2}$ years old. The **histogram** of the distribution of weights may look like Figure 1.

This picture clearly suggests that there are two clusters, one with center at y_1 and another with center at y_2 . The crisp cluster analysis will assign a fish with a weight of less than y_0 to the first cluster and a fish with a weight greater than y_0 to the second cluster. Intuitively, it is a good assignment for weights x_1 and

x_2 (assigned to the first cluster) and for weights x_5 , x_6 , and x_7 (assigned to the second cluster). But the correct crisp assignment of weights x_3 and x_4 is not so obvious, let alone the point y_0 , which is exactly in the middle of the clusters’ center.

Fuzzy cluster analysis suggests considering a *partial membership* of objects in clusters. Partial membership of object i in cluster k is a real number u_{ik} between 0 (no membership) and 1 (full membership). The crisp cluster analysis may be considered as a particular case of fuzzy cluster analysis, where the values allowed for u_{ik} are only 0 and 1 (no intermediate value is allowed). In our example, assignment to clusters depends only on weight w , and thus membership in cluster k is a function $u_k(w)$. Figure 2 shows membership functions for crisp clustering, and Figure 3 shows a desired shape of membership function for fuzzy clustering.

There are many types of fuzzy cluster analyses. One of the most important is *probabilistic fuzzy*

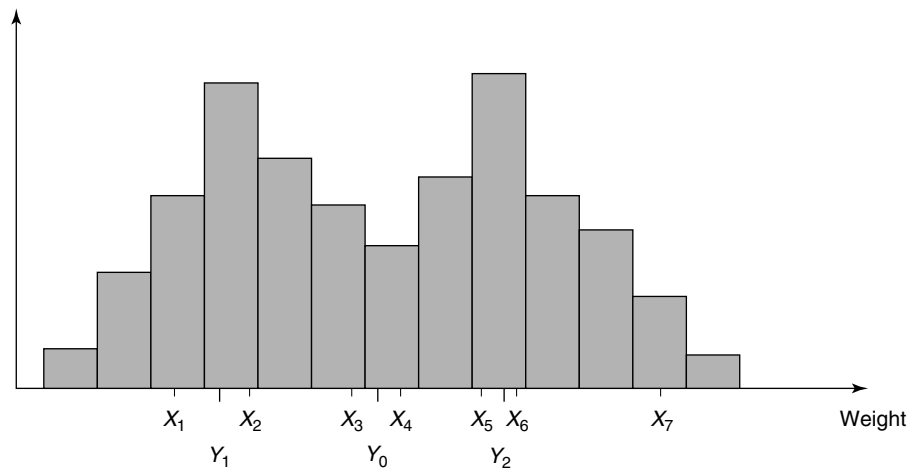


Figure 1 A Histogram of distribution of weight of fish in a pond

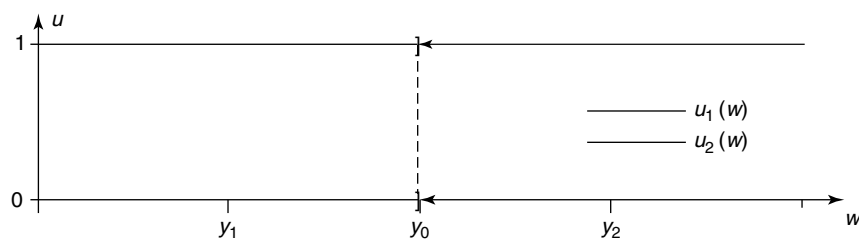


Figure 2 Membership functions for crisp clustering

2 Fuzzy Cluster Analysis

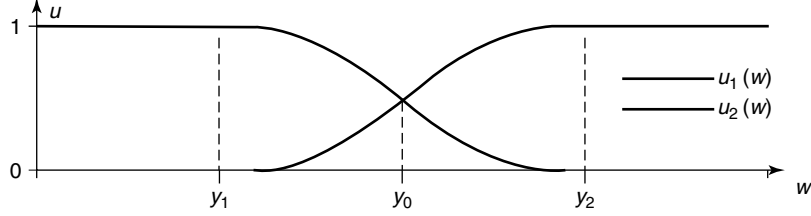


Figure 3 Membership functions for fuzzy clustering

clustering, which requires that for every object i its memberships in all clusters sum to 1, $\sum_k u_{ik} = 1$. In this case, values u_{ik} may be treated as probabilities for object i to belong to cluster k .

As in the case of crisp cluster analysis, the inputs for the fuzzy cluster analysis are the results of m measurements made on n objects, which can be represented:

$$\begin{pmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^m & x_2^m & \dots & x_n^m \end{pmatrix} \quad (1)$$

The situation differs significantly, depending on whether measurements are continuous or categorical. First, we consider continuous measurements, for which fuzzy K -means algorithm will be discussed in detail and other methods will be briefly characterized. Then we will discuss methods for analyzing categorical data which may be used for fuzzy clustering.

Fuzzy Clustering of Continuous Data: Fuzzy K-means Algorithm

When the result of every measurement is a real number, the columns of matrix (1) (which represent objects) may be considered as points in m -dimensional space. In crisp K -means clustering, the goal is to split objects into K clusters $c_1, \dots, c_K$ with centers $\mathbf{v}_1, \dots, \mathbf{v}_K$ such that

$$\sum_{k=1}^K \sum_{x_i \in c_k} d(\mathbf{x}_i, \mathbf{v}_k) \quad (2)$$

achieves its minimum (see **k-means Analysis**). This may be reformulated as the minimization of

$$\sum_{k=1}^K \sum_{i=1}^n u_{ik} d(\mathbf{x}_i, \mathbf{v}_k), \quad (3)$$

with constraints

$$\sum_{k=1}^K u_{ik} = 1 \quad (4)$$

for every $i = 1, \dots, n$. The allowed values for u_{ik} are 0 and 1; therefore, (4) means that for every i , only one value among $u_{i1}, \dots, u_{iK}$ is 1 and all others are 0. The distance $d(\mathbf{x}, \mathbf{y})$ may be chosen from a wide range of formulas, but for computational efficiency it is necessary to have a simple way to compute centers of clusters. The usual choice is the squared Euclidean distance, $d(\mathbf{x}, \mathbf{y}) = \sum_j (x^j - y^j)^2$, where the center of a cluster is its center of gravity. For the sake of simplicity, we restrict our consideration to the squared Euclidean distance.

Equation (3) suggests that fuzzy clustering may be obtained by relaxing the restriction ' u_{ik} is either 0 or 1'; rather, u_{ik} is allowed to take any value in the interval $[0,1]$ and is treated as the degree of membership of object i in cluster k . However, this is not as simple as it appears. One can show that the minimum of (3) with constraints (4) is still obtained when u_{ik} are 0s or 1s, despite the admissibility of intermediate values. In this problem, an additional parameter $f \geq 1$, called a *fuzzifier*, can be introduced in (3):

$$\sum_{k=1}^K \sum_{i=1}^n (u_{ik})^f d(\mathbf{x}_i, \mathbf{v}_k) \quad (5)$$

The fuzzifier has no effect in crisp K -means clustering (as $0^f = 0$ and $1^f = 1$), but it produces nontrivial minima of (5) with constraints (4).

Now the fuzzy clustering problem is a problem of finding the minimum of (5) under constraints (4). The fuzzy K -means algorithm searches for this minimum by alternating two steps: (a) optimizing membership degrees u_{ik} while cluster centers $\mathbf{v}_k$ are fixed; and (b) optimizing $\mathbf{v}_k$ while u_{ik} are fixed. The minimum of (5), with respect to u_{ik} , is

$$u_{ik} = \frac{1}{\sum_{k'=1}^K \left(\frac{d(\mathbf{x}_i, \mathbf{v}_k)}{d(\mathbf{x}_i, \mathbf{v}_{k'})} \right)^{\frac{1}{f-1}}}, \quad (6)$$

and the minimum of (5), with respect to $\mathbf{v}_k$, is

$$\mathbf{v}_k = \frac{\sum_{i=1}^n (u_{ik})^f \mathbf{x}_i}{\sum_{i=1}^n (u_{ik})^f}. \quad (7)$$

Equation (7) is a vector equation; it defines a center of gravity of masses $(u_{1k})^f, \dots, (u_{nk})^f$ placed at

points $\mathbf{x}_1, \dots, \mathbf{x}_n$. The right side of formula (6) is undefined if $d(\mathbf{x}_i, \mathbf{v}_{k_0})$ is 0 for some k_0 ; in this case, one lets $u_{ik_0} = 1$ and $u_{ik} = 0$ for all other k . The algorithm stops when changes in u_{ik} and $\mathbf{v}_k$ during the last step are below a predefined threshold.

The fuzziness of the cluster depends on fuzzifier f . If f is close to 1, the membership is close to a crisp one; if f tends to infinity, the fuzzy K -means algorithm tends to give equal membership in all clusters to all objects. Figures 4, 5, and 6 demonstrate membership functions for $f = 2, 3, 5$. The most common choice for the fuzzifier is $f = 2$.

This method gives nonzero membership in all clusters for any object that does not coincide with the center of one of the clusters. Some researchers, however, prefer to have a crisp membership for objects close to cluster centers and fuzzy membership for objects that are close to cluster boundaries. One possibility was suggested by Klawonn and Hoepfner [4]. Their central idea is to consider subexpression u^f as a special case of function $g(u)$. To be used in place of the fuzzifier, such functions must (a) be a

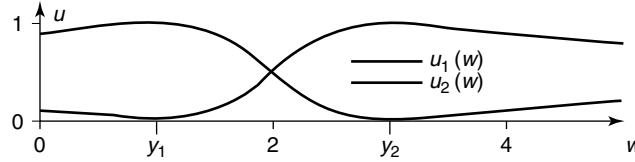


Figure 4 Membership functions for $f = 2$

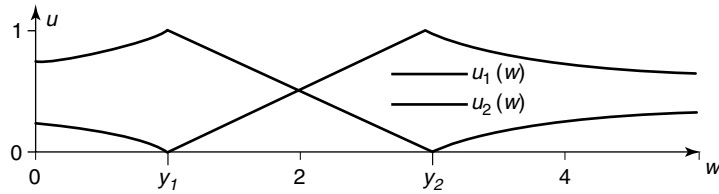


Figure 5 Membership functions for $f = 3$

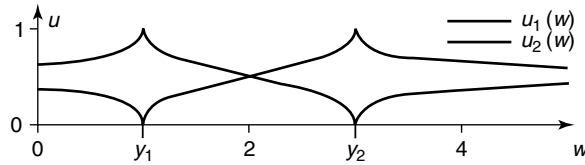


Figure 6 Membership functions for $f = 4$

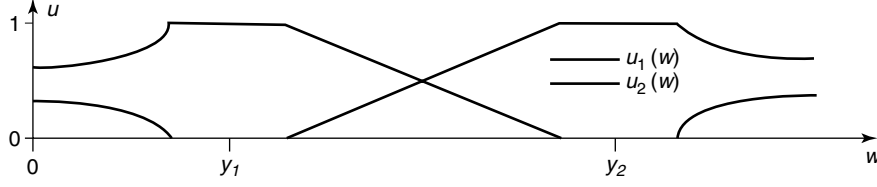


Figure 7 Membership functions for fuzzifier $g(u) = \frac{1}{2}u^2 + \frac{1}{2}u$

monotonically increasing map of the interval $[0,1]$ into itself with $g(0) = 0$, $g(1) = 1$; (b) must have a continuous increasing derivative; and (c) must satisfy $g'(0) < g'(1)$. The standard fuzzifier $g(u) = u^f$ possesses these properties, but this approach gives other choices. Klawonn and Hoepfner proved that to obtain crisp membership for objects close to cluster centers, $g(u)$ must satisfy $g'(0) > 0$. One possibility for such functions is $g(u) = \alpha u^2 + (1 - \alpha)u$ with $0 < \alpha < 1$. The membership functions for the fish example with fuzzifier $g(u) = 1/2u^2 + 1/2u$ are shown in Figure 7.

Although theoretically $g(u)$ may be chosen from a wide range of functions, to obtain a good equivalent of (6), it must have a sufficiently simple analytical form.

Fuzzy Clustering of Continuous Data: Other Approaches

Probabilistic fuzzy clustering gives approximately equal membership degrees in all clusters for objects that are far from all clusters. Moreover, ‘bad’ objects move cluster centers from the position defined by ‘good’ members. Several methods have been suggested to resolve this problem.

Noise clustering introduces an additional, aptly designated the ‘noise cluster’. The distance from any object to the noise cluster is the same large number. In this way, objects near the border between two normal clusters still receive nonzero membership in these clusters, while objects that are far away from all normal clusters become members of the noise cluster, and have no membership in normal clusters.

Possibilistic clustering tries to handle the problem by dropping the constraints (4). To avoid a trivial minimum of (5) (all degrees of membership are 0s),

(5) is modified to

$$\sum_{k=1}^K \sum_{i=1}^n (u_{ik})^f d(\mathbf{x}_i, \mathbf{v}_k) + \sum_{k=1}^K \eta_k \sum_{i=1}^n (1 - u_{ik})^f \quad (8)$$

The global minimum of (8) does not give a satisfactory solution, but local minima near a result of a fuzzy K -means clustering algorithm solution produce a good clustering. Minimization of (8) works exactly as minimization of (5) in the fuzzy K -means algorithm – with one exception: formula (6) for updating membership degrees is replaced by

$$u_{ik} = \frac{1}{1 + \left(\frac{d(\mathbf{x}_i, \mathbf{v}_k)}{\eta_k} \right)^{\frac{1}{f-1}}} \quad (9)$$

This formula also explains the meaning of coefficients η_k : it is the distance from the center of cluster $\mathbf{v}_k$, at which the membership degree equals 0.5. This suggests that the way to calculate η_k from an existing clustering is:

$$\eta_k = \alpha \frac{\sum_{i=1}^n (u_{ik})^f d(\mathbf{x}_i, \mathbf{v}_k)}{\sum_{i=1}^n (u_{ik})^f} \quad (10)$$

Usually, a possibilistic clustering is performed in three steps. First, a fuzzy K -means clustering is performed. Second, coefficients η_k are calculated using (10). Third, (8) is minimized by alternately applying formulas (9) and (7).

A different approach to fuzzy clustering arises from the theory of mixed distributions [2, 8, 10]. When objects under classification may be considered realizations of random variables (or, more generally, of random vectors), and the observed distribution law can be represented as a finite mixture of simpler distribution laws, component distribution laws may be

considered as clusters; consequently membership in a cluster is the probability of belonging to a component distribution law (conditional on observations) (see **Finite Mixture Distributions; Model Based Cluster Analysis**). In the fish example, the observed distribution law can be represented as a mixture of two normal distributions, which leads to two clusters similar to those previously considered.

The applicability of this approach is restricted in that a representation as a finite mixture may or may not exist, or may be not unique, even when an obvious decomposition into clusters is present in the data. On the other hand, although there is no obvious extension of the fuzzy K -means algorithm to categorical data, the mixed distribution approach can be applied to categorical data.

Fuzzy Clustering of Categorical Data: Latent Class Models

Latent structure analysis [3, 6, 7] deals with categorical measurements. The columns of (1) are vectors of measurements made on an object. These vectors may be considered as realizations of a random vector $\mathbf{x} = (X^1, \dots, X^m)$. We say that the distribution law of random vector $\mathbf{x}$ is independent, if component random variables $X^1, \dots, X^m$ are mutually independent.

The observed distribution law is not required to be independent; however, under some circumstances, it may be represented as a mixture of independent distribution laws. This allows considering a population as a disjointed union of classes ('latent' classes), such that the distribution law of random vector $\mathbf{x}$ in every class is independent. Probabilities for objects belonging to a class that is conditional on the outcomes of measurements can be calculated and can be considered as degrees of membership in corresponding classes (see **Latent Class Analysis**). Most widely used algorithms for construction of latent class models are based on maximizing the likelihood function and involve heavy computation.

Fuzzy Clustering of Categorical Data: Grade of Membership Analysis

Grade of membership (GoM) analysis [5, 9, 11] works with the same data as latent structure analysis. It also searches for a representation of the observed distribution as a mixture of independent distributions.

However, in GoM, though the mixture sought is allowed to be infinite, all mixed distributions must belong to a low-dimensional linear subspace Q of a space of independent distributions. Under weak conditions this linear subspace is identifiable, and the algorithm for finding this subspace reduces the problem to an **eigenvalue/eigenvector** problem.

The mixing distribution may be considered a distribution of a random vector $\mathbf{g}$ taking values in Q . Individual scores $\mathbf{g}_i$ are expectations of random vector $\mathbf{g}$ conditional on outcome of measurements $x_i^1, \dots, x_i^m$. Conditional expectations may be found as a solution of a linear system of equations. Let subspace Q be K -dimensional, $\lambda^1, \dots, \lambda^K$ be its basis, and let $g_i^1, \dots, g_i^K$ be coordinates of the vector of individual scores $\mathbf{g}_i$ in this basis. Often, for an appropriate choice of basis, g_i^k may be interpreted as a partial membership of object i in cluster k . Alternatively, a crisp or fuzzy clustering algorithm may be applied to individual scores $\mathbf{g}_i$ to obtain other classifications. The low computational complexity of the GoM algorithm makes it very attractive for analyzing data involving a large number (hundreds or thousands) of categorical variables.

Example: Analysis of Gene Expression Data

We used as our example gene expression data – the basis of Figure 2 in [1]. These authors performed a hierarchical cluster analysis on 2427 genes in the yeast *S. cerevisia*. Data were drawn at time points during several processes given in Table 1, taken from footnotes in [1]: for example, cell division after synchronization by alpha factor arrest (ALPH; 18 time points) after centrifugal elutriation (ELU; 14 time points).

Gene expression (log ratios) was measured for each of these time points and subjected to hierarchical cluster analysis. For details of their cluster analysis method, see [1]. They took the results of their cluster analysis and made plots of genes falling in various clusters. Their plots consisted of raw data values (log ratios) to which they assigned a color varying from saturated green at the small end of the scale to saturated red at the high end of the scale. The resulting plots exhibited large areas similarly colored. These areas indicated genes that clustered together.

6 Fuzzy Cluster Analysis

Table 1 Cell division processes

Process type	Time points	Process description
ALPH	18	Synchronization by alpha factor arrest
ELU	14	Centrifugal elutriation
CDC15	15	Temperature-sensitive mutant
SPO	11	Sporulation
HT	6	High-temperature shock
Dtt	4	Reducing agents
DT	4	Low temperature
DX	7	Diauxic shift

The primary purpose of this example is to describe the use of GoM for fuzzy cluster analysis. A secondary purpose was to identify some genes that cluster together and constitute an interesting set. To use the Grade of Membership (GoM) model, we categorized the data by breaking each range of expression into 5 parts – roughly according to the empirical distribution. GoM constructs K groups (types) types with different characteristics to explain heterogeneity of the data. The product form of the multinomial GoM likelihood for categorical variables x_{ij} is:

$$L = \prod_L \prod_j \prod_\ell (p_{ij\ell})^{y_{ij\ell}}, \quad (11)$$

where

$$p_{ij\ell} = \sum_{h=1}^K g_{ih} \lambda_{hj\ell} \quad (12)$$

with constraints $g_{ih}, \lambda_{hj\ell} \geq 0$ for all i, h, j, ℓ and $\sum_{h=1}^K g_{ih} = 1$ for all i ; $\sum_{\ell=1}^{L_j} \lambda_{hj\ell} = 1$ for all h, j . $y_{ij\ell}$ is the binary coding of x_{ij} .

In (11), $p_{ij\ell}$ is the probability that observation i gives rise to the response that ℓ for variable j ; $\lambda_{hj\ell}$ is the probability that an observation belonging exclusively to type h gives rise to the response ℓ for variable j ; g_{ih} is the degree to which observation i belongs to type h ; L_j is the number of possible values for variable j ; and K is the number of types needed to fully characterize the data and must be specified. The parameters $\{g_{ih}, \lambda_{hj\ell}\}$ are estimated from (11) by the principle of maximum likelihood.

In practice, one starts with a low value of $K = K_0$, usually 4 or 5. Then, analogous to the way one fits a polynomial, successively higher values of K are tried until increasing K does not improve the fit. One of the important features of GoM is that, by inspecting

the whole ensemble of $\lambda_{hj\ell}$ for one value of K , one can construct a description of each type that usually ‘makes sense’. For most of the analyses done with GoM, K works out to be between 5 and 7. For this analysis runs for K as high as 15 were done before settling on $K = 10$ for this analysis. The program DSIGoM available from dsisoft.com was used for the computations. Further analysis was done by standard statistical programs.

The sizes of the types are 205.8, 364.1, 188.7, 234.1, 386.1, 202.4, 211.2, 180.0, 187.9, and 207.2 for types 1 through 10 respectively. Although the GoM analysis makes a partial assignment of each observation or case to the 10 types, it is sometimes desirable to have crisp assignment. A forced crisp assignment can be made by assigning the case to the type k^* with the largest membership. For the i th case, define k_i^* such that $g_{ik}^* > g_{ih}$ for all $h \neq k^*$.

The GoM output consists of the profiles (or variable clusters) characterized by the $\{\lambda_{hj\ell}\}$ and the grades of membership $\{g_{ih}\}$ values along with several goodness-of-fit measures. To construct clusters from the results of the GoM analysis, one can compare the Empirical Cumulative Distribution Function (CDF) for each type with the CDF for the population frequencies. This is done for each variable in the analysis. Based on the CDFs for each variable, a determination was made whether the type upregulated or downregulated gene expression more than the population average. Results of this process are given in Table 2.

‘D’ means that the type downregulates gene expression *for that variable* more than the population. A value of ‘U’ indicates the type upregulates gene expression for the variable more than the population average. A ‘blank’ value indicates that gene regulation for the type did not differ from the population average. In Table 2, Type 4 downregulates ALPH stress test expression values for all experiment times while Type 8 upregulates the same expression values for all experiment times. Types 4 and 8 represent different dimensions of the sample space. Heterogeneity of the data is assumed to be fully described by the 10 types. Descriptions for all 10 types are given in Table 3.

The g_{ih} represent the degree to which the i th case belongs to the h th type or cluster. Each case was assigned the type with the largest g_{ih} value. The 79 data values for each case were assigned colors according to the scheme used by [1] and plotted. We

Table 2 Gene expression indication by process and type

	I	II	III	IV	V	VI	VII	VIII	IX	X
Variable										
Alpha 0	a			D	D		U	U		
Alpha 7	D			D			U	U		
Alpha 14	D			D	D		U	U		
Alpha 21	D			D			U	U		
Alpha 28				D		D	U	U		
Alpha 35	D			D			U	U		
Alpha 42		U		D		D	U	U		
Alpha 49				D				U		
Alpha 56				D				U		
Alpha 63				D			U	U		U
Alpha 70			D	D				U	U	
Alpha 77			D	D				U	U	
Alpha 84				D				U	U	
Alpha 91		D		D				U		
Alpha 98				D				U	D	U
Alpha 105				D				U	D	U
Alpha 112		U		D	D	D		U	U	
Alpha 119				D				U	D	U
Elu 0	D			D			U	D	U	D
Elu 30	D		D	D	U	U	U	D		U
Elu 60	D	D	D	D		U	U			U
Elu 90		D	D	D				U		U
Elu 120	U	D	D	D			D			U
Elu 150	U	D	D				D			U
Elu 180		D	D	D			D	U		
Elu 210		D	D				U	U		U
Elu 240	D	D	D				U			U
Elu 270	D	D	D				U			
Elu 300	D	D	D	U					U	U
Elu 330	D	D	D				U			
Elu 360	D		D				U			
Elu 390	D	D	D				U			
CDC_15 10	U		D					D	U	
CDC_15 30	U			D		U	D	D	U	
CDC_15 50	U		D			U	D		U	
CDC_15 70		D	U			D	U	U	U	
CDC_15 90	U		D	D			D		U	U
CDC_15 110		D	D					U	U	
CDC_15 130		D	D				U	U		
CDC_15 150	D			U	D		U	U	U	
CDC_15 170	D		D	U	D		U		U	
CDC_15 190		D		D			U	U		
CDC_15 210		D				U		U		
CDC_15 230	D	D						D	U	
CDC_15 250	D					U		U	U	
CDC_15 270	D		D	U			U		U	
CDC_15 290	D	U	D				U		U	
spo 0	D	D	D	D		U	U	U		U
spo 2	D					U	D	U		D
spo 5	D	U	U			U	D		D	
spo 7	D	U	U			U	D			D
spo 9	D		U			U	D	U		D

Table 2 (continued)

	I	II	III	IV	V	VI	VII	VIII	IX	X
spo 11	D					U	D			
spo5 2	D		D			U	U	U		U
spo5 7	D					U	D			
spo5 11						U	D			U
spo-erl	D		U	U		U	D	U	D	D
spo-mid	D		U			U	D	U	D	D
heat 0	D	D	U	D			U	U		
heat 10	D		U				U	D		D
heat 20	D			D						D
heat 40	D		U				U			D
heat 80	D		U				U		D	D
heat 160	D		U	D	U		U			D
dtl 15			D				U		U	U
dtl 30	D	U			D			D	U	U
dtl 60	D		D				U	U	U	D
dtl 120	D		D				U			D
cold 0			U	D			U			
cold 20	D		U					D		D
cold 40	D		U	D	U		U			D
cold 160	D				U		U			D
diau a		D	D	D			U	U	D	U
diau b			D	D			U	U	D	
diau c		D	D	D	D		U	U		U
diau d	D						U		U	
diau e	D				D	D	U		U	
diau f	D						U		U	D
diau g	D		U					D	U	D

^aVariables left blank are not different from the population values.

include the plots for Type 4 and Type 8. These plots are designated Figures 8 and 9 respectively. Since the figures are illustrative, not substantive, actual gene names are not indicated. Inspection of Figure 8 shows large areas of green for the ALPH stress tests and for the earlier Elu tests, indicating clusters of downregulated genes. There are large areas of green for heat and cold tests, indicating downregulation. One can also see smaller clusters of reddish areas, indicating upregulated genes connected to later time points for Elu tests. In Figure 9 for Type 8, there is a significant clustering of pink and red areas for ALPH indicating a group of upregulated genes. The same genes are downregulated for the early test times of Elu and CDC 15, indicated by splotches of green.

The clusters designated Type 4 and Type 8 are not the same size. In a GoM model, the sizes of the types are computed by summing all g_{ih} values over i giving 205.0 for Type 4 and 180.0 for Type 8.

Table 3 Summary type descriptions

Type 1	Downregulates early ALPH time points, late CDC_15, Spo, heat, dtt, cold, diauxic. Upregulates Elu.
Type 2	Gene expression for ALPH does not deviate substantially from the population. * Downregulates Elu; does not convincingly regulate other processes.
Type 3	Downregulates Elu, CDC_15, dtt, early diauxic. Upregulates mid-range Spo, heat, cold.
Type 4	Downregulates ALPH, early Elu, heat cold, early diauxic. For CDC_15, it appears to downregulate in the mid-range but upregulate in the extremes.
Type 5	With spotty indication of both upregulation and downregulation; it most resembles the population.
Type 6	Strongly upregulates Spo.
Type 7	Upregulates early ALPH, early and late Elu, early and late CDC. Downregulates Spo. Upregulates heat, dtt, cold, diauxic.
Type 8	Upregulates ALPH, some indication for Elu, CDC_15, Spo, and early diauxic.
Type 9	Upregulates CDC_15, dtt, Spo mid and early, diauxic.
Type 10	Upregulates mid and late ALPH, Elu, possibly early CDC_15, early dtt, early diauxic. Downregulates early and mid Spo, heat, late dtt, cold, late diauxic.

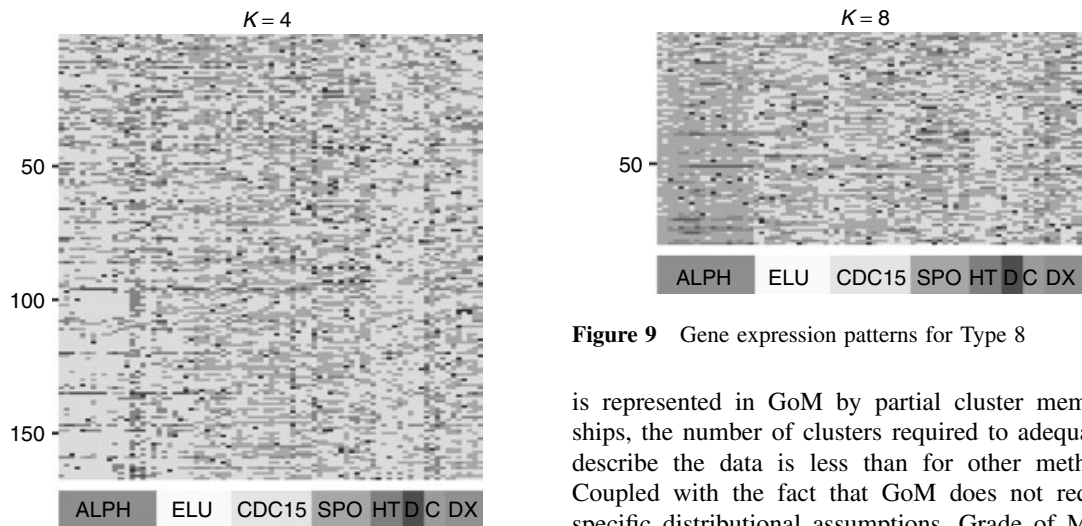


Figure 8 Gene expression patterns for Type 4

Figure 9 Gene expression patterns for Type 8

Summary

Hierarchical clustering methods require choosing a distance function value to determine how many leaf level clusters will be in the analysis. There is no available statistical test to determine how many clusters are required (*see Number of Clusters*). Grade of Membership Analysis determines the number of clusters required by standard statistical tests with known properties. Because individual variability in the data

is represented in GoM by partial cluster memberships, the number of clusters required to adequately describe the data is less than for other methods. Coupled with the fact that GoM does not require specific distributional assumptions, Grade of Membership Analysis would be usually be preferred on parsimonious and theoretical grounds. GoM simultaneously clusters both variables and cases (a property it shares with latent class analysis). The λ_{hjl} define variable clusters and g_{ih} define clusters of cases. The g_{ih} define the degree to which each case belongs to each of the clusters.

References

- [1] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceeding of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

-
- [2] Everitt, B.S. & Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman & Hall, New York.
 - [3] Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models*, Sage Publications, Thousand Oaks.
 - [4] Klawonn, F. & Höppner, F. (2003). What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier, in *Advances in Intelligent Data Analysis*, M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse & C. Borgelt, eds, Verlag Springer, Berlin, pp. 254–264.
 - [5] Kovtun, M., Akushevich, I., Manton, K.G. & Tolley, H.D. (2004). Grade of membership analysis: one possible approach to foundations, *Focus on Probability Theory*, Nova Science Publishers, New York. (to be published in 2005.).
 - [6] Langeheine, R. & Rost, J. (1988). *Latent Trait and Latent Class Models*, Plenum Press, New York.
 - [7] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton-Mifflin, Boston.
 - [8] Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications*, Institute of Mathematical Statistics, Hayward.
 - [9] Manton, K., Woodbury, M. & Tolley, H. (1994). *Statistical Applications Using Fuzzy Sets*, Wiley Interscience, New York.
 - [10] Titterton, D.M., Smith, A.F. & Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distribution*, John Wiley & Sons, New York.
 - [11] Woodbury, M.A., Clive J. (1974). Clinical pure types as a fuzzy partition, *Journal of Cybernet* **4**, 111–121.

Further Reading

- Boreiko, D. & Oesterreichische Nationalbank. (2002). *EMU and Accession Countries: Fuzzy Cluster Analysis of Membership*, Oesterreichische Nationalbank, Wien.
- Halstead, M.H. (1977). *Elements of Software Science*, Elsevier Science, New York.
- Hartigan, J. (1975). *Clustering Algorithms*, Wiley, New York.
- Höppner, F. (1999). *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition*, John Wiley, Chichester; New York.
- Jordan, B.K. (1986). A fuzzy cluster analysis of antisocial behavior: implications for deviance theory, Dissertation, Duke University, Durham North Carolina.
- Lance, G.N. & William, W.T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems, *Computer Journal* **9**, 373–380.
- Ling, R.F. (1973). A Probability Theory of Cluster Analysis, *Journal of the American Statistical Association* **68**(341), 159–164.

KENNETH G. MANTON, GENE LOWRIMORE,
ANATOLI YASHIN AND MIKHAIL KOVTUN

Galton, Francis

ROGER THOMAS

Volume 2, pp. 687–688

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Galton, Francis

Born: February 16, 1822, in Birmingham, UK.

Died: January 17, 1911, in Surrey, UK.

Sir Francis Galton (knighted in 1909) and Charles Darwin were grandsons of Erasmus Darwin, Darwin by first wife, Mary, and Galton by second wife, Elizabeth [3]. After an unsuccessful stint in the medical school of Kings College, London, Galton enrolled at Trinity College, Cambridge, and studied mathematics. Instead of working for honors, a ‘breakdown’ that he attributed to ‘overwork’ justified taking a ‘poll degree’ [4]. Nevertheless, Galton’s intellect (estimated IQ of 200; see [6]) and inherited financial independence enabled him to become so accomplished that in his obituary in *Nature* he was ranked among such leading nineteenth-century British scientists as Darwin, Kelvin, Huxley, and Clerk-Maxwell [1].

Galton had more than 300 publications including 17 books (see [4, Appendix III]) with *Hereditary Genius* [5] being one of the most important. He later regretted using the term ‘genius’, preferring instead a statistically defined ‘eminence’. His honors, interests, and inventiveness ranged widely. He received a gold medal and fellowship in the Royal Society for his geographical explorations in Africa. Fundamental contributions in meteorology included weather mapping and establishing the existence of *anti-cyclones*, a term he coined. He constructed methods for physical and psychological measurement, including composite photography of faces, in anthropology. He developed ways to identify and compare fingerprints as used in identification/investigation today. He did pioneering studies of mental measurement in psychology. He studied the efficacy of prayer, and introspected his self-induced worship of idols and self-induced paranoia. His inventions included a supersonic whistle, diving spectacles, and a periscope for peering over crowds.

In genetics, although he erred in adopting Darwin’s views of genetic blending, Galton anticipated Mendel’s work on particulate inheritance, including the distinction between **genotype** and phenotype (which Galton termed *latent* and *patent*). He was the first to use twins in investigations of morphological and behavioral genetics. He coined the term *eugenic*, and most of his work in genetics was done to support

programs to foster human eugenics. Although much tainted today, eugenics was widely popular in Great Britain, the United States, and elsewhere during Galton’s time, and, of course, it persists in milder forms today (standardized testing for university admissions, scholarships, etc.).

Galton’s contributions to statistical theory and methods were primarily in conjunction with genetics and psychology. He reversed the usual applications of the Gaussian Law of Errors to reduce variability and, instead, emphasized the importance of variability itself, which led to new directions for biological and psychological research [6]. In 1877, Galton published a numerical measure of ‘reversion’ or ‘regression’ to express relationships between certain parent–child physical characteristics. Plotting such data graphically led to his publication (1885) of the ‘elliptic contour’ (see graph in [5, p. 191]) and that led directly to his 1888 paper, ‘Co-relations and their measurement, chiefly from anthropometric data’ (see [5]). This paper provided the first means to calculate a coefficient of correlation.

Galton used the ‘*r*’ from his earlier work on ‘regression’ to symbolize the correlation coefficient, and he introduced the familiar way of expressing such coefficients as decimal fractions ranging from -1.0 to $+1.0$. However, he used the interquartile distance as his measure of variation, which would be replaced by the standard deviation in Karl **Pearson’s product-moment correlation coefficient**.

Galton’s legacy in genetics and statistics was carried forward by his friend Karl Pearson in the following ways: first, they combined Galton’s Eugenics Record Office with Pearson’s Biometric Laboratory to establish the Galton Laboratory at University College, London. Galton then provided funds to establish the journal *Biometrika*, and by his will following his death, he funded the Galton National Chair in Eugenics. He expressed his wish, which was honored, that Pearson be the first such professor [2]. As holder of the Galton chair, Pearson formed a department of genetics and statistics, a legacy that remains today as two separate departments in University College, London [7].

References

- [1] *Nature*. (1911). Francis Galton (February 16, 1822–January 17, 1911), **85**, 440–445.

2 Galton, Francis

- [2] Darwin, G.H. (1912). Galton, Sir Francis (1822–1911), *The Dictionary of National Biography, Supplement January 1901-December 1911*, Oxford University Press, Oxford.
- [3] Fancher, R.E. (1996). *Pioneers of Psychology*, 3rd Edition, Norton, New York.
- [4] Forrest, D.W. (1974). *Francis Galton: The Life and Work of a Victorian Genius*, Taplinger Publishing, New York.
- [5] Galton, F. (1869). *Hereditary Genius*, Watt, London
- [6] Gridgeman, N.T. (1972). Galton, Francis, *Dictionary of Scientific Biography* **5**, 265–267.
- [7] Smith, C.A.B. (1983). Galton, Francis, *Encyclopedia of Statistical Sciences* **3**, 274–276.

ROGER THOMAS

Game Theory

ANDREW M. COLMAN

Volume 2, pp. 688–694

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Game Theory

Background

Game theory is a formal theory of interactive decision making, used to model any decision involving two or more decision makers, called *players*, each with two or more ways of acting, called *strategies*, and well-defined preferences among the possible outcomes, represented by numerical *payoffs*.

In the theory, a player can represent an individual human decision maker or a corporate decision-making body, such as a committee or a board. A (pure) strategy is a complete plan of action, specifying in advance what *moves* a player would make – what actions the player would perform – in every contingency that might arise. If each player has only one decision to make and the players decide simultaneously, then the concept of a strategy coincides with that of a move, but in more complicated cases, a strategy is a comprehensive plan specifying moves in advance, taking account of all possible counter-moves of the coplayer(s). A *mixed strategy* is a probability distribution over a player's set of pure strategies. It is usually interpreted as a strategy chosen at random, with a fixed probability assigned to each pure strategy, as when a player tosses a coin to choose between two pure strategies with equal probabilities. In Bayesian interpretations of game theory, initiated by Harsanyi [12], a mixed strategy is sometimes interpreted as a coplayer's uncertainty about a strategy choice. Payoffs represent players' von Neumann–Morgenstern utilities, which are (roughly) their true preferences on an interval scale of measurement, as revealed by the assumedly consistent choices that they make in lotteries in which the outcomes have known probabilities assigned to them. A player's *payoff function* is a mapping that assigns a specific payoff to each outcome of the game.

The conceptual groundwork of game theory was laid by Zermelo, Borel, von Neumann, and others in the 1920s and 1930s (*see* [10]), and the first fully developed version of the theory appeared in *Theory of Games and Economic Behavior* by von Neumann and Morgenstern [31]. The theory began to have a significant impact on the behavioral and social sciences after the publication in 1957 of a more accessible text entitled *Games and Decisions* by Luce and Raiffa [18]. The early game theorists considered

the chief goal of the theory to be that of prescribing what strategies rational players *ought* to choose to maximize their payoffs. In this sense, the theory, in its classical form, is primarily *normative* rather than *positive* or *descriptive*. An additional rationality assumption, that people generally try to do the best for themselves in any given circumstances, makes the theory relevant to the empirical behavioral sciences and justifies experimental games (*see* section titled Experimental Games below); and in evolutionary game theory, the rationality assumption is replaced by replicator dynamics or adaptive learning mechanisms (*see* section titled Evolutionary Game Theory).

Basic Assumptions

In conventional decision theory, rational choice is defined in terms of maximizing expected utility (EU), or subjective expected utility (SEU), where the objective probabilities of outcomes are unknown (*see* utility theory). But this approach is problematic in games because each player has only partial control over the outcomes, and it is generally unclear how a player should choose in order to maximize EU or SEU without knowing how the other player(s) will act. Game theory, therefore, incorporates not only rationality assumptions in the form of expected utility theory, but also common knowledge assumptions, enabling players to anticipate one another's strategies to some extent, at least. The standard common knowledge and rationality (CKR) assumptions are as follows:

CKR1 (common knowledge): The specification of the game, including the players' strategy sets and payoff functions, is *common knowledge* in the game, together with everything that can be deduced logically from it and from the rationality assumption CKR2.

CKR2 (rationality): The players are rational in the sense of expected utility theory – they always choose strategies that maximize their individual expected payoffs, relative to their knowledge and beliefs – and this is *common knowledge* in the game.

The concept of *common knowledge* was introduced by Lewis [16] and formalized by Aumann [1]. A proposition is common knowledge if every player knows it to be true, knows that every other player knows it to be true, knows that every other player knows that every other player knows it to be true, and

so on. This is an everyday phenomenon that occurs, for example, whenever a public announcement is made, so that everyone present not only knows it, but knows that others know it, and so on [21].

Key Concepts

Other key concepts of game theory are most easily explained by reference to a specific example. Figure 1 depicts the best known of all strategic games, the *Prisoner's Dilemma game*. The figure shows its *payoff matrix*, which specifies the game in *normal form* (or *strategic form*), the principal alternative being *extensive form*, which will be illustrated in the section titled Subgame-perfect and Trembling-hand Equilibria. Player I chooses between the rows labeled *C* (cooperate) and *D* (defect), Player II chooses between the columns labeled *C* and *D*, and the pair of numbers in each cell represent the payoffs to Player I and Player II, in that order by convention. In *noncooperative game theory*, which is being outlined here, it is assumed that the players choose their strategies simultaneously, or at least independently, without knowing what the coplayer has chosen. A separate branch of game theory, called *cooperative game theory*, deals with games in which players are free to share the payoffs by negotiating coalitions based on binding and enforceable agreements. The rank order of the payoffs, rather than their absolute values, determines the strategic structure of a game. Replacing the payoffs 5, 3, 1, 0 in Figure 1 by 4, 3, 2, 1, respectively, or by 10, 1, -2, -20, respectively, changes some properties of the game but leaves its strategic structure (Prisoner's Dilemma) intact.

The Prisoner's Dilemma game is named after an interpretation suggested in 1950 by Tucker [30] and popularized by Luce and Raiffa [18, pp. 94–97]. Two people, charged with joint involvement in a crime, are held separately by the police, who have insufficient evidence for a conviction unless at least one of them discloses incriminating evidence. The

police offer each prisoner the following deal. If neither discloses incriminating evidence, then both will go free; if both disclose incriminating evidence, then both will receive moderate sentences; and if one discloses incriminating evidence and the other conceals it, then the former will be set free with a reward for helping the police, and the latter will receive a heavy sentence. Each prisoner, therefore, faces a choice between cooperating with the coplayer (concealing the evidence) and defecting (disclosing it). If both cooperate, then the payoffs are good for both (3, 3); if both defect, then the payoffs are worse for both (1, 1); and if one defects while the other cooperates, then the one who defects receives the best possible payoff and the cooperator the worst, with payoffs (5, 0) or (0, 5), depending on who defects.

This interpretation rests on the assumption that the utility numbers shown in the payoff matrix do, in fact, reflect the prisoners' preferences. Considerations of loyalty and a reluctance to betray a partner-in-crime might reduce the appeal of being the sole defector for some criminals, in which case that outcome might not yield the highest payoff. But the payoff numbers represent von Neumann–Morgenstern utilities, and they are, therefore, assumed to reflect the players' preferences after taking into account such feelings and everything else affecting their preferences. Many everyday interactive decisions involving cooperation and competition, trust and suspicion, altruism and spite, threats, promises, and commitments turn out, on analysis, to have the strategic structure of the Prisoner's Dilemma game [7]. An obvious example is price competition between two companies, each seeking to increase its market share by undercutting the other.

How should a rational player act in a Prisoner's Dilemma game played once? The first point to notice is that *D* is a *best reply* to both of the coplayer's strategies. A best reply (or best response) to a coplayer's strategy is a strategy that yields the highest payoff against that particular strategy. It is clear that *D* is a best reply to *C* because it yields a payoff of 5, whereas a *C* reply to *C* yields only 3; and *D* is also a best reply to *D* because it yields 1 rather than 0. In this game, *D* is a best reply to both of the coplayer's strategies, which means that defection is a best reply whatever the coplayer chooses. In technical terminology, *D* is a *dominant strategy* for both players. A dominant strategy is one that is a best

		II	
		C	D
I	C	3,3	0,5
	D	5,0	1,1

Figure 1 Prisoner's Dilemma game

reply to *all* the strategies available to the coplayer (or coplayers, if there are several).

Strategic dominance is a decisive argument for defecting in the one-shot Prisoner's Dilemma game – it is in the rational self-interest of each player to defect, whatever the other player might do. In general, if a game has a dominant strategy, then a rational player will certainly choose it. A dominated strategy, such as *C* in the Prisoner's Dilemma game, is *inadmissible*, inasmuch as no rational player would choose it. But the Prisoner's Dilemma game embodies a genuine paradox, because if both players cooperate, then each receives a better payoff (each gets 3) than if both defect (each gets 1).

Nash Equilibrium

The most important “solution concept” of game theory flows directly from best replies. A *Nash equilibrium* (or *equilibrium point* or simply *equilibrium*) is an outcome in which the players' strategies are best replies to each other. In the Prisoner's Dilemma game, joint defection is a Nash equilibrium, because *D* is a best reply to *D* for both players, and it is a unique equilibrium, because no other outcome has this property. A Nash equilibrium has strategic stability, because neither player could obtain a better payoff by choosing differently, given the coplayer's choice, and the players, therefore, have no reason to regret their own choices when the outcome is revealed.

The fundamental theoretical importance of Nash equilibrium rests on the fact that if a game has a uniquely rational solution, then it must be a Nash equilibrium. Von Neumann and Morgenstern [31, pp. 146–148] established this important result via a celebrated *indirect argument*, the most frequently cited version of which was presented later by Luce and Raiffa [18, pp. 63–65]. Informally, by CKR2, the players are expected utility maximizers, and by CKR1, any rational deduction about the game is common knowledge. Taken together, these premises imply that, in a two-person game, if it is uniquely rational for the players to choose particular strategies, then those strategies must be best replies to each other. Each player can anticipate the coplayer's rationally chosen strategy (by CKR1) and necessarily chooses a best reply to it (by CKR2); and because the strategies are best replies to each other, they are in

Nash equilibrium by definition. A uniquely rational solution must, therefore, be a Nash equilibrium.

The indirect argument also provides a proof that a player cannot solve a game with the techniques of standard (individual) decision theory (*see* strategies of decision making) by assigning subjective probabilities to the coplayer's strategies as if they were states of nature and then simply maximizing SEU. The proof is by *reductio ad absurdum*. Suppose that a player were to assign subjective probabilities and maximize SEU in the Prisoner's Dilemma game. The specific probabilities are immaterial, so let us suppose that Player I, for whatever reason, believed that Player II was equally likely to choose *C* or *D*. Then, Player I could compute the SEU of choosing *C* as $1/2(3) + 1/2(0) = 1.5$, and the SEU of choosing *D* as $1/2(5) + 1/2(1) = 3$; therefore, to maximize SEU, Player I would choose *D*. But if that were a rational conclusion, then by CKR1, Player II would anticipate it, and by CKR2, would choose (with certainty) a best reply to *D*, namely *D*. This leads immediately to a contradiction, because it proves that Player II was not equally likely to choose *C* or *D*, as assumed from the outset. The only belief about Player II's choice that escapes contradiction is that Player II will choose *D* with certainty, because joint defection is the game's unique Nash equilibrium.

Nash proved in 1950 [22] that every game with a finite number of pure strategies has at least one equilibrium point, provided that the rules of the game allow mixed strategies to be used. The problem with Nash equilibrium as a solution concept is that many games have multiple equilibria that are nonequivalent and noninterchangeable, and this means that game theory is systematically indeterminate. This is illustrated in Figure 2, which shows the payoff matrix of the *Stag Hunt game*, first outlined in 1755 by Rousseau [26, Part II, paragraph 9], introduced into the literature of game theory by Lewis [16, p. 7], brought to prominence by Aumann [1], and discussed

		II	
		<i>C</i>	<i>D</i>
I	<i>C</i>	9,9	0,8
	<i>D</i>	8,0	7,7

Figure 2 Stag Hunt game

in an influential book by Harsanyi and Selten [13, pp. 357–359]. It is named after Rousseau’s interpretation of it in terms of a hunt in which joint cooperation is required to catch a stag, but each hunter is tempted to go after a hare, which can be caught without the other’s help. If both players defect in this way, then each is slightly less likely to succeed in catching a hare, because they may end up chasing the same one.

This game has no dominant strategies, and the (C, C) and (D, D) outcomes are both Nash equilibria because, for both players, C is the best reply to C , and D is the best reply to D . In fact, there is a third Nash equilibrium – virtually all games have odd numbers of equilibria – in which both players use the mixed strategy $(7/8C, 1/8D)$, yielding expected payoffs of $63/8$ to each. The existence of multiple Nash equilibria means that formal game theory specifies no rational way of playing this game, and other psychological factors are, therefore, likely to affect strategy choices.

Payoff Dominance

Inspired by the Stag Hunt game, and in an explicit attempt to provide a method for choosing among multiple equilibria, Harsanyi and Selten’s *General Theory of Equilibrium Selection in Games* [13] introduced as axioms two auxiliary principles. The first and most important is the *payoff-dominance* principle, not to be confused with strategic dominance, discussed in the section titled Key Concepts. If e and f are two equilibria in a game, then e payoff-dominates (or Pareto-dominates) f if, and only if, e yields a strictly greater payoff to every player than f does. The payoff-dominance principle is the proposition that if one equilibrium payoff-dominates all others in a game, then the players will play their parts in it by choosing its strategies. Harsanyi and Selten suggested that this principle should be regarded as part of every player’s ‘concept of rationality’ and should be common knowledge among the players.

In the Stag Hunt game, (C, C) payoff-dominates (D, D) , and it also payoff-dominates the mixed-strategy equilibrium in which both players choose $(7/8C, 1/8D)$; therefore, the payoff-dominance principle requires both players to choose C . But this assumption requires collective reasoning that goes beyond the orthodox rationality assumption of CKR2.

Furthermore, it is not intuitively obvious that players should choose C , because, by so doing, they risk the worst possible payoff of zero. The D strategy is a far safer choice, risking a worst possible payoff of 7. This leads naturally to Harsanyi and Selten’s secondary criterion of selection among multiple equilibria, called the *risk-dominance* principle, to be used only if payoff dominance fails to yield a determinate solution. If e and f are any two equilibria in a game, then e risk-dominates f if, and only if, the minimum possible payoff resulting from the choice of e is strictly greater than the minimum possible payoff resulting from the choice of f , and players who follow the risk-dominance principle choose its strategies. In the Stag Hunt game, D risk-dominates C for each player, but the payoff-dominance principle takes precedence, because, in this game, it yields a determinate solution.

Subgame-perfect and Trembling-hand Equilibria

Numerous refinements of the Nash equilibrium concept have been suggested to deal with the problem of multiple Nash equilibria and the consequent indeterminacy of game theory. The most influential of these is the *subgame-perfect equilibrium*, introduced by Selten [27]. Selten was the first to notice that some Nash equilibria involve strategy choices that are clearly irrational when examined from a particular point of view. A simple example is shown in Figure 3.

The payoff matrix in Figure 3(a) specifies a game in which both (C, C) and (D, D) are Nash equilibria, but only (C, C) is subgame perfect. The Nash equilibrium (D, D) is not only weak (because 3 is not *strictly* greater than 3 for Player II) but also imperfect, because it involves an irrational choice from

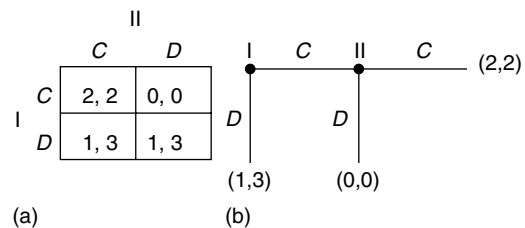


Figure 3 Subgame-perfect equilibrium

Player II. This emerges most clearly from an examination of the *extensive form* of the game, shown in Figure 3(b). The extensive form is a *game tree* depicting the players' moves as if they occurred sequentially. This extensive form is read from Player I's move on the left. If the game were played sequentially, and if the second decision node were reached, then a utility-maximizing Player II would choose *C* at that point, to secure a payoff of 2 rather than zero. At the first decision node, Player I would anticipate Player II's reply, and would, therefore, choose *C* rather than *D*, to secure 2 rather than 1. This form of analysis, reasoning backward from the end, is called *backward induction* and is the basic method of finding subgame-perfect equilibria. In this game, it shows that the (*D*, *D*) equilibrium could not be reached by rational choice in the extensive form, and that means that it is imperfect in the normal form. By definition, a subgame-perfect equilibrium is one that induces payoff-maximizing choices in every branch or subgame of its extensive form.

In a further refinement, Selten [28] introduced the concept of the *trembling-hand equilibrium* to identify and eliminate imperfect equilibria. At every decision node in the extensive form or game tree, there is assumed to be a small probability ε (epsilon) that the player acts irrationally and makes a mistake. The introduction of these error probabilities, generated by a random process, produces a *perturbed game* in which every move that could be played has some positive probability of being played. Assuming that the players' trembling hands are common knowledge in a game, Selten proved that only the subgame-perfect equilibria of the original game remain equilibria in the perturbed game, and they continue to be equilibria as the probability ε tends to zero. According to this widely accepted refinement of the equilibrium concept, the standard game-theoretic rationality assumption (CKR2) is reinterpreted as a limiting case of incomplete rationality.

Experimental Games

Experimental games have been performed since the 1950s in an effort to understand the strategic interaction of human decision makers with bounded rationality and a variety of nonrational influences on their behavior (for detailed reviews, see [6, 11, 15, Chapters 1–4], [17, 25]). Up to the end of the 1970s,

experimental attention focused largely on the Prisoner's Dilemma and closely related games. The rise of behavioral economics in the 1980s led to experiments on a far broader range of games – see [4, 5].

The experimental data show that human decision makers deviate widely from the rational prescriptions of orthodox game theory. This is partly because of bounded rationality and severely limited capacity to carry out indefinitely iterated recursive reasoning ('I think that you think that I think...') (see [8, 14, 29]), and partly for a variety of unrelated reasons, including a strong propensity to cooperate, even when cooperation cannot be justified on purely rational grounds [7].

Evolutionary Game Theory

The basic concepts of game theory can be interpreted as elements of the theory of natural selection as follows. Players correspond to individual organisms, strategies to organisms' genotypes, and payoffs to the changes in their Darwinian fitness – the numbers of offspring resembling themselves that they transmit to future generations. In evolutionary game theory, the players do not choose their strategies rationally, but natural selection mimics rational choice. Maynard Smith and Price [20] introduced the concept of the *evolutionarily stable strategy* (ESS) to handle such games. It is a strategy with the property that if most members of a population adopt it, then no mutant strategy can invade the population by natural selection, and it is, therefore, the strategy that we should expect to see commonly in nature. An ESS is invariably a symmetric Nash equilibrium, but not every symmetric Nash equilibrium is an ESS.

The standard formalization of ESS is as follows [19]. Suppose most members of a population adopt strategy *I*, but a small fraction of mutants or invaders adopt strategy *J*. The expected payoff to an *I* individual against a *J* individual is written $E(I, J)$, and similarly for other combinations strategies. Then, *I* is an ESS if *either* of the conditions (1) or (2) below is satisfied:

$$E(I, I) > E(J, I) \quad (1)$$

$$E(I, I) = E(J, I), \text{ and } E(I, J) > E(J, J) \quad (2)$$

Condition (1) or (2) ensures that J cannot spread through the population by natural selection. In addition, differential and difference equations called *replicator dynamics* have been developed to model the evolution of a population under competitive selection pressures. If a population contains k genetically distinct types, each associated with a different pure strategy, and if their proportions at time t are $x(t) = (x_1(t), \dots, x_k(t))$, then the replicator dynamic equation specifies the population change from $x(t)$ to $x(t+1)$.

Evolutionary game theory turned out to solve several long-standing problems in biology, and it was described by Dawkins as ‘one of the most important advances in evolutionary theory since Darwin’ [9, p. 90]. In particular, it helped to explain the evolution of cooperation and altruistic behavior – conventional (ritualized) rather than escalated fighting in numerous species, alarm calls by birds, distraction displays by ground-nesting birds, and so on.

Evolutionary game theory is also used to study adaptive learning in games repeated many times. Evolutionary processes in games have been studied analytically and computationally, sometimes by running simulations in which strategies are pitted against one another and transmit copies of themselves to future generations in proportion to their payoffs (see [2, 3, Chapters 1, 2; 23, 24]).

References

- [1] Aumann, R.J. (1976). Agreeing to disagree, *Annals of Statistics* **4**, 1236–1239.
- [2] Axelrod, R. (1984). *The Evolution of Cooperation*, Basic Books, New York.
- [3] Axelrod, R. (1997). *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration*, Princeton University Press, Princeton.
- [4] Camerer, C.F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton.
- [5] Camerer, C.F., Loewenstein, G. & Rabon, M., eds (2004). *Advances in Behavioral Economics*, Princeton University Press, Princeton.
- [6] Colman, A.M. (1995). *Game Theory and its Applications in the Social and Biological Sciences*, 2nd Edition, Routledge, London.
- [7] Colman, A.M. (2003a). Cooperation, psychological game theory, and limitations of rationality in social interaction, *The Behavioral and Brain Sciences* **26**, 139–153.
- [8] Colman, A.M. (2003b). Depth of strategic reasoning in games, *Trends in Cognitive Sciences* **7**, 2–4.
- [9] Dawkins, R. (1976). *The Selfish Gene*, Oxford University Press, Oxford.
- [10] Dimand, M.A. & Dimand, R.W. (1996). *A History of Game Theory (Vol 1): From the Beginnings to 1945*, Routledge, London.
- [11] Foddy, M., Smithson, M., Schneider, S. & Hogg, M., eds (1999). *Resolving Social Dilemmas: Dynamic, Structural, and Intergroup Aspects*, Psychology Press, Hove.
- [12] Harsanyi, J.C. (1967–1968). Games with incomplete information played by “Bayesian” players, Parts I–III, *Management Science* **14**, 159–182, 320–334, 486–502.
- [13] Harsanyi, J.C. & Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge.
- [14] Hedden, T. & Zhang, J. (2002). What do you think I think you think? Strategic reasoning in matrix games, *Cognition* **85**, 1–36.
- [15] Kagel, J.H. & Roth, A.E., eds (1995). *Handbook of Experimental Economics*, Princeton University Press, Princeton.
- [16] Lewis, D.K. (1969). *Convention: A Philosophical Study*, Harvard University Press, Cambridge.
- [17] Liebrand, W.B.G., Messick, D.M. & Wilke, H.A.M., eds (1992). *Social Dilemmas: Theoretical Issues and Research Findings*, Pergamon, Oxford.
- [18] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.
- [19] Maynard Smith, J. (1982). *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.
- [20] Maynard Smith, J. & Price, G.R. (1973). The logic of animal conflict, *Nature* **246**, 15–18.
- [21] Milgrom, P. (1981). An axiomatic characterization of common knowledge, *Econometrica* **49**, 219–222.
- [22] Nash, J.F. (1950). Equilibrium points in n -person games, *Proceedings of the National Academy of Sciences of the United States of America* **36**, 48–49.
- [23] Nowak, M.A., May, R.M. & Sigmund, K. (1995). The arithmetics of mutual help, *Scientific American* **272**(6), 76–81.
- [24] Nowak, M.A. & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner’s Dilemma game, *Nature* **364**, 56–58.
- [25] Pruitt, D.G. & Kimmel, M.J. (1977). Twenty years of experimental gaming: critique, synthesis, and suggestions for the future, *Annual Review of Psychology* **28**, 363–392.
- [26] Rousseau, J.-J. (1755). Discours sur l’origine d’inégalité parmi les hommes [discourse on the origin of inequality among men], in *Oeuvres Complètes*, Vol. 3, J.-J. Rousseau, Edition Pléiade, Paris.
- [27] Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit [Game-theoretic treatment of an oligopoly model with demand inertia], *Zeitschrift für die Gesamte Staatswissenschaft* **121**, 301–324, 667–689.
- [28] Selten, R. (1975). Re-examination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4**, 25–55.

- [29] Stahl, D.O. & Wilson, P.W. (1995). On players' models of other players: theory and experimental evidence, *Games and Economic Behavior* **10**, 218–254.
- [30] Tucker, A. (1950/2001). A two-person dilemma, in *Readings in Games and Information*, E. Rasmusen, ed., Blackwell, Oxford, pp. 7–8 (Reprinted from a handout distributed at Stanford University in May 1950).
- [31] von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, 2nd Edition, 1947; 3rd Edition, 1953, Princeton University Press, Princeton.

ANDREW M. COLMAN

Gauss, Johann Carl Friedrich

RANDALL D. WIGHT AND PHILIP A. GABLE

Volume 2, pp. 694–696

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Gauss, Johann Carl Friedrich

Born: April 30, 1777, in Brunswick, Germany.

Died: February 23, 1855, in Göttingen, Germany.

Born to humble means in the Duchy of Brunswick (now in Germany), Carl Friedrich Gauss's contributions spanned a lifetime and earned the epithet 'Prince of Mathematicians'. As a mathematical scientist, often ranked with Archimedes and Newton, Gauss is among the elite of any era. In 1784, he entered a Brunswick gymnasium that recognized his talent. By age 14, the Duke of Brunswick granted Gauss a stipend and he entered the Collegium Carolinum, where he studied modern languages as well as the works of Newton, Euler, and Lagrange. In 1795, he entered the University of Göttingen. Shortly thereafter, entries in Gauss's personal notebooks contained groundbreaking insights into number theory. Within a year, Gauss also constructed a 17-sided regular polygon using only ruler and compass – the first so constructed since antiquity. Gauss returned to Brunswick in 1798 and entered the University of Helmstedt, from which he earned a doctorate the following year. The dissertation, published in 1801, contains Gauss's first proof of the fundamental theorem of algebra. Also in 1801, without publishing his method, Gauss correctly predicted the location of the first-known, recently discovered asteroid, Ceres. His brilliance emerged at an early age.

Gauss's contributions to statistics revolve around the conceptual convergence known as the Gauss–Laplace synthesis. Occurring in the years following 1809, this foundational merger advanced effective methods for combining data with the ability to quantify error. Gauss's role in this synthesis centers in his account of least squares, his use of the normal curve, and the influence this work had on **Pierre-Simon Laplace**.

During the mid-seventeenth century, astronomers wanted to know how best to combine a number of independent but varying observations of the same phenomenon. Among the most promising procedures was the method of least squares (*see* **Least Squares Estimation**), which argues that the minimal distance to the true value of a distribution of observations is the sum of squared deviations from the mean. In

1807, Gauss became director of the University of Göttingen's observatory. Despite early applications of least squares insights, Gauss failed to publish an account until 1809, and even then only in the last chapter of a major contribution to celestial mechanics [2]. A priority dispute arose with Adrien Marie Legendre, who first published a least squares discussion in 1805.

However, the respect afforded Gauss always gave pause to quick dismissals of his priority claims. Even if least squares publication priority belongs to Legendre, Gauss offered a sophisticated elaboration of the method. In addition to developing least squares, the 1809 publication contained another seminal contribution to statistics, that is, use of the normal distribution to describe measurement error. Here, Gauss employed Laplace's probability curve for sums and means to describe the measurement of random deviations around the true measurement of an astronomical event. Because of this insight, by the end of the nineteenth century, what we know today as the normal distribution came to be known as the Gaussian distribution. Although Gauss was not the first to describe the normal curve, he was the first to use it to assign precise probabilities to errors. In honor of this insight, the 10 Deutschmark would one day bear both an image of Gauss and the normal curve's geometric and formulaic expression.

The 1809 Gaussian insights proved a conceptual catalyst. In 1810, Laplace presented what we know today as the **central limit theorem**: the distribution of any sufficiently sampled variable can be expressed as the sum of small independent observations or variables approximating a normal curve. When Laplace read Gauss's 1809 book later that year, he recognized the connection between his theorem, the normal distribution, and least squares estimates. If errors are aggregates, then errors should distribute along the normal curve with least squares providing the smallest expected error estimate. The coming years were a conceptual watershed as additional work by Gauss, Laplace, and others converged to produce the Gauss–Laplace synthesis. As Youden ([5], p. 55) later observed "The normal law of error stands out... as one of the broadest generalizations of natural philosophy... It is an indispensable tool for the analysis and the interpretation of the basic data obtained by observation and experiment". Following 1809, these insights spread from astronomy to physics and the military. Absorption into the social sciences

took more time. By the time the use of least squares flowered in the social sciences, **Galton**, **Pearson**, and **Yule** had uniquely transformed the procedure into the techniques of regression (*see* **Multiple Linear Regression**) and **analysis of variance** [3, 4].

In addition to the Gauss–Laplace synthesis, Gauss’s more general contributions include the fundamental theorems of arithmetic and algebra and development of the algebra of congruence. He published important work on actuarial science, celestial mechanics, differential geometry, geodesy, magnetism, number theory, and optics. He invented a heliotrope, magnetometer, photometer, and telegraph. *Sub rosa*, he was among the first to investigate non-Euclidean geometry and, in 1851, approved Riemann’s doctoral thesis. Indeed a titan of science [1], Gauss was extraordinarily productive throughout his life, although his personal life was not without turmoil. After developing heart disease, Gauss died in his sleep in late February, 1855.

References

- [1] Dunnington, G.W. (1955). *Carl Friedrich Gauss, Titan of Science: A Study of his Life and Work*, Exposition Press, New York.
- [2] Gauss, C.F. (1809/1857/2004). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium*, [Theory of the motion of the heavenly bodies moving about the sun in conic sections] Translated by, C.H. Davis, ed., Dover Publications, Mineola, (Original work published 1809; original translation published 1857).
- [3] Stigler, S.M. (1986). *The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge.
- [4] Stigler, S.M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press, Cambridge.
- [5] Youden, W.J. (1994). *Experimentation and Measurement*, U.S. Department of Commerce, Washington.

RANDALL D. WIGHT AND PHILIP A. GABLE

Gene-Environment Correlation

DANIELLE M. DICK

Volume 2, pp. 696–698

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Gene-Environment Correlation

Nature versus Nurture

The history of behavioral sciences has been plagued by a seeming rivalry between genetic influences and environmental influences as the primary determinant of behavior. Today, there is widespread support for the idea that most behavior results from a complex interaction of genetic and environmental effects (*see Gene-Environment Interaction*). However, it is a mistake to assume that these effects are independent sources of influence. Much of an individual's environment is not random in nature; rather, genes can influence an individual's exposure to certain environments, how that individual experiences the environment, and the degree of influence that certain environments exert [2, 13]. This phenomenon is called *gene-environment correlation*. It refers to the fact that an individual's environment is influenced by their genetic predispositions, making genes and the environment inexorably entwined.

Three specific ways by which genes may exert an effect on the environment have been delineated [9, 14]: (a) Passive gene-environment effects refer to the fact that among biologically related relatives (i.e., nonadoptive families), parents provide not only their child's genotypes, but also their rearing environment. Therefore, the child's genotype and home environment are correlated. (b) Evocative gene-environment effects refer to the idea that individual's genotypes influence the responses they receive from others. For example, a child who is predisposed to having an outgoing, cheerful disposition might be more likely to receive positive attention from others than a child who is predisposed to timidity and tears. A person with a grumpy, abrasive temperament is more likely to evoke unpleasant responses from coworkers and others with whom they interact. Thus, evocative gene-environment effects can influence the way an individual experiences the world. (c) Finally, active gene-environment effects refer to the fact that an individual actively selects certain environments and takes away different things from his/her environment, and these processes are influenced by an individual's genotype. Therefore, an individual predisposed to high sensation-seeking may be more prone to attend

parties and meet new people, thereby actively influencing the environments he/she experiences.

Evidence exists in literature for each of these processes. Support for passive gene-environment effects can be found in a study of more than 600 adoptive families recruited from across the United States [6]. Comparisons of adoptive and biological children's correlations between family functioning and adolescent outcome are informative for examining passive gene-environment effects, because only biological children share genes with their parents and are affected by these passive gene-environment effects (both will be affected by evocative and active gene-environment effects). Correlations between mother ratings of family functioning and child ratings of adjustment were substantially higher in biological offspring than in adoptive offspring, supporting passive gene-environment correlation.

The adoption design (*see Adoption Studies*) has also been used to examine evocative gene-environment effects [7]. With data from the Colorado Adoption Project, adoptive children who were or were not at genetic risk for antisocial behavior, based on their biological mother's antisocial history, were compared. Children who were at risk for antisocial behavior consistently evoked more negative control, as self-reported by their adoptive parents, than did adoptees not at risk, from age 5 to 12. These results suggest an evocative gene-environment effect. Children who were genetically predisposed to higher levels of antisocial behavior displayed more externalizing behavior, which evoked more negative control responses from their parents.

Finally, support for active gene-environment effects can be found in another type of study. In order to study active gene-environment effects, one must study individuals outside the home, to assess the degree to which genes may be actively influencing an individual's selection of various environmental niches [14]. As part of the **Nonshared Environment** in Adolescent Development project, genetic influences on individuals' social interactions with peers and teachers were assessed using a twin, full-sibling, and step-sibling design [5]. Significant genetic influence was found for adolescents' reports of positive interactions with friends and teachers. Additionally, **heritability** estimates were quite high for parents' reports of peer group characteristics, suggesting active gene-environment effects in which an

2 Gene-Environment Correlation

individual's genotype influenced the group of individuals they selected as peers. Peer environments are known to then play a significant role in adolescent outcomes [1].

A number of other findings exist supporting genetic influence on environmental measures. Substantial genetic influence has been reported for adolescents' reports of family warmth [10–12]. Genes have been found to influence the degree of anger and hostility that children receive from their parents [11]. They influence the experience of life events [2, 3], and exposure to trauma [4]. In fact, genetic influence has been found for nearly all of the most widely used measures of the environment [8]. Perhaps, even more convincing is that these environmental measures include not only reports by children, parents, and teachers, but also observations by independent observers [11].

Thus, many sources of behavioral influence that we might consider 'environmental' are actually under a degree of genetic influence. An individual's family environment is correlated with their genotype when they are reared among biological relatives. Furthermore, genes influence an individual's temperament and personality, which impacts both the way that other people react to the individual and the environments that person seeks out and experiences. Thus, environmental experiences are not always random, but can be influenced by a person's genetic predispositions. It is important to note that in standard **twin designs**, the effects of gene-environment correlation are included in the genetic component. For example, if genetic influences enhance the likelihood that delinquent youth seek out other delinquents for their peers, and socialization with these peers further contributes to the development of externalizing behavior, that effect could be subsumed in the genetic component of the model, because genetic effects led to the risky environment, which then influenced behavioral development. Thus, genetic estimates may represent upper bound estimates of direct genetic effects on the disorders because they also include gene-environment correlation effects.

References

- [1] Harris, J.R. (1998). *The Nurture Assumption*, The Free Press, New York.
- [2] Kendler, K.S. (1995). Adversity, stress and psychopathology: a psychiatric genetic perspective, *International Journal of Methods in Psychiatric Research* **5**, 163–170.
- [3] Kendler, K.S., Neale, M., Kessler, R., Heath, A. & Eaves, L. (1993). A twin study of recent life events and difficulties, *Archives of General Psychiatry* **50**, 789–796.
- [4] Lyons, M.J., Goldberg, J., Eisen, S.A., True, W., Tsuang, M.T., Meyer, J.M., Hendersen, W.G. (1993). Do genes influence exposure to trauma? A twin study of combat, *American Journal of Medical Genetics* **48**, 22–27.
- [5] Manke, B., McGuire, S., Reiss, D., Hetherington, E.M. & Plomin, R. (1995). Genetic contributions to adolescents' extrafamilial social interactions: teachers, best friends, and peers, *Social Development* **4**, 238–256.
- [6] McGue, M., Sharma, A. & Benson, P. (1996). The effect of common rearing on adolescent adjustment: evidence from a U.S. adoption cohort, *Developmental Psychology* **32**(6), 604–613.
- [7] O'Connor, T.G., Deater-Deckard, K., Fulker, D., Rutter, M. & Plomin, R. (1998). Genotype-environment correlations in late childhood and early adolescence: antisocial behavioral problems and coercive parenting, *Developmental Psychology* **34**, 970–981.
- [8] Plomin, R. & Bergeman, C.S. (1991). The nature of nurture: genetic influence on "environmental" measures, *Behavioral and Brain Sciences* **14**, 373–427.
- [9] Plomin, R., DeFries, J.C. & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior, *Psychological Bulletin* **84**, 309–322.
- [10] Plomin, R., McClearn, G.E., Pedersen, N.L., Nesselroade, J.R. & Bergeman, C.S. (1989). Genetic influences on adults' ratings of their current family environment, *Journal of Marriage and the Family* **51**, 791–803.
- [11] Reiss, D. (1995). Genetic influence on family systems: implications for development, *Journal of Marriage and the Family* **57**, 543–560.
- [12] Rowe, D. (1981). Environmental and genetic influences on dimensions of perceived parenting: a twin study, *Developmental Psychology* **17**, 209–214.
- [13] Rutter, M.L. (1997). Nature-nurture integration: the example of antisocial behavior, *American Psychologist* **52**, 390–398.
- [14] Scarr, S. & McCartney, K. (1983). How people make their own environments: a theory of genotype-environment effects, *Child Development* **54**, 424–435.

DANIELLE M. DICK

Gene-Environment Interaction

KRISTEN C. JACOBSON

Volume 2, pp. 698–701

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Gene-Environment Interaction

Gene x environment interaction, or GxE, as it is commonly called in behavioral genetic literature, is the acknowledgement that genetic and environmental influences on behaviors and traits do not act additively and independently, but may instead be dependent upon one another. One of the working hypotheses among researchers is the diathesis-stress model [7], whereby genetic variants confer an underlying diathesis, or vulnerability to a certain behavior or trait. These genetic vulnerabilities may only impact upon development when accompanied by a specific environmental stressor. Other hypotheses for GxE interactions include the protective effect of genetic variants on environmental risk, and genetic sensitivity to environmental exposure. There are three main strategies for assessing GxE interactions in behavioral genetic research, primarily through the use of **adoption studies**, **twin studies**, and studies of genotyped individuals. Each method has its relative strengths and weaknesses.

The Adoption Study Method

Some of the earliest examples of gene x environment interaction appear in the adoption literature in the early 1980s. Theoretically, adoption studies are ideal methods for assessing GxE, as they allow for a clean separation of genetic influence (via the biological parent characteristics) from salient environmental characteristics (defined by adoptive parent characteristics). Figure 1 shows an example of gene x environment interaction using the adoption design for the development of adolescent antisocial behavior [1]. In this study, genetic risk was defined as the presence of alcoholism or antisocial behavior in the biological parent, and environmental risk was defined as being raised in an adoptive family with significant psychopathology in adoptive siblings or parents, and/or the presence of adoptive parent marital separation or divorce. Standard multivariate regression analyses (*see* **Multivariate Multiple Regression**) were performed assessing the independent effects of genetic and environmental risk, as

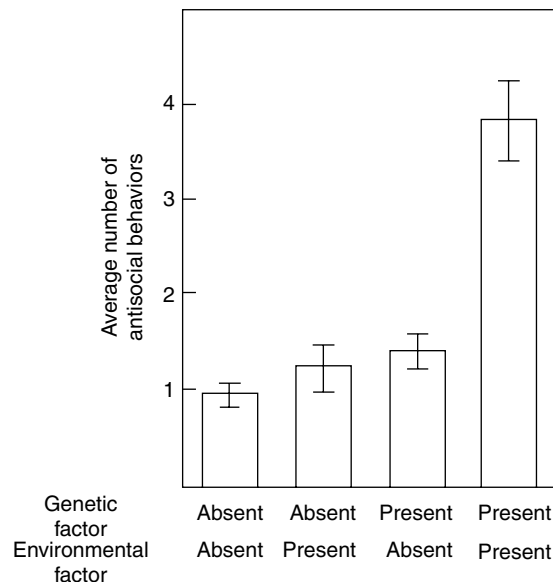


Figure 1 Least-squares means ($\pm$ SE) for simple genetic, environmental, and interaction effects (Iowa 1980 data; $N = 367$). (Figure reproduced from Kluwer Academic Publishers Behavior Genetics, **13**, 1983, p. 308, *Evidence for Gene-environment Interaction in the Development of Adolescent Antisocial Behavior*, R.J. Cadoret, C.A. Cain, and R.R. Crowe, Figure 1, copyright 1983, Plenum Publishing Corporation, with kind permission of Springer Science and Business Media)

well as the interaction term between the two variables. Figure 1 illustrates that neither the presence of genetic risk nor the presence of environmental risk was sufficient, in and of itself, to cause an increase in the average number of antisocial behaviors in adolescent adoptees, compared with adoptees with neither genetic nor environmental risk. In contrast, the presence of *both* genetic and environmental risk factors was related to a higher mean number of antisocial behaviors, compared to all other groups.

Adoption studies have the advantage over other methods that use samples of related individuals of being able to more cleanly separate genetic risk from environmental risk, as adoptees typically have limited or no contact with their biological parents. Thus, in theory, genetic risk in the biological parent is unlikely to be correlated with environmental risk in the adoptive home environment via passive gene-environment

correlation. However, as shown in more recent studies, results from adoption studies can still potentially be confounded by evocative **gene-environment correlation**. For example, Ge et al. [4] reported that hostile and aggressive parenting from the adoptive parent was, in fact, correlated with psychopathology in the biological parent. This relationship was largely mediated through the adoptees' own hostile and aggressive behavior, demonstrating that gene-environment correlation can occur when adoptive parents respond to the behavior of their adopted child (which is, in turn, partly influenced by genetic factors). Thus, genetic-influenced behaviors of the adoptee can evoke a gene-environment correlation. Other limitations to the adoption study method have typically included: (1) the fact that adoptive parents are screened prior to placement, indicating that the range of environmental factors within adoptive samples is likely to be truncated, and that severe environmental deprivation therefore is unlikely; and (2) the limited generalizability of results from adoptive samples to the general population (*see Adoption Studies*).

The Twin Study Method

Twin studies typically estimate the proportion of variation in a given behavior or trait that is due to latent genetic and environmental factors (*see ACE Model*). In GxE studies using twin samples, the central question is generally whether genetic variation on behavior or traits changes across some level of a measured environmental variable. Methods to assess GxE interaction in twin studies include extensions of the DeFries–Fulker regression model (*see DeFries–Fulker Analysis*), testing whether heritabilities (*see Heritability*) are the same or different among individuals in two different groups (e.g., the finding that the heritability of alcohol use is higher among unmarried vs. married women [5]), or through the inclusion of a measured environmental variable as a continuous moderator of genetic and environmental influences in the standard *ACE model*. Examples of replicated GxE interactions using twin data include the finding that the heritability of cognitive ability is greater among adolescents from more advantaged socioeconomic backgrounds [9, 11]. In both of these studies, the absolute magnitude of shared environmental influences on variation was stronger among adolescents from poorer and/or less educated homes.

Conversely, the absolute magnitude of genetic variation was greater among adolescents from higher socioeconomic status families. Both of these factors contributed to the finding that the heritability of cognitive ability (which is defined as the proportion of phenotypic variance attributed to genetic factors) was higher among adolescents in more educated homes.

Advantages to the twin method include the fact that these studies call into question the assumption that heritability is a constant, and can identify salient aspects of the environment that may either promote or inhibit genetic effects. In addition, there are many large twin studies in existence, which make replication of potential GxE interactions possible. The primary disadvantage to this method is that genetic factors are defined as latent variables. Thus, these studies cannot identify the specific genetic variants that may confer greater or lesser risk at different levels of the environment.

Studies of Genotyped Individuals

Arguably, perhaps the 'gold standard' for assessing GxE interaction are studies that investigate whether a specific genetic variant interacts with a measured environmental characteristic. One of the first examples of these studies is the finding that a polymorphism in the monoamine oxidase A (MAOA) gene interacts with child maltreatment to influence mean levels of antisocial behavior [2]. Figure 2 shows the relevant results from this study for four measures of antisocial behavior. As can be seen in this figure, maltreated children with the genetic variant of the MAOA gene that confers high levels of MAOA activity showed mean levels of antisocial behavior that were not significantly different from mean levels of antisocial behavior among non-maltreated children, indicating that this genetic variant had a protective influence against the effects of child maltreatment. Interestingly, although there was a main effect of child maltreatment in these analyses, there was no main effect of the MAOA genotype, indicating that genetic variants that confer low levels of MAOA activity are not, in and of themselves, a risk factor for antisocial behavior.

Advantages to this method are many. Analysis is relatively straightforward, requiring simply the use of multivariate regression techniques. Studies can be done using any sample of genotyped

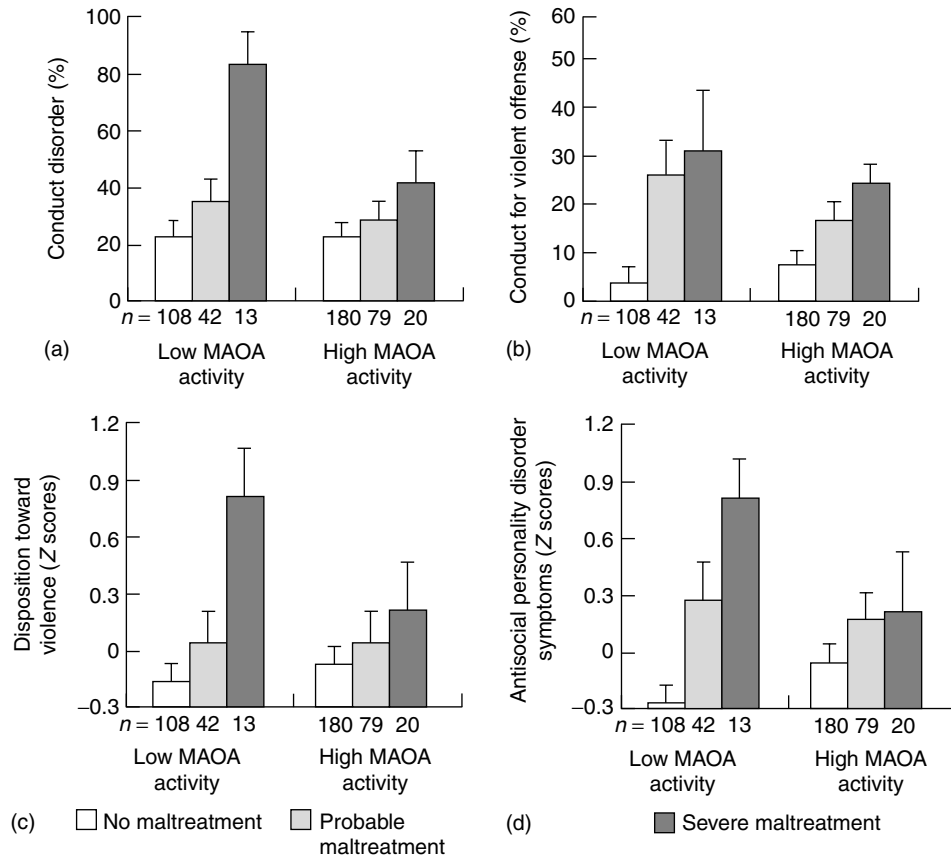


Figure 2 The association between childhood maltreatment and subsequent antisocial behavior as a function of *MAOA* activity. (a) Percentage of males (and standard errors) meeting diagnostic criteria for Conduct Disorder between ages 10 and 18. In a hierarchical logistic regression model, the interaction between maltreatment and *MAOA* activity was in the predicted direction, $b = -0.63$, $SE = 0.33$, $z = 1.87$, $P = 0.06$. Probing the interaction within each genotype group showed that the effect of maltreatment was highly significant in the low-*MAOA* activity group ($b = 0.96$, $SE = 0.27$, $z = 3.55$, $P < 0.001$), and marginally significant in the high-*MAOA* group ($b = 0.34$, $SE = 0.20$, $z = 1.72$, $P = 0.09$). (b) Percentage of males convicted of a violent crime by age 26. The $G \times E$ interaction was in the predicted direction, $b = -0.83$, $SE = 0.42$, $z = 1.95$, $P = 0.05$. Probing the interaction, the effect of maltreatment was significant in the low-*MAOA* activity group ($b = 1.20$, $SE = 0.33$, $z = 3.65$, $P < 0.001$), but was not significant in the high-*MAOA* group ($b = 0.37$, $SE = 0.27$, $z = 1.38$, $P = 0.17$). (c) Mean z scores ($M = 0$, $SD = 1$) on the Disposition Toward Violence Scale at age 26. In a hierarchical ordinary least squares (OLS) regression model, the $G \times E$ interaction was in the predicted direction ($b = -0.24$, $SE = 0.15$, $t = 1.62$, $P = 0.10$); the effect of maltreatment was significant in the low-*MAOA* activity group ($b = 0.35$, $SE = 0.11$, $t = 3.09$, $P = 0.002$) but not in the high-*MAOA* group ($b = 0.12$, $SE = 0.07$, $t = 1.34$, $P = 0.17$). (d) Mean z scores ($M = 0$, $SD = 1$) on the Antisocial Personality Disorder symptom scale at age 26. The $G \times E$ interaction was in the predicted direction ($b = -0.31$, $SE = 0.15$, $t = 2.02$, $P = 0.04$); the effect of maltreatment was significant in the low-*MAOA* activity group ($b = 0.45$, $SE = 0.12$, $t = 3.83$, $P = 0.001$) but not in the high-*MAOA* group ($b = 0.14$, $SE = 0.09$, $t = 1.57$, $P = 0.12$). (Reprinted with permission from Caspi et al. (2002). *Role of Genotype in the Cycle of Violence in Maltreated Children*. Science, **297**, 851–854. Copyright 2002 AAAS)

individuals – there is no special adoptive or family samples required. Because these studies rely on measured genotypes, they can pinpoint more precisely the genetic variants and the potential

underlying biological mechanism that confer risk or protective effects across different environments. On the other hand, the effects of any one individual gene (both additively and/or interactively)

on variation in behavior or traits is likely to be quite small, which may limit the power to detect such interactions in these studies, and may further require some *a priori* knowledge of how specific genes may influence behavior or traits. In addition, because genotypes are inherited from parents, there may be significant gene-environment correlations that bias the interpretation of results (see Rutter & Silberg [10] and Purcell [8] for discussion of methodological issues in assessing GxE interactions in the presence of gene-environment correlation).

Conclusions

Recent behavioral genetic studies have taken a much-needed departure from standard studies of additive genetic and environmental influences on variation in human development by both acknowledging and testing for the interaction between genes and environments. Although these issues have a long history in behavioral genetic thinking [3, 6, 7], it is only more recently, with statistical, methodological, and molecular genetic advances, that these interesting research questions about gene-environment interplay have become tractable. The three methods reviewed rely on different means of assessing genetic influence. Specifically, adoption designs rely on the presence of high-risk phenotypes in biological parents of adoptees to assess genetic risk, traditional twin studies estimate changes in genetic influence as measured at the latent trait level, and studies of genotyped individuals focus on specific genetic variants. Nonetheless, all three methods require the inclusion of measured variables in the analyses, suggesting that behavioral genetic researchers should continue to include and refine their definitions of possible environmental influences on behavior, so that these more interesting and complex questions of gene x environment interaction can be explored.

References

- [1] Cadoret, R.J., Cain, C.A. & Crowe, R.R. (1983). Evidence for gene-environment interaction in the development of adolescent antisocial behavior, *Behavior Genetics* **13**, 301–310.
- [2] Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., Craig, I.W., Taylor, A., & Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children, *Science* **297**, 851–854.
- [3] Eaves, L. (1976). A model for sibling effects in man, *Heredity* **36**, 205–214.
- [4] Ge, X., Conger, R.D., Cadoret, R.J., Neiderhiser, J.M., Yates, W., Troughton, E., & Stewart, J.M. et al. (1996). The developmental interface between nature and nurture: a mutual influence model of child antisocial behavior and parent behaviors, *Developmental Psychology* **32**, 574–589.
- [5] Heath, A.C., Jardine, R. & Martin, N.G. (1989). Interactive effects of genotype and social environment on alcohol consumption in female twins, *Journal of Studies on Alcohol* **50**, 38–48.
- [6] Jinks, J.L. & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of human behavior, *Psychological Bulletin* **73**, 311–349.
- [7] Kendler, K.S. & Eaves, L.J. (1986). Models for the joint effect of genotype and environment on liability to psychiatric illness, *The American Journal of Psychiatry* **143**, 279–289.
- [8] Purcell, S. (2003). Variance components models for gene-environment interaction in twin analysis, *Twin Research* **5**, 554–571.
- [9] Rowe, D.C., Jacobson, K.C. & Van den Oord, E.J.C.G. (1999). Genetic and environmental influences on vocabulary IQ: parental education as moderator, *Child Development* **70**, 1151–1162.
- [10] Rutter, M. & Silberg, J. (2002). Gene-environment interplay in relation to emotional and behavioral disturbance, *Annual Review of Psychology* **53**, 463–490.
- [11] Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B.D. & Gottesman, I.I. (2003). Socioeconomic status modifies heritability of IQ in young children, *Psychological Science* **14**, 623–628.

KRISTEN C. JACOBSON

Generalizability

VANCE W. BERGER

Volume 2, pp. 702–704

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalizability

Due to various limitations in necessary resources, researchers generally study a phenomenon in a specific setting, and then try to generalize the obtained results to other settings. So, for example, while a target population may be infinite, the study sample will be a small (finite) subset of this target population. The degree to which the results of a study can be generalized from one setting to other settings is referred to as ‘generalizability’. Here, we will first present a set of commonly accepted rules for generalizing information, and then we will discuss the applications of these rules in behavioral sciences. We will then argue that the presented rules are subject to many exceptions, and that generalization requires abstraction, and depends on the human thought process, rather than being a mechanistic or statistical process [6].

There are some commonly accepted rules for generalizability of information. We will discuss three. First, a prerequisite for generalizability of the results is that the results of the study be valid within the study population (**internal validity**) [1, 2]. Consider, for example, a study comparing the efficacy of treatments *A* and *B*, in which the researchers allocate all, or even most, of the patients with a good prognosis to treatment *A*. Such a comparison is obviously invalid even internally, and so generalization of its results would be irrelevant, the possibility of a sampling bias that precisely compensates for this allocation bias [2] notwithstanding. To enhance the internal validity of studies, researchers often select a homogeneous group of humans or animals as their study population. Often, however, such a homogeneous population is not representative of the target population.

The extent to which a study population is homogeneous is usually a trade-off between the quest for internal validity and representativeness; hence, making a decision about the study population is usually a personal judgment of the researcher. Bench scientists and animal researchers tend to emphasize higher internal validity, while social and behavioral scientists and epidemiologists tend to prefer enhanced representativeness to enhanced internal validity.

Second, experience has shown that the more similar (among units in the target population) the contributors to the causal pathway are, the more reproducible the results will be. Physical phenomena are generally highly reproducible; for example, an

experiment conducted to measure the acceleration of a falling object due to gravity will likely obtain results that generalize to anywhere (on earth, at least). Likewise for an experiment performed to determine the freezing point of water, which will depend on pressure, but little else. If it is known that pressure matters, then it can be controlled for. Biologic phenomena tend to be more easily generalized from population to population than social or behavioral phenomena can be.

Third, in human studies, there is an important difference between generalizability of results of surveys and generalizability of results of association studies [6, 7]. In surveys, the target population is usually well defined. For the results of a survey to be generalizable, then, the study population must be representative of this well-defined target population. For this purpose, one usually selects a sufficiently large random sample, which confers confidence in the extent to which the sample is representative of the target population, with respect to both known and unknown factors. There are a few caveats worthy of mention. For one thing, the population from which the sample is randomly selected tends to differ from the target population. This forces one to consider not only the degree to which the sample represents the sampled population, but also the degree to which the sampled population represents the target population, and, ultimately, the degree to which the sample represents the target population.

Years ago, when telephones were not ubiquitous, telephone surveys tended to create biases in that they would overrepresent those wealthy enough to afford telephones. This would create a problem only if the target population included those without telephones. A famous example of this type of phenomenon occurred during the presidential race of 1936, when *The Literary Digest* mistakenly oversampled Republicans, and confidently predicted that Alf Landon would beat Franklin Roosevelt and win the presidency. Sometimes, the distortion between the sampled population and the target population is created intentionally, as when a run-in is used prior to a randomized trial for the purposes of excluding those subjects who do not respond well to the active treatment offered during the run-in [2]. This step can help create the illusion of a treatment effect.

Another consideration is the distinction between randomized and representative. It is true that a random sample may be used in hopes of creating

2 Generalizability

a representative sample, but if presented with a sample, one could check the extent to which this sample represents the target population (assuming that characteristics of this target population are also known). Having gone through the step of assessing the extent to which the sample is representative of the target population, would one then care if this sample were obtained at random? Certainly, other sampling schemes, such as convenience sampling, may create a sample that appears to represent the target population well, at least with respect to the dimensions of the sample that can be checked. For example, it may be feasible to examine the gender, age, and annual salary of the study subject for representativeness, but possibly not their political belief. It is conceivable that unmeasured factors contribute to results of survey questions, and ignoring them may lead to unexpected errors. From this perspective, then, randomization does confer added benefits beyond those readily checked and classified under the general heading of representativeness.

Of course, one issue that remains, even with a sample that has been obtained randomly, is a variant of the Heisenberg uncertainty principle [1]. Specifically, being in the study may alter the subjects in ways that cannot be measured, and the sample differs from the population at large with respect to a variable that may assume some importance. That is, if X is a variable that takes the value 1 for subjects in the sample, and the value 0 for subjects not in the sample, then the sample differs from the target population maximally with respect to X . Of course, prior to taking the sample, each subject in the target population had a value of $X = 0$, but for those subjects in the sample, the value of X was changed to 1, in time for X to exert whatever influence it may have on the primary outcomes of the study. This fact has implications for anyone not included in a survey.

If, for example, a given population (say male smokers over the age of 50 years) is said to have a certain level of risk regarding a given disease (say lung cancer), then what does this mean to a male smoker who was not included in the studies on which this finding was based? Hypothetically, suppose that this risk is 25%. Does this then mean that each male smoker over 50, whether in the sample or not, has a one in four chance of contracting lung cancer? Or does it mean that there is some unmeasured variable, possibly a genetic mutation, which we will call a predisposition towards lung cancer (for lack of a

better term), which a quarter of the male smokers over 50 happens to have? Presumably, there is no recognizable subset of this population, male smokers over 50, which would allow for greater separation (as in those who exercise a certain amount have 15% risk while those who do not have 35% risk).

Suppose, further, that one study finds a 20% risk in males and a 35% risk in smokers, but that no study had been done that cross-classified by both gender and smoking status. In such a case, what would the risk be for a male smoker? The most relevant study for any given individual is a study performed in that individual, but the resources are not generally spent towards such studies of size one. Even if they were, the sample size would still be far too small to study all variables that would be of interest, and so there is a trade-off between the specificity of a study (for a given individual or segment of the population) and the information content of a study.

In contrast to surveys, association studies require not only that the sample be representative of the study population but also that it be homogeneous. As mentioned previously, the use of run-ins, prior to randomization, to filter out poor responders to an active treatment creates a distortion that may result in a spurious association [2]. That is, there may well be an association, among this highly select group of randomized subjects who are already known to respond well to the active treatment, between treatment received and outcome, but this association may not reflect the reality of the situation in the population at large. But even if the sample is representative of the target population, there is still a risk of spurious association that arises from pooling heterogeneous segments of the population together. Suppose, for example, that one group tends to be older and to smoke more than another group, but that within either group there is no association between smoking status and age. Ignoring the group, and studying only age and smoking status, would lead to the mistaken impression that these two variables are positively associated. This is the ecological fallacy [5].

When trying to generalize associations in behavioral sciences, one needs to consider different characteristics of exposure, effect modifiers, confounders, and outcome. Duration, dose, route, and age at exposure may all be important. In general, extrapolating the results obtained from a certain range of exposure to values outside that range may be very misleading. While short-term low-dose stress may be

stimulating, very high levels of stress may inhibit productivity. Effect modifiers may vary among different populations. Single parenthood may be a stigma in some societies, and, therefore, may lead to behavioral abnormalities in the children. However, societies that show high support for single parent families may modify and lower such detrimental effects. Differences in the distribution of confounding factors result in failure of replication.

For example, higher education may be associated with higher income in some societies, but not in others. A clear definition of exposure and outcome is necessary, and these definitions should be maintained when generalizing the results. Sufficient variability in both exposure and outcome is also important. Family income may not be a predictor of future educational success when studied in a select group of families that all have an annual income between \$80 000–100 000, but it may be a strong predictor in a wider range of families.

Despite the fact that the term ‘generalizability’ is frequently used, and the rules mentioned above are commonly taught in methodology classes, the meaning of generalizability is often not clear, and these rules give us only a vague idea about how and when we are allowed to generalize information. One reason for such vagueness is that generalizability is a continuum rather than a dichotomous phenomenon, and the degree of acceptable similarity is not well defined. For example, suppose that in comparing treatments *A* and *B* for controlling high blood pressure, treatment *A* is more effective by 20 mmHg on average

in one population, but only 10 mmHg more effective on average in another population. Although treatment *A* is better than treatment *B* in both populations, the magnitude of the blood pressure reduction is different. Are the results obtained from one population generalizable to the other? There is no clear answer. Despite centuries of thinking and examination, the process of synthesis of knowledge from individual observations is not well understood [3, 4, 6]. This process is neither mechanical nor statistical; that is, the process requires abstraction [7].

References

- [1] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [2] Berger, V.W., Rezvani, A. & Makarewicz, V.A. (2003). Direct effect on validity of response run-in selection in clinical trials, *Controlled Clinical Trials* **24**, 156–166.
- [3] Chalmers, A.F. (1999). *What is this Thing Called Science?* 3rd Edition, Hackett publishing company, Indianapolis.
- [4] Feyerabend, P. (1993). *Against Method*, 3rd Edition, Verso Books, New York.
- [5] Piantidosi, S., Byar, D. & Green, S. (1988). The ecological fallacy, *American Journal of Epidemiology* **127**(5), 893–904.
- [6] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott–Raven, Philadelphia.
- [7] Szklo, M. (1998). Population-based cohort studies, *Epidemiologic Reviews* **20**, 81–90.

VANCE W. BERGER

Generalizability Theory: Basics

JOHN R. BOULET

Volume 2, pp. 704–711

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalizability Theory: Basics

From a practical perspective, scientists have always been interested in quantifying **measurement errors**. Regardless of the professional discipline, be it psychology, biology, sociology, medicine, and so on, gathering measurements and determining their precision is a fundamental task. A testing agency would certainly want to know the precision of any ability estimates, and which measurement conditions (e.g., test length, number of raters) best achieved the goal of obtaining valid and reliable examinee scores. For these studies and programs, and many others, there is often a fundamental need to ascertain the degree of measurement error that is involved in the process of gathering data. Once this is accomplished, one can then tailor the data-gathering process to minimize potential errors, thereby enhancing the precision of the measures.

Historically, classical test theory (CTT) (*see Classical Test Models*) [15] was employed as a framework for understanding and quantifying measurement errors. In CTT, a person's observed score (X) is made up of 'true' (T) component and error (E). The reliability of a measure (r_{xx}) is simply the ratio of true score variance (σ^2_{true}) to observed score variance ($\sigma^2_{\text{observed}}$). True scores and true score variances are typically estimated via methods such as internal consistency and test-retest. If there is no error in the measurement process, then the observed and 'true' measurement will be the same, reflecting perfect reliability. For most real-world measurement problems, this is unlikely to be the case. More important, an undifferentiated E is frequently of little value except to quantify the consistencies and inconsistencies in the data. From a practical perspective, it is often essential to know the exact conditions of measurement needed for acceptably precise measurement. To do this, one must be able to determine, quantify, and study multiple potential sources of error.

Generalizability (G) theory (*see Generalizability Theory: Overview*) offers a broad conceptual framework and associated statistical procedures for investigating various measurement issues. Unlike CTT, generalizability theory does not conceptualize measurement error as a unitary concept. Error can be attributed to multiple sources, and experiments can

be devised to estimate how much variation arises from each source. In effect, generalizability theory liberalizes, or extends, CTT. **Analysis of variance (ANOVA)** is used to disentangle multiple sources of error that contribute to the unitary E in CTT. As a result, generalizability analyses can be used to understand the relative importance of various sources of error and to define efficient measurement procedures. It should be noted, however, that although generalizability theory is rooted in ANOVA-type designs, there are key differences in emphasis (e.g., estimation of variance components as opposed to tests of statistical significance) and terminology.

While the conceptual framework for generalizability theory is relatively straightforward, there are some unique features that require explanation. These include universes of admissible observations, G (generalizability) studies, universes of generalization, D (decisions) studies, and universe scores. The universe of admissible observations for a particular measure is based on what the decision-maker is willing to treat as interchangeable for the purposes of making a decision. It is characterized by the sources of variation in universe scores (expected value of observations for the person in the stated universe) that are to be explicitly evaluated. For example, a researcher may be interested in evaluating the clinical skills of physicians. To do this, he/she could identify potential performance exercises (e.g., take a patient history, perform a required physical examination) and observers (physician, or expert, raters). For this hypothetical investigation, the universe of admissible observations contains an exercise, or task, facet (take a history, perform a physical examination) and a rater facet. If any of the tasks (t) could be paired with any of the raters (r), then the universe of admissible observations is said to be crossed (denoted $t \times r$). In generalizability theory, the term universe is reserved for conditions of measurement, and is simply the set of admissible observations to which the decision maker would like to generalize. The word population is used for objects of measurement. In the example noted above, the researcher would also need to specify the population (i.e., persons to be evaluated). The next step would be to collect and analyze data to estimate the relevant **variance components**. This is known as a G (generalizability) study. For example, one could envision a design where a sample of raters (n_r) evaluates the performances of a sample of persons (n_p) on a sample of clinical exercises, or

2 Generalizability Theory: Basics

tasks (n_t). This is a two-facet design and is denoted $p \times t \times r$ (person by task by rater). Where each level of one facet (rater) is observed in combination with each level of the other (task), the result is a crossed design. If levels of one facet are observed in combination with specific level(s) of another, the design is said to be nested. For example, variance components can be estimated for people, tasks, raters, and the associated interactions. These components are simply estimates of differences in scores attributable to a given facet or interaction of sources.

The purpose of a G study is to obtain estimates of variance components associated with the universe of admissible observations. These estimates can be used in D (decision) studies to design efficient measurement procedures. For D studies, the researcher must specify a universe of generalization. This could contain all facets in the universe of admissible observations (e.g., $p \times T \times R$; for D study designs, facets are typically denoted by capital letters) or be otherwise restricted. For the scenario above, one may want to generalize persons' scores based on the specific tasks and raters used in the G study to persons' scores for a universe of generalization that involves many other tasks and raters. The sample sizes in the D study (n'_t, n'_r) need not be the same as the sample sizes in the G study (n_t, n_r). Also, the focus of the D study is on mean scores for persons rather than single person by task by rater observations. If a person's score is based on his or her mean score over $n'_t n'_r$ observations, the researcher can explore, through various D studies, the specific conditions that can make the measurement process more efficient.

It is conceivable to obtain a person's mean score for every instance of the measurement procedure (e.g., tasks, raters) in the universe of generalization. The expected value of these mean scores in the stated universe is the person's universe score. The variance of universe scores over all persons in the population is called the *universe score variance*. More simply, it is the sum of all variance components that contribute to differences in observed scores. Universe score variance is conceptually similar to true score variance (T) in classical test theory.

As mentioned previously, once the G study variance components are estimated, various D studies can be completed to determine the optimal conditions for measurement. Unlike CTT, where observed score variance can only be partitioned into two parts (σ^2_{true} and $\sigma^2_{\text{observed}}$), generalizability theory affords

the opportunity to further partition error variance. More important, since some error sources are only critical with respect to relative decisions (e.g., rank ordering people based on scores), and others can influence absolute decisions (e.g., determining mastery based on defined standards or cutoffs), it is essential that they can be identified and disentangled. Once this is accomplished, both error 'main effects' and error 'interaction effects' can be studied. For example, in figure-skating, multiple raters are typically used to judge the performance of skaters across multiple programs (short, long). Measurement error can be introduced as a function of the choice of rater, the type of program (task), and, most important, the interaction between the two. For this situation, if we accept that any person in the population can participate in any program in the universe and can be evaluated by any rater in the universe, the observable score for a single program evaluated by a single rater can be represented:

$$X_{\text{ptr}} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{\text{ptr}}. \quad (1)$$

For this design, μ is the grand mean in the population and universe and ν specifies any one of the seven components. From this, the total observed score variance can be decomposed into seven independent variance components:

$$\begin{aligned} \sigma^2(X_{\text{ptr}}) = & \sigma^2_{(p)} + \sigma^2_{(t)} + \sigma^2_{(r)} + \sigma^2_{(pt)} \\ & + \sigma^2_{(pr)} + \sigma^2_{(tr)} + \sigma^2_{(\text{ptr})} \end{aligned} \quad (2)$$

The variance components depicted above are for single person by programs (tasks) by rater combinations. From a CTT perspective, one could collapse scores over the raters and estimate the consistency of person scores between the long and short programs. Likewise, one could collapse scores over the two programs and estimate error attributable to the raters. While these analyses could prove useful, only generalizability theory evaluates the interaction effects that introduce additional sources of measurement error.

In generalizability theory, reliability-like coefficients can be computed both for situations where scores are to be used for relative decisions and for conditions where absolute decisions are warranted. For both cases, relevant measurement error variances (determined by the type of decision, relative or absolute) are pooled. The systematic variance (universe score variance) is then divided by the sum of the

systematic and the measurement error variance to estimate reliability. When relative decisions are being considered, only measurement error variances that could affect the rank orderings of the scores are important. For this use of scores, the ratio of systematic variance to the total variance, known as a generalizability coefficient ($E\rho^2$), is the reliability estimate. This is simply a quantification of how well persons' observed scores correspond to the universe scores. When absolute decisions are considered (i.e., where scores are interpreted in relation to a standard or cutoff), all the measurement error variances can impact the reliability of the scores. Here, the reliability coefficient, Phi (Φ), is also calculated as the ratio of systematic variance to total variance. If all the measurement error variances that are uniquely associated with absolute decisions are zero, then the generalizability and Phi coefficients will be equal.

The defining treatment of generalizability theory is provided by Cronbach et al. [10]. Brennan provides a history of the theory [2]. For the interested reader, there are numerous books and articles, both technical and nontechnical, covering all aspects of the theory [17].

Purpose

The purpose of this entry is to familiarize the reader with the basic concepts of generalizability theory. For the most part, the treatment is nontechnical and concentrates on the utility of the theory and associated methodology for handling an assortment of measurement problems. In addition, only univariate models are considered. For more information on specific estimation procedures, multivariate specifications, **confidence intervals** for estimated variance components, and so on, the reader should consult Brennan [3]. For this entry, the basic concepts of generalizability theory are illustrated through the analysis of assessment data taken from an examination developed to evaluate the critical-care skills of physicians [1].

Measurement Example

Fully Crossed Design

The data for this example came from a performance-based assessment, designed to evaluate the emergency

care skills of physicians training in anesthesiology. The assessment utilized a sensorized, life-size electromechanical patient mannequin that featured, amongst other things, breath sounds, heart sounds, and pulses. Computer-driven physiologic and pharmacologic models determine cardiac and respiratory responses, and are used to simulate acute medical conditions. The simulator offers simple as well as advanced programming actions to create and then save a unique scenario for repeated evaluation of performances. A variety of additional features (e.g., heart rate, lung compliance, vascular resistance) can be manipulated independently to create a unique, but reproducible event that effectively tests the skill level of the medical provider. Six acute care scenarios (cases) were developed. Each simulated scenario was constructed to model a medical care situation that required a rapid diagnosis and acute intervention in a brief period of time.

Twenty-eight trainees were recruited and evaluated. Each of the 28 participants was assessed in each of the six simulated scenarios. Each trainee's performance was videotaped and recorded. Three raters independently observed and scored each of the performances from the videotaped recordings. A global score, based on the time to diagnosis and treatment as well as potentially egregious or unnecessary diagnostic or therapeutic actions, was obtained. The raters were instructed to make a mark on a 10-cm horizontal line based on their assessment of the trainee's performance. The global rating system was anchored by the lowest value 0 (unsatisfactory) and the highest value 10 (outstanding).

Analysis. From a generalizability standpoint, the G study described above was fully crossed ($p \times t \times r$). All of the raters ($n_r = 3$) provided a score for each of the six ($n_t = 6$) scenarios (referred to as tasks) across all 28 trainees (objects of measurement). The person by rater by task design can be used to investigate the sources of measurement error in the simulation scores. Here, it was expected that the principle source of variance in scores would be associated with differences in individual resident's abilities, not choice of task or choice rater.

Generalizability (G) Study. The analysis of G study, including the provision of estimated variance components, could be done by hand. There are, however, a number of available software packages that

4 Generalizability Theory: Basics

Table 1 Estimated variance components, standard errors of measurements, generalizability, and dependability coefficients for simulation scores (G and D studies)

G Study		D studies – mean variance component			
Component	Estimate ^a	Component	t = 6, r = 3	t = 6, r = 2	t = 8, r = 2
Person ($\hat{\sigma}_p^2$)	1.28	Person ($\hat{\sigma}_p^2$)	1.28	1.28	1.28
Task ($\hat{\sigma}_t^2$)	0.51	Task ($\hat{\sigma}_t^2$)	0.09	0.09	0.06
Rater ($\hat{\sigma}_r^2$)	0.24	Rater ($\hat{\sigma}_r^2$)	0.08	0.12	0.12
$\hat{\sigma}_{pt}^2$	2.09	$\hat{\sigma}_{pT}^2$	0.35	0.35	0.26
$\hat{\sigma}_{pr}^2$	0.30	$\hat{\sigma}_{pR}^2$	0.10	0.15	0.15
$\hat{\sigma}_{tr}^2$	0.11	$\hat{\sigma}_{TR}^2$	0.01	0.01	0.01
$\hat{\sigma}_{ptr}^2$	1.07	$\hat{\sigma}_{pTR}^2$	0.06	0.09	0.07
		$\hat{\sigma}^2(\Delta)$	0.69	0.81	0.67
		$\hat{\sigma}(\Delta)$	0.83	0.90	0.82
		$\hat{\sigma}^2(\delta)$	0.51	0.59	0.48
		$\hat{\sigma}(\delta)$	0.71	0.77	0.69
		Φ	0.65	0.61	0.66
		$E\rho^2$	0.72	0.68	0.73

^aestimate for single person by task by rater combinations.

make this task much less cumbersome [4, 5, 9]. In addition, commonly used statistical programs typically have routines for estimating variance components for a multitude of G study designs. For this example, the SAS PROC VARCOMP routine was used [16].

The estimated variance components for the G study are presented in Table 1. The person (trainee) variance component ($\hat{\sigma}^2(p)$) is an estimate of the variance across trainees of trainee-level mean scores. If one could obtain the person's expected score over all tasks and raters in the universe of admissible observations, the variance of these scores would be $\hat{\sigma}^2(p)$. Ideally, most of the variance should be here, indicating that individual abilities account for differences in observed scores. The other "main effect" variance components include task ($\hat{\sigma}^2(t)$) and rater ($\hat{\sigma}^2(r)$). The task component is the estimated variance of scenario mean scores. Since the estimate is greater than zero, we know that the six tasks vary somewhat in average difficulty. Not surprisingly, mean performance, by simulation scenario, ranged from a low of 5.7 to a high of 8.2. The rater component is the variance of the rater mean scores. The nonzero value indicates that raters vary somewhat in terms of average stringency. The mean rater scores, on a scale from 0 to 10, were 7.4, 6.3, and 7.0, respectively. Interestingly, the task variance component is approximately twice as large as the rater component. We can, therefore, conclude that raters differ much less

in average stringency than simulation scenarios differ in average difficulty.

The largest interaction variance component was person by task ($\hat{\sigma}^2(pt)$). The magnitude of this component suggests that there are considerably different rank orderings of examinee mean scores for each of the various simulation scenarios. The relatively small person by rater component suggests that the various raters rank order persons similarly. Likewise, the small rater by task component indicates that the raters rank order the difficulty of the simulation scenarios similarly. The final variance component is the residual variance that includes the triple-order interactions and all other unexplained sources of variation.

Decision (D) Studies. The G study noted above was used to derive estimates of variance components associated with a universe of admissible observations. Decision (D) studies can use these estimates to design efficient measurement procedures for future operations. To do this, one must specify universe of generalization. For the simulation assessment, we may want to generalize trainees' scores based on the six tasks and three raters to trainees' scores for a universe of generalization that includes many other tasks and many other raters. In this instance, the universe of generalization is "infinite" in that we wish to generalize to any set of raters and any set of tasks. Here, consistent with ANOVA terminology, both the rater and task facets are said to be random as opposed to fixed.

For the simulation assessment, we may decide that we want each trainee to be assessed in each of the six encounters (tasks; $n'_t = 6$) and each of the tasks be rated by three independent raters ($n'_r = 3$). Although the sample sizes for the D study are the same as those for the G study, this need not be the case. Unlike the G study, which focused on single trainee by task by rater observations, the D study focuses on mean scores for persons.

The D study variance components can be easily estimated using the G study variance components in Table 1 (see Table 1). The estimated random effects variance components are for person mean scores over $n'_t = 6$ tasks and $n'_r = 3$ raters. The calculation of the variance components for this fully crossed design is relatively straightforward. The estimated universe score variance ($\hat{\sigma}_p^2$) stays the same. To obtain means, variance components that contain t but not r are divided by n'_t . Components that contain r but not t are divided by n'_r . And components that contain both t and r are divided by $n'_t n'_r$.

Since an infinite universe of generalization has been defined, all variance components other than $\hat{\sigma}^2(p)$ contribute to one or more types of error variance. If the trainee scores are going to be used for mastery decisions (e.g., pass/fail), then all sources of error are important. Here, both simulation scenario difficulty and rater stringency are potential sources of error in estimating "true" ability. Absolute error is simply the difference between a trainee's observed and universe score. Variance of the absolute errors $\sigma^2(\Delta)$ is the sum of all variance components except $\sigma^2(p)$ (see Table 1). The square root of this value $\hat{\sigma}(\Delta)$ is interpretable as the absolute standard error of measurement (SEM). On the basis of the D study described above, $\hat{\sigma}(\Delta) = 0.83$. As a result, $X_{pTR} \pm 1.62$ constitutes an approximate 95% confidence interval for trainees' universe scores.

If the purpose of the simulation assessment is simply to rank order the trainees, then some component variances will not contribute to error. For these measurement situations, relative error variance $\hat{\sigma}^2(\delta)$ similar to error variance in CTT is central. For the $p \times T \times R$ D study with $n'_t = 6$ and $n'_r = 3$ relative error variance is the sum of all variance components, excluding $\sigma^2(p)$, that contain p (i.e., $\hat{\sigma}_{pT}^2, \hat{\sigma}_{pR}^2, \hat{\sigma}_{pTR}^2$). These are the only sources of variance that can impact the relative ordering of trainee scores. The calculated value ($\hat{\sigma}^2(\delta) = 0.51$) is necessarily lower than $\hat{\sigma}^2(\Delta)$, in that fewer variance components are

considered. The square root of the relative error variance ($\hat{\sigma}(\delta) = 0.71$) is interpretable as an estimate of the relative SEM. As would be expected, and borne out by the data, absolute interpretations of a trainee's score are more error-prone than relative ones.

In addition to calculating error variances, two types of reliability-like coefficients are widely used in generalizability theory. The generalizability coefficient ($E\rho^2$), analogous to a reliability coefficient in CTT, is the ratio of universe score variance to itself plus error variance:

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}. \quad (3)$$

For $n'_t = 6$ and $n'_r = 3$, $E\rho^2 = 0.72$. An index of dependability (Φ) can also be calculated:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}. \quad (4)$$

This is the ratio of universe score variance to itself plus absolute score variance. For $n'_t = 6$ and $n'_r = 3$, $\Phi = 0.65$. The dependability coefficient is apropos when absolute decisions about scores are being made. For example, if the simulation scores, in conjunction with a defined standard, were going to be used for licensure or certification decisions, then all potential error sources are important, including those associated with variability in task difficulty and rater stringency.

The $p \times T \times R$ design with two random facets (tasks, $n'_t = 6$; raters, $n'_r = 3$) was used for illustrative purposes. However, based on the relative magnitudes of the G study variance components, it is clear that the reliability of the simulation scores is generally more dependent on the number of simulation scenarios as opposed to the number of raters. One could easily model a different D study design and calculate mean variance components for $n'_t = 6$ and $n'_r = 2$ (see Table 1). By keeping the same number of simulated encounters, and decreasing the number of raters per case ($n'_r = 2$), the overall generalizability and dependability coefficients are only slightly lower. Increasing the number of tasks ($n'_t = 8$) while decreasing the number of raters per task ($n'_r = 2$) has the effect of lowering both absolute and relative error variance. Ignoring the specific costs associated with developing simulation exercises, testing trainees, and rating performances, increasing the number of tasks, as opposed to raters per given task, would appear to

be the most efficient means of enhancing the precision of examinee scores.

Fixed Facets. The D studies described above ($p \times T \times R$) involved two random facets and a universe of generalization that was infinite. Here, we were attempting to generalize to any other set of simulation exercises and any other group of raters. This does not, however, always have to be the case. If we were only interested in generalizing to the six simulation scenarios that were initially modeled, but some set of randomly selected raters, then a mixed model results. The task facet is said to be fixed, as opposed to random, and the universe of generalization is thereby restricted. In essence, we are considering the six simulation scenarios to be the universe of all simulation scenarios. On the basis of the fully crossed $p \times t \times r$ G study design, the variance components for a design with a fixed T can be calculated quite easily. For D study sample sizes ($n'_t = 6$ and $n'_r = 3$), with T fixed, the generalizability coefficient $E\rho^2$ is estimated to be 0.91. Although this value is much larger than that estimated for a design with a random task facet ($E\rho^2 = 0.72$), one cannot generalize to situations in which other simulation exercises are used.

Nested Designs. The G study employed a fully crossed $p \times t \times r$ design with an infinite universe of admissible observations. Here, all tasks (simulation scenarios) were evaluated by all raters. Given that the cost of physician raters is high and the G study variance components associated with the rater were comparatively low, one could also envision a situation where there were multiple tasks but only a single (different) rater for each performance. This describes a nested design ($p \times (R:T)$), where R:T denotes rater nested in task. For a design with $n'_t = 6$ and $n'_r = 1$, the random effects D study variance components, including $\hat{\sigma}^2(R:T)$ and $\hat{\sigma}^2(pR:T)$, can be calculated from the appropriate G study variance components. For the simulation study, the estimated values of Φ and $E\rho^2$ would be 0.64 and 0.69, respectively. Interestingly, these values are only slightly lower than those for a random model with $n'_t = 6$ and $n'_r = 3$ raters (per task). From a measurement perspective, this suggests that a design involving multiple tasks and a single rater per task would be comparatively efficient.

Conclusion

From a descriptive point of view, generalizability theory involves the application of ANOVA techniques to measurement procedures. Its major contribution is that it permits the decision-maker to pinpoint sources of measurement error and change the appropriate number of observations accordingly in order to obtain a certain level of generalizability. Unlike CTT, which considers a single source of error (E), generalizability theory allows for multiple sources of error, permits direct comparison of these error sources, allows for different 'true' (universe) scores, and provides an analysis framework for determining optimal measurement conditions to attain desired precision.

Generalizability theory has been applied to many real-world measurement problems, including standard setting and equating exercises [11, 13], computerized scoring applications [8], and the design of various performance-based assessments [7]. There have also been numerous advances in generalizability theory, including work related to model fit and estimation methods [6], sampling issues [14], and applications of multivariate specifications [12].

References

- [1] Boulet, J.R., Murray, D., Kras, J., Woodhouse, J., McAllister, J. & Ziv, A. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents, *Anesthesiology* **99**(6), 1270–1280.
- [2] Brennan, R.L. (1997). A perspective on the history of generalizability theory, *Educational Measurement: Issues and Practice* **16**(4), 14–20.
- [3] Brennan, R.L. (2001a). *Generalizability Theory*, Springer-Verlag, New York.
- [4] Brennan, R.L. (2001b). *Manual for urGENOVA (Version 2.1)*, Iowa Testing Programs, University of Iowa, Iowa City.
- [5] Brennan, R.L. (2001c). *Manual for mGENOVA (Version 2.1)*, Iowa Testing Programs, University of Iowa, Iowa City.
- [6] Brennan, R.L. & Gao, X. (2001). Variability of estimated variance components and related statistics in a performance assessment, *Applied Measurement in Education* **14**(2), 191–203.
- [7] Brennan, R.L. & Johnson, E.G. (1995). Generalizability of performance assessments, *Educational Measurement: Issues and Practice* **14**, 9–12.
- [8] Clauser, B., Harik, P. & Clyman, S. (2000). The generalizability of scores for a performance assessment scored

-
- with a computer-automated scoring system, *Journal of Educational Measurement* **37**(3), 245–262.
- [9] Crick, J.E. & Brennan, R.L. (1983). *Manual for GENOVA: A Generalized Analysis of Variance System*, American College Testing, Iowa City.
- [10] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, John Wiley, New York.
- [11] Fitzpatrick, A.R. & Lee, G. (2003). The effects of a student sampling plan on estimates of the standard errors for student passing rates, *Journal of Educational Measurement* **40**(1), 17–28.
- [12] Hartman, B.W., Fuqua, D.R. & Jenkins, S.J. (1988). Multivariate generalizability analysis of three measures of career indecision, *Educational and Psychological Measurement* **48**, 61–68.
- [13] Hurtz, N.R. & Hurtz, G.M. (1999). How many raters should be used for establishing cutoff scores with the Angoff method: a generalizability theory study, *Educational and Psychological Measurement* **59**(6), 885–897.
- [14] Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling, *Journal of Educational Measurement* **39**(2), 165–181.
- [15] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [16] SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6*, Vol. 2, 4th Edition, SAS Institute Inc., Cary.
- [17] Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: a Primer*, Sage, Newbury Park.

JOHN R. BOULET

Generalizability Theory: Estimation

PIET F. SANDERS

Volume 2, pp. 711–717

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalizability Theory: Estimation

Introduction

Generalizability theory (G theory) is a conceptual and statistical framework for the analysis and construction of measurement instruments. The first extensive treatment of G theory was presented by Cronbach, Gleser, Nanda, and Rajaratnam [3] and the most recent one by Brennan [2]. Less extensive introductions to G theory can be found in publications by Shavelson and Webb [6] and Brennan [1].

In most testing situations, it is not the particular sample of items or the particular sample of raters involved in an evaluation of a person's performance that is of interest. Different samples of items and different samples of raters would be equally acceptable or admissible. All the admissible items and raters together constitute the universe of admissible observations, which we would ideally like to administer to determine a person's universe score. This being unfeasible, we have to work with the person's performance on a particular sample of items evaluated by a particular sample of raters, that is, the person's observed score. The question is to what extent can we generalize from observed score to universe score. In G theory, the answer to this generalizability question is quantified by generalizability coefficients.

Conceptual Framework of G Theory

In G theory, behavioral measurements or observations are described in terms of conditions; a set of similar conditions is referred to as a *facet*. An achievement test with 40 multiple-choice items is said to have 40 conditions of the facet 'items'; a performance test with 10 tasks evaluated by two raters has 10 conditions of the facet 'tasks' and 2 conditions of the facet 'raters.' The objects being measured, usually persons, are not regarded as a facet.

Responses are obtained by means of designs where objects of measurement have to respond to the conditions of one or more facets. Designs differ not only in the number of facets (one or more) and the nature of the facets (random or fixed) but also

in how the conditions of the facet are administered (crossed or nested). The term-crossed design means that the persons have to respond to all the conditions of all the facets; with a nested design, they have to respond to only a selection of the conditions of the facets.

One-facet Design

In a one-facet crossed design, the observed score for one person on one item, X_{pi} , can be decomposed as

$$\begin{aligned} X_{pi} &= \mu && \text{(grand mean)} \\ &+ \mu_p - \mu && \text{(person effect)} \\ &+ \mu_i - \mu && \text{(item effect)} \\ &+ X_{pi} - \mu_p - \mu_i + \mu && \text{(residual effect)} \end{aligned} \quad (1)$$

The model in (1) has three parameters. The first parameter, the grand mean or the mean over the population of persons and the universe of items, is defined as $\mu \equiv \varepsilon_p \varepsilon_i X_{pi}$. The second parameter, the universe score of a person or the mean score over the universe of items, is defined as $\mu_p \equiv \varepsilon_i X_{pi}$. The third parameter, the population mean of an item or the mean score over the population of persons, is defined as $\mu_i \equiv \varepsilon_p X_{pi}$. Except for the grand mean, the three effects in (1) have a distribution with a mean of zero and a positive variance. For example, the mean of the person effect is $\varepsilon_p(\mu_p - \mu) = \varepsilon_p(\mu_p) - \varepsilon_p(\mu) = \mu - \mu = 0$. Each effect or component has its own variance component. The **variance components** for persons, items, and the residual or error are defined as $\sigma_p^2 = \varepsilon_p(\mu_p - \mu)^2$, $\sigma_i^2 = \varepsilon_i(\mu_i - \mu)^2$, and $\sigma_{pi,e}^2 = \varepsilon_p \varepsilon_i (X_{pi} - \mu_p - \mu_i + \mu)^2$. The variance of the observed scores is defined as $\sigma_X^2 = \sigma^2(X_{pi}) = \varepsilon_p \sigma_i^2 (X_{pi} - \mu)^2$. It can be shown that σ_X^2 , the total variance, is equal to the sum of the three variance components: $\sigma_X^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2$.

Generalizability and Decision Study

In generalizability theory, a distinction is made between a generalizability study (G study) and a decision study (D study). In a G study, the variance components are estimated using procedures from analysis of variance. In a D study, these variance components are used to make decisions on, for example, how many items should be included in the test or how

2 Generalizability Theory: Estimation

many raters should evaluate the responses. A G study of a one-facet crossed random effects design is presented below. A D study of this design is discussed in the next section.

Generalizability Study One-facet Design

The **Analysis of Variance** table of a crossed one-facet random effects design, a design where a random sample of n_p persons from a population of persons responds to a random sample of n_i items from a universe of items, is presented in Table 1.

From Table 1, we can see that we first have to compute the sums of squares in order to estimate the variance components. For that, we substitute the three parameters μ , μ_p , and μ_i in (1) with their observed counterparts, which results in the following decomposition:

$$\begin{aligned} X_{pi} &= \bar{X} + (\bar{X}_p - \bar{X}) + (\bar{X}_i - \bar{X}) \\ &\quad + (X_{pi} - \bar{X}_p - \bar{X}_i + \bar{X}) \\ &= X_{pi} - \bar{X} = (\bar{X}_p - \bar{X}) + (\bar{X}_i - \bar{X}) \\ &\quad + (X_{pi} - \bar{X}_p - \bar{X}_i + \bar{X}) \end{aligned} \quad (2)$$

By squaring and summing the observed deviation scores in (2), four sums of squares are obtained: the total sums of squares and the sums

of squares of persons, items, and interactions. The total sums of squares, $\sum_p \sum_i (X_{pi} - \bar{X})^2$, is equal to the sum of the three other sums of squares: $\sum_p \sum_i (X_p - \bar{X})^2 + \sum_p \sum_i (X_i - \bar{X})^2 + \sum_p \sum_i (X_{pi} - \bar{X}_p - \bar{X}_i + \bar{X})^2$. The former is also written as $SS_{\text{tot}} = SS_p + SS_i + SS_{pi,e}$. The mean squares can be computed from the sums of squares. Solving the equations of the expected mean squares for the variance components and replacing the observed mean squares by their expected values results in the following estimators for the variance components: $\hat{\sigma}_p^2 = (MS_p - MS_{pi,e})/n_i$, $\hat{\sigma}_i^2 = (MS_i - MS_{pi,e})/n_p$, and $MS_{pi,e} = \hat{\sigma}_{pi,e}^2$.

The artificial example in Table 2 was used to obtain the G study results presented in Table 3.

Table 2 contains the scores (0 or 1) of four persons on three items, the mean scores of the persons, the mean scores of the items, and the general mean. The mean scores of the persons vary between a perfect mean score of 1 and a mean score of 0. The mean scores of the items range from an easy item of 0.75 to a difficult item of 0.25.

The last column of Table 3 contains the estimated variance components which are variance components of scores of single persons on single items. Since the size of the components depends on the score scale of the items, the absolute size of the variance components does not yield very useful information.

Table 1 Analysis of variance table of a crossed one-facet random effects design

Effects	Sums of squares	Degrees of freedom	Mean squares	Expected mean squares
Persons (p)	SS_p	$df_p = n_p - 1$	$MS_p = SS_p/df_p$	$\Sigma(MS_p) = \sigma_{pi,e}^2 + n_i \sigma_p^2$
Items (i)	SS_i	$df_i = n_i - 1$	$MS_i = SS_i/df_i$	$\Sigma(MS_i) = \sigma_{pi,e}^2 + n_p \sigma_i^2$
Residual (pi, e)	$SS_{pi,e}$	$df_{pi,e} = (n_p - 1) \times (n_i - 1)$	$MS_{pi,e} = SS_{pi,e}/df_{pi,e}$	$\Sigma(MS_{pi,e}) = \sigma_{pi,e}^2$

Table 2 The item scores of four persons on three items, the mean score per person and per item, and the general mean

Person	Item			$\bar{X}_p$
	1	2	3	
1	1	1	1	1.000
2	1	1	0	0.667
3	1	0	0	0.333
4	0	0	0	0.000
$\bar{X}_i$	0.75	0.50	0.25	0.500 = $\bar{X}$

Table 3 Results generalizability study for the example from Table 2

Effects	Sums of squares	Degrees of freedom	Mean squares	Estimates of variance components
Persons (p)	1.667	3	0.556	$\hat{\sigma}_p^2 = 0.139$ (45.5%)
Items (i)	0.500	2	0.250	$\hat{\sigma}_i^2 = 0.028$ (9%)
Residual (pi, e)	0.833	6	0.139	$\hat{\sigma}_{pi,e}^2 = 0.139$ (45.5%)

It is therefore common practice to report the size of the component as a percentage of the total variance. Since the items are scored on a 0 to 1 score scale, the variance component cannot be larger than 0.25. The reason for the large universe score variance is the large differences between the mean scores of the four persons. The estimated variance component for the items is relatively small. This can be confirmed by taking the square root of the variance components, resulting in a standard deviation of 0.17, which is approximately one-sixth of the range for items scored on a dichotomous score scale. This value is what we might expect under a normal distribution of the scores.

Decision Study One-facet Design

The model in (1) and its associated variance components relate to scores of single persons on single items from the universe of admissible observations. However, the evaluation of a person's performance is never based on the score obtained on a single item, but on a test with a number of items. What the effect is of increasing the number of items on the variance components was investigated in a D study.

The linear model for the decomposition of the mean score of a person on a test with n_i items, denoted by X_{pI} , is

$$X_{pI} = \mu + (\mu_p - \mu) + (\mu_I - \mu) + (X_{pI} - \mu_p - \mu_I + \mu). \quad (3)$$

In (3), the symbol I is used to indicate the mean score on a number of items. In (3), the universe score is defined as $\mu_p \equiv \Sigma_I X_{pI}$, the expected value of X_{pI} over random parallel tests. By taking the expectation over I in (3), the universe score variance σ_p^2 does not change; the two other variance components do change and are defined as $\sigma_I^2 = \sigma_i^2/n_i'$ and $\sigma_{pi,e}^2 = \sigma_{pi,e}^2/n_i'$. The total variance, $\sigma_X^2 = \sigma^2(X_{pI})$, is equal to $\sigma_X^2 = \sigma_p^2 + \sigma_I^2 + \sigma_{pi,e}^2$.

Table 4 contains the variance components from the G study and the D study with three items.

The results in Table 4 show how the variance component of the items and the variance component of the interaction or error component change if we increase the number of items. To gauge the effect of using three more items from the universe of admissible observations, we have to divide the appropriate G-study variance components by 6.

The purpose of many tests is to determine the position of a person in relation to other persons. In generalizability theory, the relative position of a person is called the *relative universe score* and defined as $\mu_p - \mu$. The relative universe score is estimated by $X_{pI} - X_{PI}$, the difference between the mean test score of a person and the mean test score of the sample of persons. The deviation between $X_{pI} - X_{PI}$ and $\mu_p - \mu$ is called *relative error* and is defined as $\delta_{pI} = (X_{pI} - X_{PI}) - (\mu_p - \mu)$. The estimated relative error variance is equal to $\hat{\sigma}_\delta^2 = \hat{\sigma}_{pi,e}^2$. (Note that the prime is used to indicate the sample sizes in a D study.) For the crossed one-facet random effects design, the estimate of the

Table 4 Results of G study and D study for the example from Table 2

Effects	Variance components G study	Variance components D study
Persons (p)	$\hat{\sigma}_p^2 = 0.139$	$\hat{\sigma}_p^2 = 0.139$
Items (i)	$\hat{\sigma}_i^2 = 0.028$	$\hat{\sigma}_i^2 = 0.028/3 = 0.009$
Residual (pi, e)	$\hat{\sigma}_{pi,e}^2 = 0.139$	$\hat{\sigma}_{pi,e}^2 = 0.139/3 = 0.046$

4 Generalizability Theory: Estimation

generalizability coefficient for relative decisions, $\hat{\rho}^2$, is defined as

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\delta^2}.$$

Table 4 shows that the universe score variance for the test with three items is equal to 0.139 and the relative error variance is equal to 0.046. The generalizability coefficient, which has a lower limit of 0 and an upper limit of 1, is equal to 0.75. There are two possible interpretations of this coefficient. The first is that the coefficient is approximately equal to the expected value of the squared correlation between observed and universe score. The second is that the coefficient is approximately equal to the correlation between pairs of two randomly parallel tests.

It can be shown [4] that $\hat{\rho}^2$ is equal to the reliability coefficient KR-20 for dichotomous item scores and equal to Cronbach's coefficient alpha for polytomous scores. In addition to the reliability and generalizability coefficient, the standard error of measurement is also used as an indicator for the reliability of measurement instruments. The relative standard error of measurement is obtained by taking the square root of the relative error variance and can be shown to be equal to the standard error of measurement from classical test theory.

The purpose of a measurement instrument can also be to determine a person's absolute universe score. For example, if we want to know that a person can give a correct answer to at least 80% of the test items. The absolute universe score, μ_p , is estimated by X_{pI} . The deviation between X_{pI} and μ_p is called *absolute error* and is defined as $\Delta_{pI} = X_{pI} - \mu_p = (\mu_I - \mu) + (X_{pI} - \mu_p - \mu_I + \mu)$. The estimated absolute error variance is equal to $\hat{\sigma}_\Delta^2 = \hat{\sigma}_I^2 + \hat{\sigma}_{pI,e}^2$. For a crossed one-facet random effects design, the estimate of the generalizability coefficient, $\hat{\phi}$, for absolute decisions is defined as

$$\hat{\phi} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\Delta^2}.$$

For the example in Table 2 with three items, the generalizability coefficient for absolute decisions is equal to 0.72.

Two-facet Design

In a crossed two-facet design, the observed score of person p on item i awarded by rater r , X_{pir} , can be

decomposed into seven components:

$$\begin{aligned} X_{pir} &= \mu && \text{(grand mean)} \\ + \mu_p - \mu &&& \text{(person effect)} \\ + \mu_i - \mu &&& \text{(item effect)} \\ + \mu_r - \mu &&& \text{(rater effect)} \\ + \mu_{pi} - \mu_p - \mu_i + \mu &&& \text{(person} \times \text{item effect)} \\ + \mu_{pr} - \mu_p - \mu_r + \mu &&& \text{(person} \times \text{rater effect)} \\ + \mu_{ir} - \mu_i - \mu_r + \mu &&& \text{(item} \times \text{rater effect)} \\ + X_{pir} - \mu_{pi} - \mu_{pr} &&& \\ - \mu_{ir} + \mu_p + \mu_i &&& \\ + \mu_r - \mu &&& \text{(residual effect)} \end{aligned} \quad (4)$$

The seven parameters in (4) are defined as $\mu \equiv \varepsilon_p \varepsilon_i \varepsilon_r X_{pir}$, $\mu_p \equiv \varepsilon_i \varepsilon_r X_{pir}$, $\mu_i \equiv \varepsilon_p \varepsilon_r X_{pir}$, $\mu_r \equiv \varepsilon_p \varepsilon_i X_{pir}$, $\mu_{pi} \equiv \varepsilon_r X_{pir}$, $\mu_{pr} \equiv \varepsilon_i X_{pir}$, and $\mu_{ir} \equiv \varepsilon_p X_{pir}$.

A crossed two-facet design has a total of seven variance components. The total variance is equal to $\sigma_X^2 = \sigma_p^2 + \sigma_i^2 + \sigma_r^2 + \sigma_{pi}^2 + \sigma_{pr}^2 + \sigma_{ir}^2 + \sigma_{pir,e}^2$. Estimates of variance components can be obtained using procedures comparable to the one described for the crossed one-facet design.

Generalizability Study Two-facet Design

Table 5 contains an example, taken from Thorndike [7], of a crossed two-facet design where two raters have awarded a score to the answers on four items of six persons.

The results of the generalizability study for the example in Table 5 are presented in Table 6.

The last column in Table 6 contains the estimates of the variance components and the contribution of each component to the total variance in terms of percentage. In this example, the variance component for raters is negative. Negative estimates can result from using the wrong model or too small a sample. It should be noted that standard errors of variance components with small numbers of persons and conditions are very large. To obtain acceptable standard errors according to Smith [8], the sample should be at least a hundred persons. There are different approaches to dealing with negative estimates. One of them is to set the negative estimates to zero. The relatively large contribution of the variance component of items can be ascribed to the large differences between the mean scores of the items. The contribution of the interaction component between persons

Table 5 The item scores of six persons on four items and two raters, per rater the mean score per item and per person, the mean score per rater, the mean score per person, and the general mean

Pers.	Rater 1					$\bar{X}_{p1}$	Rater 2					$\bar{X}_{p2}$	$\bar{X}_p$
	Item:	1	2	3	4		Item:	1	2	3	4		
1		9	6	6	2	5.75		8	2	8	1	4.75	5.25
2		9	5	4	0	4.50		7	5	9	5	6.50	5.50
3		8	9	5	8	7.50		10	6	9	10	8.75	8.13
4		7	6	5	4	5.50		9	8	9	4	7.50	6.50
5		7	3	2	3	3.75		7	4	5	1	4.25	4.00
6		10	8	7	7	8.00		7	7	10	9	8.25	8.13
	$\bar{X}_{i1}$	8.33	6.17	4.83	4.00	5.83	$\bar{X}_{i2}$	8.00	5.33	8.33	5.00	6.67	$\bar{X} = 6.25$

Table 6 Results of the generalizability study for example from Table 5

Effects	Sums of squares	Degrees of freedom	Mean squares	Estimates of variance components
Persons (p)	109.75	5	21.95	$\hat{\sigma}_p^2 = 2.16$ (28%)
Items (i)	85.17	3	28.39	$\hat{\sigma}_i^2 = 1.26$ (15%)
Raters (r)	8.33	1	8.33	$\hat{\sigma}_r^2 = -0.15$ (0%)
Persons $\times$ Items (pi)	59.08	15	3.94	$\hat{\sigma}_{pi}^2 = 0.99$ (12%)
Persons $\times$ Raters (pr)	13.42	5	2.68	$\hat{\sigma}_{pr}^2 = 0.18$ (2%)
Items $\times$ Raters (ir)	33.83	3	11.28	$\hat{\sigma}_{ir}^2 = 1.55$ (19%)
Residual (pir, e)	29.42	15	1.96	$\hat{\sigma}_{pir,e}^2 = 1.96$ (24%)

and items is much larger than the interaction component between persons and raters. Interaction between persons and items means that persons do not give consistent reactions to different questions with the result that depending on the question the relative position of the persons differs.

Decision Study Two-facet Design

The linear model for the decomposition of the average score of a person on a test with n_i items of which the answers were rated by n_r raters, denoted by X_{pIR} , is

$$\begin{aligned}
 X_{pIR} = & \mu + (\mu_p - \mu) + (\mu_I - \mu) + (\mu_R - \mu) \\
 & + (\mu_{pI} - \mu_p - \mu_I + \mu) \\
 & + (\mu_{pR} - \mu_p - \mu_R + \mu) \\
 & + (\mu_{IR} - \mu_I - \mu_R + \mu) \\
 & + (X_{pIR} - \mu_{pI} - \mu_{pR} - \mu_{IR} \\
 & + \mu_p + \mu_I + \mu_R - \mu).
 \end{aligned}$$

The seven variance components associated with this model are $\sigma_p^2, \sigma_I^2 = \sigma_i^2/n_i', \sigma_R^2 = \sigma_r^2/n_r', \sigma_{pI}^2 = \sigma_{pi}^2/n_i', \sigma_{pR}^2 = \sigma_{pr}^2/n_r', \sigma_{IR}^2 = \sigma_{ir}^2/n_i'n_r', \sigma_{pIR,e}^2 = \sigma_{pir,e}^2/n_i'n_r'$.

The total variance is equal to $\sigma_X^2 = \sigma_p^2 + \sigma_I^2 + \sigma_R^2 + \sigma_{pI}^2 + \sigma_{pR}^2 + \sigma_{IR}^2 + \sigma_{pIR,e}^2$.

The estimate of the generalizability coefficient for relative decisions for the crossed two-facet random effects design is defined as

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{pr}^2}{n_r'} + \frac{\hat{\sigma}_{pir,e}^2}{n_i'n_r'}}.$$

The denominator of this coefficient has three variance components that relate to interactions with persons. Interaction between persons and items means that on certain items a person performs better than other persons, while on certain other items the performance is worse. This inconsistent performance by persons on items contributes to error variance. Interaction between persons and raters

means that a person is awarded different scores by different raters. This inconsistent rating by raters contributes to error variance. The residual variance component is by definition error variance and the interaction component between persons, items, and raters.

For the example in Table 5, with four items and two raters, the generalizability coefficient is equal to $2.16/(2.16 + 0.99/4 + 0.18/2 + 1.96/8) = 0.79$. This generalizability coefficient can be improved by increasing the number of observations, that is, the product of the number of items and the number of raters. Having more items, however, will have a much greater effect than more raters because the variance component of the interaction between persons and items is much larger than the variance component of the interaction between persons and raters. This example shows that the Spearman–Brown formula from classical theory does not apply to multifacet designs from generalizability theory. Procedures for selecting the optimal number of conditions in multifacet designs have been presented by Sanders, Theunissen, and Baas [5].

The estimate of the generalizability coefficient for absolute decisions for the crossed two-facet random effects design is defined as

$$\hat{\phi} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_i^2}{n'_i} + \frac{\hat{\sigma}_r^2}{n'_r} + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{pr}^2}{n'_r} + \frac{\hat{\sigma}_{ir}^2}{n'_i n'_r} + \frac{\hat{\sigma}_{pir,e}^2}{n'_i n'_r}}.$$

For making absolute decisions, it does matter whether we administer a test with difficult items or a test with easy items or have the answers rated by lenient or strict raters. Therefore, the variance components of the items and the raters, and the variance component of the interaction between items and raters also contribute to the error variance. For the example in Table 5, the generalizability coefficient for absolute decisions is equal to $2.16/(2.16 + 1.26/4 + 0.0/2 + 0.99/4 + 0.18/2 + 1.55/8 + 1.96/8) = 0.66$.

Other Designs

In the previous sections, it was shown that modifying the number of items and/or raters could affect the generalizability coefficient. However, the generalizability coefficient can also be affected by changing

the universe to which we want to generalize. We can, for example, change the universe by interpreting a random facet as a fixed facet. If the items in the example with four items and two raters are to be interpreted as a fixed facet, only these four items are admissible. If the facet ‘items’ is interpreted as a fixed facet, generalization is no longer to the universe of random parallel tests with four items and two raters, but to the universe of random parallel tests with two raters. Interpreting a random effect as a fixed facet means that fewer variance components can be estimated. In a crossed two-facet mixed effects design, the three variance components that can be estimated, expressed in terms of the variance components of the crossed two-facet random effects design, are $\hat{\sigma}_{p*}^2 = \hat{\sigma}_p^2 + \hat{\sigma}_{pi}^2/n'_i$, $\hat{\sigma}_{r*}^2 = \hat{\sigma}_r^2 + \hat{\sigma}_{ir}^2/n'_i$, and $\hat{\sigma}_{pr,e*}^2 = \hat{\sigma}_{pr}^2 + \hat{\sigma}_{pir,e}^2/n'_i$. The estimate of the generalizability coefficient for relative decisions for the crossed two-facet mixed effects design, originally derived by Maxwell and Pilliner [4], is defined as

$$\begin{aligned} \hat{\rho}^2 &= \frac{\hat{\sigma}_{p*}^2}{\hat{\sigma}_{p*}^2 + \hat{\sigma}_{pr,e*}^2/n'_r} \\ &= \frac{\hat{\sigma}_p^2 + \hat{\sigma}_{pi}^2/n'_i}{\hat{\sigma}_p^2 + \hat{\sigma}_{pi}^2/n'_i + \hat{\sigma}_{pr}^2/n'_r + \hat{\sigma}_{pir,e}^2/n'_i n'_r}. \end{aligned}$$

With the facet ‘items’ fixed, the generalizability coefficient for our example is equal to 0.88, compared to a generalizability coefficient of 0.79 with the facet ‘items’ being random. This increase of the coefficient is expected since, by restricting the universe, the relative decisions about persons will be more accurate.

In G theory, nested designs can also be analyzed. Our example with two facets would be a nested design if the first two questions were evaluated by the first rater and the other two questions by the second rater. In a design where raters are nested within questions, the variance component of raters and the variance component of the interaction between persons and raters cannot be estimated. The estimate of the generalizability coefficient for relative decisions for the nested two-facet random effects design is defined as

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{pr,pir,e}^2}{n'_i n'_r}}.$$

The estimates of variance components of the crossed two-facet random effects design can be used to estimate the variance components of not only a nested two-facet random effects design but also those of a nested two-facet mixed effects design. Because of their versatility, crossed designs should be given preference.

G theory is not limited to the analysis of univariate models; multivariate models where persons have more than one universe score can also be analyzed. In G theory, persons as well as facets can be selected as objects of measurement, making G theory a conceptual and statistical framework for a wide range of research problems from different disciplines.

References

- [1] Brennan, R.L. (1992). *Elements of Generalizability Theory*, ACT, Iowa.
- [2] Brennan, R.L. (2001). *Generalizability Theory*, Springer, New York.
- [3] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley, New York.
- [4] Maxwell, A.E. & Pilliner, A.E.G. (1968). Deriving coefficients of reliability and agreement, *The British Journal of Mathematical and Statistical Psychology* **21**, 105–116.
- [5] Sanders, P.F., Theunissen, T.J.J.M. & Baas, S.M. (1989). Minimizing the numbers of observations: a generalization of the Spearman-Brown formula, *Psychometrika* **54**, 587–598.
- [6] Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*, Sage Publications, Newbury Park.
- [7] Thorndike, R.L. (1982). *Applied Psychometrics*, Houghton Mifflin Company, Boston.
- [8] Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies, *Journal of Educational Statistics* **3**, 319–346.

PIET F. SANDERS

Generalizability Theory: Overview

NOREEN M. WEBB AND RICHARD J. SHAVELSON

Volume 2, pp. 717–719

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalizability Theory: Overview

Generalizability (G) theory is a statistical theory for evaluating the dependability ('reliability') of behavioral measurements [2]; see also [1], [3], and [4]. G theory pinpoints the sources of measurement error, disentangles them, and estimates each one. In G theory, a behavioral measurement (e.g., a test score) is conceived of as a sample from a *universe of admissible observations*, which consists of all possible observations that decision makers consider to be acceptable substitutes for the observation in hand. Each characteristic of the measurement situation that a decision maker would be indifferent to (e.g., test form, item, occasion, rater) is a potential source of error and is called a *facet of a measurement*. The universe of admissible observations, then, is defined by all possible combinations of the levels (called *conditions*) of the facets. In order to evaluate the dependability of behavioral measurements, a *generalizability (G) study* is designed to isolate and estimate as many facets of measurement error as is reasonably and economically feasible.

Consider a two-facet crossed person $\times$ item $\times$ occasion G study design where items and occasions have been randomly selected. The *object of measurement*, here persons, is not a source of error and, therefore, is not a facet. In this design with generalization over all admissible test items and occasions taken from an indefinitely large universe, an observed score for a particular person on a particular item and occasion is decomposed into an effect for the grand mean, plus effects for the person, the item, the occasion, each two-way interaction (see **Interaction Effects**), and a residual (three-way interaction plus unsystematic error). The distribution of each component or 'effect', except for the grand mean, has a mean of zero and a variance σ^2 (called the **variance component**). The variance component for the person effect is called the *universe-score variance*. The variance components for the other effects are considered error variation. Each variance component can be estimated from a traditional **analysis of variance** (or other methods such as **maximum likelihood**). The relative magnitudes of the estimated variance components provide information about sources of error influencing a behavioral measurement. Statistical tests are not

used in G theory; instead, standard errors for variance component estimates provide information about sampling variability of estimated variance components.

The *decision (D) study* deals with the practical application of a measurement procedure. A D study uses variance component information from a G study to design a measurement procedure that minimizes error for a particular purpose. In planning a D study, the decision maker defines the universe that he or she wishes to generalize to, called the *universe of generalization*, which may contain some or all of the facets and their levels in the universe of admissible observations. In the D study, decisions usually will be based on the mean over multiple observations (e.g., test items) rather than on a single observation (a single item).

G theory recognizes that the decision maker might want to make two types of decisions based on a behavioral measurement: relative ('norm-referenced') and absolute ('criterion- or domain-referenced'). A *relative decision* focuses on the rank order of persons; an *absolute decision* focuses on the level of performance, regardless of rank. Error variance is defined differently for each kind of decision. To reduce error variance, the number of conditions of the facets may be increased in a manner analogous to the Spearman–Brown prophecy formula in classical test theory and the standard error of the mean in sampling theory. G theory distinguishes between two reliability-like summary coefficients: a Generalizability (G) Coefficient for relative decisions and an Index of Dependability (Phi) for absolute decisions.

Generalizability theory allows the decision maker to use different designs in G and D studies. Although G studies should use crossed designs whenever possible to estimate all possible variance components in the universe of admissible observations, D studies may use nested designs for convenience or to increase estimated generalizability.

G theory is essentially a random effects theory. Typically, a random facet is created by randomly sampling levels of a facet. A fixed facet arises when the decision maker: (a) purposely selects certain conditions and is not interested in generalizing beyond them, (b) finds it unreasonable to generalize beyond the levels observed, or (c) when the entire universe of levels is small and all levels are included in the measurement design (see **Fixed and Random Effects**). G theory typically treats fixed facets by averaging over the conditions of the fixed facet and examining

2 Generalizability Theory: Overview

the generalizability of the average over the random facets. Alternatives include conducting a separate G study within each condition of the fixed facet, or a multivariate analysis with the levels of the fixed facet comprising a vector of dependent variables.

As an example, consider a G study in which persons responded to 10 randomly selected science items on each of 2 randomly sampled occasions. Table 1 gives the estimated variance components from the G study. The large $\hat{\sigma}_p^2$ (1.376, 32% of the total variation) shows that, averaging over items and occasions, persons in the sample differed systematically in their science achievement. The other estimated variance components constitute error variation; they concern the item facet more than the occasion facet. The non-negligible $\hat{\sigma}_i^2$ (5% of total variation) shows that items varied somewhat in difficulty level. The large $\hat{\sigma}_{pi}^2$ (20%) reflects different relative standings of persons across items. The small $\hat{\sigma}_o^2$ (1%) indicates that performance was stable across occasions, averaging over persons and items. The nonnegligible $\hat{\sigma}_{po}^2$ (6%) shows that the relative standing of persons differed somewhat across occasions. The zero $\hat{\sigma}_{io}^2$ indicates that the rank ordering of item difficulty was similar across

occasions. Finally, the large $\hat{\sigma}_{pio,e}^2$ (36%) reflects the varying relative standing of persons across occasions and items and/or other sources of error not systematically incorporated into the G study.

Because more of the error variability in science achievement scores came from items than from occasions, changing the number of items will have a larger effect on the estimated variance components and generalizability coefficients than will changing the number of occasions. For example, the estimated G and Phi coefficients for 4 items and 2 occasions are 0.72 and 0.69, respectively; the coefficients for 2 items and 4 occasions are 0.67 and 0.63, respectively. Choosing the number of conditions of each facet in a D study, as well as the design (nested vs. crossed, fixed vs. random facet), involves logistical and cost considerations as well as issues of dependability.

References

- [1] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.
- [2] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*, Wiley, New York.
- [3] Shavelson, R.J. & Webb, N.M. (1981). Generalizability theory: 1973–1980, *British Journal of Mathematical and Statistical Psychology* **34**, 133–166.
- [4] Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*, Sage Publications, Newbury Park.

(See also **Generalizability Theory: Basics; Generalizability Theory: Estimation**)

NOREEN M. WEBB AND RICHARD J.
SHAVELSON

Table 1 Estimated variance components in a generalizability study of science achievement (p × i × o design)

Source	Variance component	Estimate	Total variability (%)
Person (p)	σ_p^2	1.376	32
Item (i)	σ_i^2	0.215	05
Occasion (o)	σ_o^2	0.043	01
p × i	σ_{pi}^2	0.860	20
p × o	σ_{po}^2	0.258	06
i × o	σ_{io}^2	0.001	00
p × i × o, e	$\sigma_{pio,e}^2$	1.548	36

Generalized Additive Model

BRIAN S. EVERITT

Volume 2, pp. 719–721

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalized Additive Model

The **generalized linear model (GLM)** can accommodate nonlinear functions of the explanatory variables, for example, quadratic or cubic terms, if these are considered to be necessary to provide an adequate fit of a model for the observations. An alternative approach is to use a model in which the relationships between the response variable and the explanatory variables are modeled by **scatterplot smoothers**. This leads to generalized additive models described in detail in [1]. Such models are useful where

- the relationship between the variables is expected to be complex, not easily fitted by standard linear or nonlinear models;
- there is no *a priori* reason for using a particular model;
- we would like the data themselves to suggest the appropriate functional form.

Such models should be regarded as being philosophically closer to the concepts of **exploratory data**

analysis in which the form of any functional relationship emerges from the data rather than from a theoretical construct. In psychology, this can be useful because it reflects the uncertainty of the correct model to be applied in many situations.

In generalized additive models, the $\beta_i x_i$ term of **multiple linear regression** and **logistic regression** is replaced by a ‘smooth’ function of the explanatory variable x_i , as suggested by the observed data. Generalized additive models work by replacing the regression coefficients found in other regression models by the fit from one or other of these ‘smoothers’. In this way, the strong assumptions about the relationships of the response to each explanatory variable implicit in standard regression models are avoided. Details of how such models are fitted to data are given in [1].

Generalized additive models provide a useful addition to the tools available for exploring the relationship between a response variable and a set of explanatory variables. Such models allow possible nonlinear terms in the latter to be uncovered and then, perhaps, to be modeled in terms of a suitable, more familiar, low-degree polynomial. Generalized additive models can deal with nonlinearity in covariates that are not of main interest in a study and can ‘adjust’ for such effects appropriately.

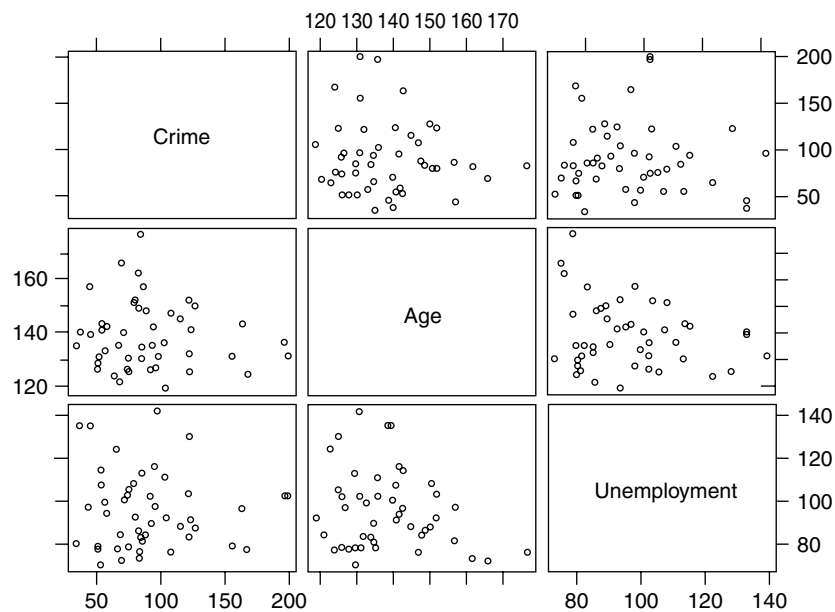


Figure 1 Scatterplot matrix of the data on crime in the United States

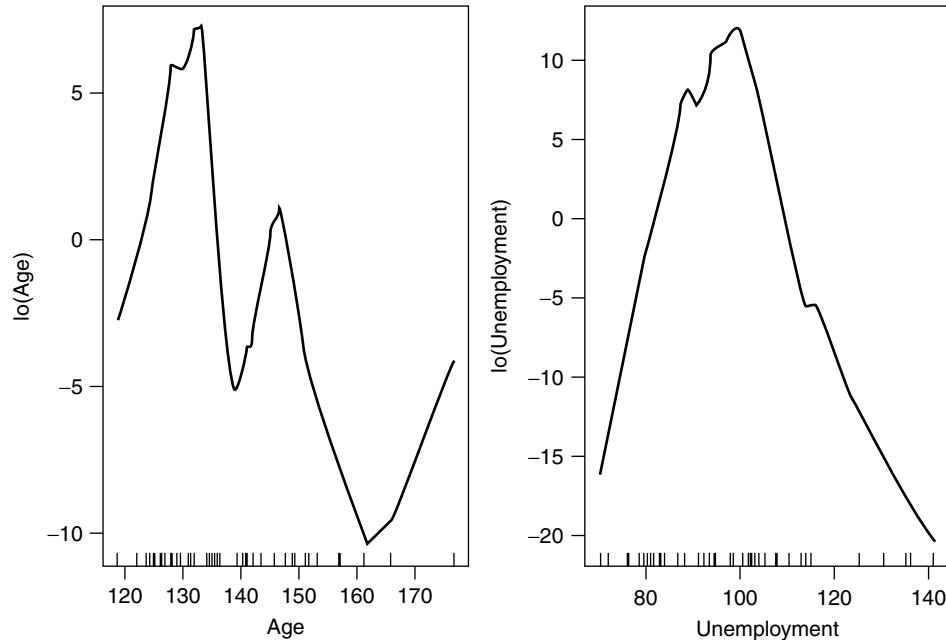


Figure 2 Form of locally weighted regression fit for crime rate and age [$\text{lo}(\text{age})$ represents the lowess fit—see scatterplot smoothers], and crime rate and unemployment [$\text{lo}(\text{unemployment})$ represents the lowess fit] of locally weighted regression fit for crime rate and age and crime rate and unemployment

As an example of the application of GAMs, we consider some data on crime rates in the United States given in [2]. The question of interest is how crime rate (number of offenses known to the police per one million population) in different states of the United States is related to the age of males in the age group 14 to 24 per 1000 of the total state population and to unemployment in urban males per 1000 population in the age group 14 to 24. A **scatterplot matrix** of the data is shown in Figure 1 and suggests that the relationship between crime rate and each of the other two variables *may* depart from linearity in some subtle fashion that is worth investigating using a GAM. Using a locally weighted regression to model the relationship between crime rate and each of the explanatory variables, the model can be fitted simply using software available in, for example, *SAS* or *S-PLUS* (see **Software for Statistical Analyses**). Rather than giving the results in detail, we simply show the locally weighted fits of crime rate on age and unemployment in Figure 2.

The locally weighted regression fit for age suggests, perhaps, that a linear fit for crime rate on age might be appropriate, with crime declining with an increasingly aged state population. But the relationship between crime rate and unemployment is clearly nonlinear. Use of the GAM suggests, perhaps, that crime rate might be modeled by a multiple regression approach with a linear term for age and a quadratic term for unemployment.

References

- [1] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, CRC Press/Chapman & Hall, Boca Raton.
- [2] Vandehe, W. (1978). Participation in illegitimate activities: Erlich revisited, in *Deterrence and Incapacitation*, A. Blumstein, J. Cohen & D. Nagin, eds, National Academy of Science, Washington.

BRIAN S. EVERITT

Generalized Estimating Equations (GEE)

JAMES W. HARDIN

Volume 2, pp. 721–729

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalized Estimating Equations (GEE)

Introduction

The usual practice in model construction is the specification of the systematic and random components of variation. Classical **maximum likelihood** models then rely on the validity of the specified components. Model construction proceeds from the (components of variation) specification to a likelihood and, ultimately, an *estimating equation*. The estimating equation for maximum likelihood estimation is obtained by equating zero to the derivative of the log-likelihood with respect to the parameters of interest. Point estimates of unknown parameters are obtained by solving the estimating equation.

Generalized Linear Models

The theory and an algorithm appropriate for obtaining maximum likelihood estimates where the response follows a distribution in the exponential family (see **Generalized Linear Models (GLM)**) was introduced in [16]. This reference introduced the term generalized linear models (GLMs) to refer to a class of models which could be analyzed by a single algorithm. The theoretical and practical application of (GLMs) has since received attention in many articles and books; see especially [14].

GLMs encompass a wide range of commonly used models such as linear regression (see **Multiple Linear Regression**), **logistic regression** for binary outcomes, and Poisson regression (see **Generalized Linear Models (GLM)**) for count data outcomes. The specification of a particular GLM requires a link function that characterizes the relationship of the mean response to a vector of covariates. In addition, a GLM requires specification of a variance function that relates the variance of the outcomes as a function of the mean.

The derivation of the iteratively reweighted least squares (see **Generalized Linear Mixed Models**) (IRLS) algorithm appropriate for fitting GLMs begins with the likelihood specification for the exponential family. Within an iterative algorithm, an updated estimate of the coefficient vector may be obtained via

weighted ordinary least squares where the weights are related to the link and variance specifications. The estimation is then iterated to convergence where convergence may be defined, for example, as the change in the estimated coefficient vector being smaller than some tolerance.

For any response that follows a member of the exponential family of distributions, $f(y) = \exp\{[y\theta - b(\theta)]/\phi + c(y, \phi)\}$, where θ is the canonical parameter and ϕ is a proportionality constant, we can obtain maximum likelihood estimates of the $p \times 1$ regression coefficient vector β by solving the estimating equation given by

$$\Psi_\beta = \sum_{i=1}^n \Psi_i = \sum_{i=1}^n \mathbf{x}_i^T \left\{ \frac{y_i - \mu_i}{\phi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right\} = \mathbf{0}_{p \times 1}. \quad (1)$$

In the estimation equation $\mathbf{x}_i$ is the i th row of an $n \times p$ matrix of covariates $\mathbf{X}$, $\mu_i = g^{-1}(\mathbf{x}_i \beta)$ represents the expected outcome $E(y) = b'(\theta)$ in terms of a transformation of the linear predictor $\eta_i = \mathbf{x}_i \beta$ via a monotonic (invertible) link function $g()$, and the variance $V(\mu_i)$ is a function of the expected value proportional to the variance of the outcome $V(y_i) = \phi V(\mu_i)$. The estimating equation is also known as the score equation since it equates the score vector Ψ_β to zero.

Modelers are free to choose a link function (see **Generalized Linear Models (GLM)**) as well as a variance function. If the link-variance pair of functions are chosen from a common member of the exponential family of distributions, the resulting estimates are equivalent to maximum likelihood estimates. However, modelers are not limited to these choices. When one selects variance and link functions that do not coincide to a particular exponential family member distribution, the estimating equation is said to imply a quasilielihood, (see **Generalized Linear Models (GLM)**) and the resulting estimates are referred to as maximum quasilielihood estimates.

The link function that equates the canonical parameter θ with the linear predictor $\eta_i = \mathbf{x}_i \beta$ is called the canonical link. If this link is selected, the estimating equation simplifies to

$$\Psi_\beta = \sum_{i=1}^n \Psi_i = \sum_{i=1}^n \mathbf{x}_i^T \left\{ \frac{y_i - \mu_i}{\phi} \right\} = \mathbf{0}_{p \times 1}. \quad (2)$$

One advantage of the canonical link over other link functions is that the expected Hessian matrix is

equal to the observed Hessian matrix. This means that the model-based variance estimate (inverse of the expected Hessian) usually provided by the IRLS algorithm for GLMs will be the same as the model-based variance estimate (inverse of the observed Hessian) usually provided from a maximum likelihood algorithm. One should note, however, that this property does not automatically mean that the canonical link function is the best choice for a given dataset.

The large sample covariance matrix of the estimated regression coefficients $\hat{\beta}$ may be estimated using the inverse of the expected **information matrix** (the expectation of the matrix outer product of the scores $\sum_{i=1}^n \Psi_i \Psi_i^T$), or the inverse of the observed information matrix (matrix of derivatives of the score vector, $\partial \Psi_{\beta} / \partial \beta$). These two variance estimators, evaluated at $\hat{\beta}$, are the same if the canonical link is used.

The Independence Model

A basic individual-level model is written in terms of the n individual observations y_i for $i = 1, \dots, n$. When observations may be clustered (*see Clustered Data*), owing to repeated observations on the sampling unit or because the observations are related to some cluster identifier variable, the model may be written in terms of the observations y_{it} for the clusters $i = 1, \dots, n$ and the within-cluster repeated, or related, observations $t = 1, \dots, n_i$. The total number of observations is then $N = \sum_i n_i$. The clusters may also be referred to as panels, subjects, or groups. In this presentation, the clusters i are independent, but the within-clusters observations it may be correlated. An independence model, however, assumes that the within-cluster observations are not correlated.

The independence model is a special case of more sophisticated correlated data approaches (such as GEE). This model assumes that there is no correlation within clusters. Therefore, the model specification is in terms of the individual observations y_{it} . While the independence model assumes that the repeated measures are independent, the model still provides consistent estimators in the presence of correlated data. Of course, this approach is paid for through inefficiency, though the efficiency loss is not always large as investigated by Glonek et al. [5]. As such, this model remains an attractive alternative because of its computational simplicity. The independence

model also serves as a reference model in the derivation of diagnostics for more sophisticated models for clustered data (such as GEE models).

Analysts can use the independence model to obtain point estimates $\hat{\beta}$ along with standard errors based on the modified sandwich variance estimator to ensure that inference is robust to any type of within-cluster correlation. While the inference regarding marginal effects is valid (assuming that the model for the mean is correctly specified), the estimator from the independence model is not efficient when the data are correlated.

Modified Sandwich Variance Estimator

The validity of the (naive) model-based variance estimators, using the inverse of either the observed or expected Hessian, depends on the correct specification of the variance; in turn this depends on the correct specification of the working correlation model. A formal justification for an alternative estimator known as the sandwich variance estimator is given in [9].

The sandwich variance estimator is presented in the general form $\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-T}$. Here $\mathbf{A}^{-1}$ (the so-called ‘bread’ of the sandwich) is the standard model-based (naive) variance estimator which can be based on the expected Hessian or the observed Hessian (*see Information Matrix*). The $\mathbf{B}$ variance estimator is the sum of the cross-products of the scores.

The $\mathbf{B}$ variance estimator does not depend on the correct specification of the assumed model and is given by $\mathbf{B} = \sum_{i=1}^n \sum_{t=1}^{n_i} \Psi_{it} \Psi_{it}^T$. As the expected value of the estimating equation is zero, this formula is similar to the usual variance estimator. A generalization is obtained by squaring the sums of the terms for each cluster (since we assume that the clusters are independent) instead of summing the squares of the terms for each observation. This summation over clusters $\mathbf{B} = \sum_{i=1}^n [\sum_{t=1}^{n_i} \Psi_{it}] [\sum_{t=1}^{n_i} \Psi_{it}^T]$ is what adds the *modified* adjective to the modified sandwich variance estimator.

The beneficial properties of the sandwich variance estimator, in the usual or the modified form, make it a popular choice for many analysts. However, the acceptance of this estimator is not without some controversy. A discussion of the decreased efficiency and increased variability of the sandwich estimator in common applications is presented in [11], and [3]

argues against blind application of the sandwich estimator by considering an independent samples test of means.

It should be noted that assuming independence is not always conservative; the model-based (naive) variance estimates based on the observed or expected Hessian matrix are not always smaller than those of the modified sandwich variance estimator. Since the sandwich variance estimator is sometimes called the robust variance estimator, this result may seem counterintuitive. However, it is easily seen by assuming negative within-cluster correlation leading to clusters with both positive and negative residuals. The cluster-wise sums of those residuals will be small and the resulting modified sandwich variance estimator will yield smaller standard errors than the model-based Hessian variance estimators.

Subject-specific (SS) versus Population-averaged (PA) Models

There are two main approaches to dealing with correlation in repeated or **longitudinal data**. One approach focuses on the marginal effects averaged across the individuals (*see Marginal Models for Clustered Data*) (population-averaged approach), and the second approach focuses on the effects for given values of the random effects by fitting parameters of the assumed random-effects distribution (subject-specific approach). Formally, we specify a generalized linear mixed model and include a source of the nonindependence. We can then either explicitly model the conditional expectation given the random effects γ_i using $\mu_{it}^{SS} = E(y_{it} | \mathbf{x}_{it}, \gamma_i)$, or we can focus on the marginal expectation (integrated over the distribution of the random effects) as $\mu_{it}^{PA} = E_{\gamma_i}[E(y_{it} | \mathbf{x}_{it}, \gamma_i)]$. The responses in these approaches are characterized by

$$\begin{aligned} g(\mu_{it}^{SS}) &= \mathbf{x}_{it} \boldsymbol{\beta}^{SS} + \mathbf{z}_{it} \gamma_i \\ V(y_{it} | \mathbf{x}_{it}, \gamma_i) &= V(\mu_{it}^{SS}) \\ g(\mu_{it}^{PA}) &= \mathbf{x}_{it} \boldsymbol{\beta}^{PA} \\ V(y_{it} | \mathbf{x}_{it}) &= \phi V(\mu_{it}^{PA}). \end{aligned} \quad (3)$$

The population-averaged approach models the average response for observations sharing the same covariates (across all of the clusters or subjects). The superscripts are used to emphasize that the

fitted coefficients are not the same. The subject-specific approach explicitly models the source of heterogeneity so that the fitted regression coefficients have an interpretation in terms of the individuals.

The most commonly applied GEE is described in [12]. This is a population-averaged approach. It is possible to derive subject-specific GEE models, but such models are not currently part of software packages and so do not appear nearly as often in the literature.

(Population-averaged) Generalized Estimating Equations

The genesis of population-averaged generalized estimating equations is presented in [12]. The basic idea behind this novel approach is illustrated as follows. We consider the estimating equation for a model specifying the exponential family of distributions

$$\begin{aligned} \Psi_{\beta} &= \sum_{i=1}^n \Psi_i = \sum_{i=1}^n \mathbf{X}_i^T \left[\left\{ \mathbf{D} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \right) [\mathbf{V}(\boldsymbol{\mu}_i)]^{-1} \right. \right. \\ &\quad \left. \left. \times \left(\frac{\mathbf{y}_i - \boldsymbol{\mu}_i}{\phi} \right) \right\} \right] = \mathbf{0}_{p \times 1}, \end{aligned} \quad (4)$$

where $\mathbf{D}(\mathbf{d}_i)$ denotes a diagonal matrix with diagonal elements given by the $n_i \times 1$ vector $\mathbf{d}_i$, $\mathbf{X}_i$ is the $n_i \times p$ matrix of covariates for cluster i , and $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})$ are $n_i \times 1$ vectors for cluster i . Assuming independence, $\mathbf{V}(\boldsymbol{\mu}_i)$ is clearly an $n_i \times n_i$ diagonal matrix which can be factored into

$$\mathbf{V}(\boldsymbol{\mu}_i) = [\mathbf{D}(V(\mu_{it}))^{1/2} \mathbf{I}_{(n_i \times n_i)} \mathbf{D}(V(\mu_{it}))^{1/2}]_{n_i \times n_i}, \quad (5)$$

where $\mathbf{D}(d_{it})$ is a $n_i \times n_i$ diagonal matrix with diagonal elements d_{it} for $t = 1, \dots, n_i$. This presentation makes it clear that the estimating equation treats each observation within a cluster as independent. A (pooled) model associated with this estimating equation is called the *independence model*.

There are two other aspects of the estimating equation to note. The first aspect is that the estimating equation is written in terms of $\boldsymbol{\beta}$ while the scale parameter ϕ is treated as ancillary. For discrete families, this parameter is theoretically equal to one, while for continuous families ϕ is a scalar multiplying the assumed variance (ϕ is estimated in this case).

4 Generalized Estimating Equations (GEE)

The second aspect of the estimating equation to note is that it is written in terms of the clusters i instead of the observations it .

The genesis of the original population-averaged generalized estimating equations is to replace the identity matrix with a parameterized *working correlation matrix* $\mathbf{R}(\alpha)$.

$$\mathbf{V}(\mu_i) = [\mathbf{D}(V(\mu_{it}))^{1/2} \mathbf{R}(\alpha)_{(n_i \times n_i)} \times \mathbf{D}(V(\mu_{it}))^{1/2}]_{n_i \times n_i}. \quad (6)$$

To address correlated data, the working correlation matrix is parameterized via α in order to specify structural constraints section ‘Estimating the Working Correlation Matrix’. In this way, the independence model is a special case of the GEE specifications where $\mathbf{R}(\alpha)$ is an identity matrix.

Formally, [12] introduces a second estimating equation for the parameters of the working correlation matrix. The authors then establish the properties of the estimators resulting from the solution of these estimating equations. The GEE moniker was applied as the model is derived through a generalization of the estimating equation rather than a derivation from some assumed distribution. Example applications of these models in behavioral statistics studies can be found in [4] and [1].

GEE is a generalization of the quasilielihood approach to GLMs which merely uses first and second moments and does not require a likelihood. There are several software packages that support estimation of these models. These packages include R, SAS, S-PLUS, Stata, and SUDAAN. R and S-PLUS users can easily find user-written software tools for fitting GEE models, while such support is included in the other packages (*see Software for Statistical Analyses*).

Estimating the Working Correlation Matrix

One should carefully consider the parameterization of the working correlation matrix since including the correct parameterization leads to more efficient estimates. We want to carefully consider this choice even if we employ the modified sandwich variance estimator in the calculation of standard errors and confidence intervals for the regression parameters. While the use of the modified sandwich variance estimator

assures robustness in the case of misspecification of the working correlation matrix, the advantage of more efficient point estimates is still worth this effort.

There is no controversy as to the fact that the GEE estimates are consistent, but there is some controversy as to how efficient they are. This controversy centers on how well the correlation parameters can be estimated.

The full generalized estimating equation for population-averaged GEEs is given in partitioned form by $\Psi = (\Psi_\beta, \Psi_\alpha) = (\mathbf{0}, \mathbf{0})$, where the regression β and correlation α components are given by

$$\begin{aligned} \Psi_\beta &= \sum_{i=1}^n \mathbf{X}_i^T \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \mathbf{V}^{-1}(\mu_i) \left(\frac{\mathbf{y}_i - \mu_i}{\phi} \right) = \mathbf{0} \\ \Psi_\alpha &= \sum_{i=1}^n \left(\frac{\partial \xi_i}{\partial \alpha} \right) \mathbf{H}_i^{-1} (\mathbf{W}_i - \xi_i) = \mathbf{0}, \end{aligned} \quad (7)$$

where $\mathbf{W}_i = (r_{i1}r_{i2}, r_{i1}r_{i3}, \dots, r_{in_i-1}r_{in_i})^T$, $\mathbf{H}_i = \mathbf{D}(V(\mathbf{W}_i))$, and $\xi_i = E(\mathbf{W}_i)$. From this specification (using r_{it} for the it th Pearson residual), it is clear that the parameterization of the working correlation matrix enters through the specification of ξ . For example, the specification $\xi = (\alpha, \alpha, \dots, \alpha)$ signals a single unknown correlation; we assume that the conditional correlations for all pairs of observations within a given cluster are the sample. For instance, the correlations do not depend on a time lag.

Typically a careful analyst chooses some small number of candidate parameterizations. The quasilielihood information criterion (QIC) measures for choosing between candidate parameterizations is discussed in [17]. This criterion measure is similar to the well known **Akaike information criterion** (AIC).

The most common choices for parameterizing the working correlation $\mathbf{R}$ matrix are then given by parameterizing the elements of the matrix as

$$\begin{aligned} \text{independent} & R_{uv} = 0 \\ \text{exchangeable} & R_{uv} = \alpha \\ \text{autocorrelated} & \\ \quad - \text{AR}(1) & R_{uv} = \alpha^{|u-v|} \\ \text{stationary}(k) & R_{uv} = \begin{cases} \alpha_{|u-v|} & \text{if } |u-v| \leq k \\ 0 & \text{otherwise} \end{cases} \\ \text{nonstationary}(k) & R_{uv} = \begin{cases} \alpha_{uv} & \text{if } |u-v| \leq k \\ 0 & \text{otherwise} \end{cases} \\ \text{unstructured} & R_{uv} = \alpha_{uv} \end{aligned} \quad (8)$$

for $u \neq v$; $R_{uu} = 1$.

The independence model admits no extra parameters and the resulting model is equivalent to a generalized linear model specification. The exchangeable correlation parameterization admits one extra parameter and the unstructured working correlation parameterization admits $\binom{M}{2} - M$ extra parameters where $M = \max\{n_i\}$. The exchangeable correlation specification is also known as *equal correlation*, *common correlation*, and *compound symmetry* (see **Sphericity Test**).

The elements of the working correlation matrix are estimated using the Pearson residuals from the current fit. Estimation alternates between estimating the regression parameters β for the current estimates of α , and then using those β estimates to obtain residuals to update the estimate of α .

In addition to estimating (α, β) , the continuous families also require estimation of the scale parameter ϕ ; this is the same scale parameter as in generalized linear models. Discrete families theoretically define this parameter to be 1, but one can optionally estimate this parameter in the same manner as is required by continuous exponential family members. Software documentation should specify the conditions under which the parameter is either assumed to be known or is estimated.

The usual approach in GLMs for $N = \sum_i n_i$ total observations is to estimate the ϕ scale parameter as $1/N \sum_{i=1}^n \sum_{j=1}^{n_i} r_{ij}^2$, though some software packages will use $(N - p)$, where p is the dimension of β , as the denominator. Software users should understand that this seemingly innocuous difference will lead to slightly different answers in various software packages. The scale parameter is the denominator in the estimation of the correlation parameters and a change in the estimates of the correlation parameters α will lead to slightly different regression coefficient estimates β .

Extensions to the Population-averaged GEE Model

The GEE models described in [12] are so commonly used that analyses simply refer to their application as GEE. However, GEE derivations are not limited to population-averaged models. In fact, generalized estimating equations methods can be applied in the construction of subject-specific

models; see section ‘Subject-specific (SS) versus Population-averaged (PA) Models’ in this entry and [25].

Several areas of research have led to extensions of the original GEE models. The initial extensions were to regression models not usually supported in generalized linear models. In particular, generalized logistic regression models for multinomial logit, cumulative logistic regression models, and ordered outcome models (ordered logistic and ordered probit) have all found support in various statistical software packages.

An extension of the quasiliikelihood such that both partial derivatives have score-like properties is given in [15], and then [7], and later [6], derive an extended generalized estimating equation (EGEE) model from this extended quasiliikelihood. To give some context to this extension, the estimating equation for β does not change, but the estimating equation for α is then

$$\Psi_{\alpha} = \sum_{i=1}^n \left[-(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \frac{\partial \mathbf{V}(\boldsymbol{\mu}_i)^{-1}}{\partial \boldsymbol{\alpha}} (\mathbf{y}_i - \boldsymbol{\mu}_i) + \text{tr} \left(\mathbf{V}(\boldsymbol{\mu}_i) \frac{\partial \mathbf{V}(\boldsymbol{\mu}_i)^{-1}}{\partial \boldsymbol{\alpha}} \right) \right] = \mathbf{0}. \quad (9)$$

The EGEE model is similar to the population-averaged GEE model in that the two estimating equations are assumed to be orthogonal; it is assumed that $\text{Cov}(\beta, \alpha) = \mathbf{0}$ – a property usually referred to in the literature as GEE1.

At the mention of GEE1, it should be obvious that there is another extension to the original GEE model known as GEE2. A model derived from GEE2 does not assume that β and α are uncorrelated. The GEE2, which is not robust against misspecification of the correlation, is a more general approach that has less restrictions and which provides standard errors for the correlation parameters α . Standard errors are not generally available in population-averaged GEE models though one can calculate bootstrap standard errors.

One other extension of note is the introduction of estimating methods that are resistant to **outliers**. One such approach by Preisser and Qaqish [19] generalizes GEE model estimation following the ideas in robust regression. This generalization down-weights outliers to remove exaggerated influence. The estimating equation for the regression coefficients becomes

$$\Psi_{\beta} = \sum_{i=1}^n \mathbf{D} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i} \right) \mathbf{V}(\boldsymbol{\mu}_i)^{-1} \times \left(\mathbf{w}_i \frac{\mathbf{y}_i - \boldsymbol{\mu}_i}{\phi} - \mathbf{c}_i \right) = \mathbf{0}_{p \times 1}. \quad (10)$$

The usual GEE is a special case where, for all i , the weights $\mathbf{w}_i$ are given by an $n_i \times n_i$ identity matrix and $\mathbf{c}_i$ by a vector of zeros. Typical approaches use Mallows-type weights calculated from influence diagnostics, though other approaches are possible.

Missing Data

Population-averaged GEE models are derived for complete data. If there are missing observations, the models are still applicable if the data are missing completely at random (MCAR).

Techniques for dealing with **missing data** are a source of active research in all areas of statistical modeling, but methods for dealing with missing data are difficult to implement as turnkey solutions. This means that software packages are not likely to support specific solutions to every research problem. An investigation into the missingness of data requires, as a first step, the means for communicating the nature of the missing data.

If data are not missing completely at random, then an application of GEE analysis is performed under a violation of assumptions leading to suspect results and interpretation. Analyses that specifically address data that do not satisfy the MCAR assumption are referred to as *informatively missing* methods; for further discussion see [22] for applications of inverse probability weighting and [10] for additional relevant discussion.

A formal study for modeling missing data due to dropouts is presented in [13], while [22] and [21] each discuss the application of sophisticated semi-parametric methods under non-ignorable missingness mechanisms which extend usual GEE models to provide consistent estimators. One of the assumptions of GEE is that if there is dropout, the dropout mechanism (*see Dropouts in Longitudinal Studies: Methods of Analysis*) does not depend on the values of the outcomes (outcome-dependent dropout), but as [13] points out, such missingness may depend on the values of the fixed covariates (covariate-dependent dropout).

Diagnostics

One of the most prevalent measures for model adequacy is the Akaike information criterion or AIC. An extension of this measure, given in [17], is called the quaslikelihood information criterion (QIC). This measure is useful for comparing models that differ only in the assumed correlation structure. For choosing covariates in the model, [18] introduces the QIC_u measure that plays a similar role for covariate selection in GEE models as the adjusted R^2 plays in regression.

Since the MCAR is an important assumption in GEE models, [2] provides evidence of the utility of the Wald–Wolfowitz nonparametric **run test**. This test provides a formal approach for assessing compliance of a dataset to the MCAR assumption. While this test is useful, one should not forget the basics of exploratory data analysis. The first assessment of the data and the missingness of the data should be subjectively illustrated through standard graphical techniques.

As in GLMs, the careful investigator looks at influence measures of the data. Standard DFBETA and DFFIT residuals introduced in the case of linear regression are generalized for clustered data analysis by considering deletion diagnostics based on deleting a cluster i at a time rather than an observation it at a time. For goodness-of-fit, [26] provides discussion of measures based on entropy (as a proportional reduction in variation), along with discussion in terms of the concordance correlation.

A χ^2 goodness-of-fit test for GEE binomial models is presented in [8]. The basic idea of the test is to group results into deciles and investigate the frequencies as a χ^2 test of the expected and observed counts. As with the original test, analysts should use caution if there are many ties at the deciles since breaking the ties will be a function of the sort order of the data. In other words, the results will be random.

Standard Wald-type hypothesis tests of regression coefficients can be performed using the estimated covariance matrix of the regression parameters. In addition, [23] provides alternative extensions of Wald, Rao (score), and likelihood ratio tests (deviance difference based on the independence model). These tests are available in the SAS commercial packages via specified contrasts.

Example

To highlight the interpretation of GEE analyses and point out the alternate models, we focus on a simple example. The data are given in Table 1.

Table 1 Number of seizures for four consecutive time periods for 59 patients

id	age	trt	base	s1	s2	s3	s4
1	31	0	11	5	3	3	3
2	30	0	11	3	5	3	3
3	25	0	6	2	4	0	5
4	36	0	8	4	4	1	4
5	22	0	66	7	18	9	21
6	29	0	27	5	2	8	7
7	31	0	12	6	4	0	2
8	42	0	52	40	20	23	12
9	37	0	23	5	6	6	5
10	28	0	10	14	13	6	0
11	36	0	52	26	12	6	22
12	24	0	33	12	6	8	5
13	23	0	18	4	4	6	2
14	36	0	42	7	9	12	14
15	26	0	87	16	24	10	9
16	26	0	50	11	0	0	5
17	28	0	18	0	0	3	3
18	31	0	111	37	29	28	29
19	32	0	18	3	5	2	5
20	21	0	20	3	0	6	7
21	29	0	12	3	4	3	4
22	21	0	9	3	4	3	4
23	32	0	17	2	3	3	5
24	25	0	28	8	12	2	8
25	30	0	55	18	24	76	25
26	40	0	9	2	1	2	1
27	19	0	10	3	1	4	2
28	22	0	47	13	15	13	12
29	18	1	76	11	14	9	8
30	32	1	38	8	7	9	4
31	20	1	19	0	4	3	0
32	20	1	10	3	6	1	3
33	18	1	19	2	6	7	4
34	24	1	24	4	3	1	3
35	30	1	31	22	17	19	16
36	35	1	14	5	4	7	4
37	57	1	11	2	4	0	4
38	20	1	67	3	7	7	7
39	22	1	41	4	18	2	5
40	28	1	7	2	1	1	0
41	23	1	22	0	2	4	0
42	40	1	13	5	4	0	3
43	43	1	46	11	14	25	15
44	21	1	36	10	5	3	8
45	35	1	38	19	7	6	7

Table 1 (continued)

id	age	trt	base	s1	s2	s3	s4
46	25	1	7	1	1	2	4
47	26	1	36	6	10	8	8
48	25	1	11	2	1	0	0
49	22	1	151	102	65	72	63
50	32	1	22	4	3	2	4
51	25	1	42	8	6	5	7
52	35	1	32	1	3	1	5
53	21	1	56	18	11	28	13
54	41	1	24	6	3	4	0
55	32	1	16	3	5	4	3
56	26	1	22	1	23	19	8
57	21	1	25	2	3	0	1
58	36	1	13	0	0	0	0
59	37	1	12	1	4	3	2

The data have been analyzed in many forums; data values are also available in [24] (along with other covariates). The data are from a panel study on Progabide treatment of epilepsy. Baseline measures of the number of seizures in an eight-week period were collected and recorded as baseline for 59 patients. Four follow-up two-week periods also counted the number of seizures; these were recorded as s1, s2, s3, and s4. The baseline variable was divided by four in our analyses to put it on the same scale as the follow-up counts. The age variable records the patient's age in years, and the trt variable indicates whether the patient received the Progabide treatment (value recorded as one) or was part of the control group (value recorded as zero).

An obvious approach to analyzing the data is to hypothesize a Poisson model for the number of seizures. Since we have repeated measures (*see Repeated Measures Analysis of Variance*), we can choose a number of alternative approaches. In our illustrations of these alternative models, we utilize the baseline measure as a covariate along with the time and age variables.

Table 2 contains the results of several analyses. For each covariate, we list the estimated incidence rate ratio (exponentiated coefficient). Following the incidence rate ratio estimates, we list the classical and sandwich-based estimated standard errors. We did not calculate sandwich-based standard errors for the gamma-distributed random-effects model.

We emphasize again that the independence model coupled with standard errors based on the modified sandwich variance estimator is a valid approach

8 Generalized Estimating Equations (GEE)

Table 2 Estimated incidence rate ratios and standard errors for various Poisson models

Model	time	trt	age	baseline
Independence	0.944 (0.019,0.033)	0.832 (0.039,0.143)	1.019 (0.003,0.010)	1.095 (0.002,0.006)
Gamma RE	0.944 (0.019)	0.810 (0.124)	1.013 (0.011)	1.116 (0.015)
Normal RE	0.944 (0.019,0.033)	0.760 (0.117,0.117)	1.011 (0.011,0.009)	1.115 (0.012,0.011)
GEE(exch)	0.944 (0.015,0.033)	0.834 (0.058,0.141)	1.019 (0.005,0.010)	1.095 (0.003,0.006)
GEE(ar1)	0.939 (0.019,0.019)	0.818 (0.054,0.054)	1.021 (0.005,0.003)	1.097 (0.003,0.003)
GEE(unst)	0.951 (0.017,0.041)	0.832 (0.055,0.108)	1.019 (0.005,0.009)	1.095 (0.003,0.005)

to modeling data of this type. The weakness of the approach is that the estimators will not be as efficient as a model including the true underlying within-cluster correlation structure. Another standard approach to modeling this type of repeated measures is to hypothesize that the correlations are due to individual-specific random intercepts (*see Generalized Linear Mixed Models*). These random effects (one could also hypothesize fixed effects) will lead to alternate models for the data.

Results from two different random-effects models are included in the table. The gamma-distributed random-effects model is rather easy to program and fit to data as the log-likelihood of the model is in analytic form. The normally distributed random-effects model on the other hand has a log-likelihood specification that includes an integral. Sophisticated numeric techniques are required for the calculation of this model; see [20].

We could hypothesize that the correlation follows an autoregressive process since the data are collected over time. However, this is not always the best choice in an experiment since we must believe that the hypothesized correlation structure applies to both the treated and untreated groups.

The QIC values for the independence, exchangeable, ar1, and unstructured correlation structures are respectively given by -5826.23 , -5826.25 , -5832.20 , and -5847.91 . This criterion measure indicates a preference for the unstructured model over the autoregressive model. The fitted correlation matrices for these models (printing only the bottom half of the symmetric matrices) are given by

$$R_{AR(1)} = \begin{bmatrix} 1.00 & & & \\ 0.51 & 1.00 & & \\ 0.26 & 0.51 & 1.00 & \\ 0.13 & 0.26 & 0.51 & 1.00 \end{bmatrix}$$

$$R_{unst} = \begin{bmatrix} 1.00 & & & \\ 0.25 & 1.00 & & \\ 0.42 & 0.68 & 1.00 & \\ 0.22 & 0.28 & 0.58 & 1.00 \end{bmatrix}. \quad (11)$$

References

- [1] Alexander, J.A., D'Aunno, T.A. & Succi, M.J. (1996). Determinants of profound organizational change: choice of conversion or closure among rural hospitals, *Journal of Health and Social Behavior* **37**, 238–251.
- [2] Chang, Y.-C. (2000). Residuals analysis of the generalized linear models for longitudinal data, *Statistics in Medicine* **19**, 1277–1293.
- [3] Drum, M. & McCullagh, P., Comment on Fitzmaurice, G.M., Laird, N. & Rotnitzky, A. (1993). Regression models for discrete longitudinal responses, *Statistical Science* **8**, 284–309.
- [4] Ennet, S.T., Flewelling, R.L., Lindrooth, R.C. & Norton, E.C. (1997). School and neighborhood characteristics associated with school rates of alcohol, cigarette, and marijuana use, *Journal of Health and Social Behavior* **38**(1), 55–71.
- [5] Glonek, G.F.V. & McCullagh, R. (1995). Multivariate logistic models, *Journal of the Royal Statistical Society – Series B* **57**, 533–546.
- [6] Hall, D.B. (2001). On the application of extended quasilielihood to the clustered data case, *The Canadian Journal of Statistics* **29**(2), 1–22.
- [7] Hall, D.B. & Severini, T.A. (1998). Extended generalized estimating equations for clustered data, *Journal of the American Statistical Association* **93**, 1365–1375.
- [8] Horton, N.J., Bebhuk, J.D., Jones, C.L., Lipsitz, S.R., Catalano, P.J., Zahner, G.E.P. & Fitzmaurice, G.M. (1999). Goodness-of-fit for GEE: an example with mental health service utilization, *Statistics in Medicine* **18**, 213–222.
- [9] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 221–233.
- [10] Ibrahim, J.G. & Lipsitz, S.R. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, *Journal of the Royal Statistical Society – Series B* **61**(1), 173–190.

-
- [11] Kauermann, G. & Carroll, R.J. (2001). The sandwich variance estimator: efficiency properties and coverage probability of confidence intervals, *Journal of the American Statistical Association* **96**, 1386–1397.
 - [12] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
 - [13] Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated measures studies, *Journal of the American Statistical Association* **90**, 1112–1121.
 - [14] McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*, 2nd Edition, Chapman & Hall, London.
 - [15] Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**, 221–232.
 - [16] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society – Series A* **135**(3), 370–384.
 - [17] Pan, W. (2001a). Akaike’s information criterion in generalized estimating equations, *Biometrics* **57**, 120–125.
 - [18] Pan, W. (2001b). Model selection in estimating equations, *Biometrics* **57**, 529–534.
 - [19] Preisser, J.S. & Qaqish, B.F. (1999). Robust regression for clustered data with application to binary responses, *Biometrics* **55**, 574–579.
 - [20] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *Statistical Journal* **2**, 1–21.
 - [21] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**(427), 846–866.
 - [22] Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**(429), 106–121.
 - [23] Rotnitzky, A. & Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data, *Biometrika* **77**(3), 485–497.
 - [24] Thall, P.F. & Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion, *Biometrics* **46**, 657–671.
 - [25] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.
 - [26] Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data, *Statistics in Medicine* **19**, 1265–1275.

JAMES W. HARDIN

Generalized Linear Mixed Models

DONALD HEDEKER

Volume 2, pp. 729–738

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalized Linear Mixed Models

Introduction

Generalized linear models (GLMs) represent a class of fixed effects regression models for several types of dependent variables (i.e., continuous, dichotomous, counts). McCullagh and Nelder [32] describe these in great detail and indicate that the term ‘generalized linear model’ is due to Nelder and Wedderburn [35] who described how a collection of seemingly disparate statistical techniques could be unified. Common Generalized linear models (GLMs) include linear regression, **logistic regression**, and Poisson regression.

There are three specifications in a GLM. First, the linear predictor, denoted as η_i , of a GLM is of the form

$$\eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \quad (1)$$

where $\mathbf{x}_i$ is the vector of regressors for unit i with fixed effects $\boldsymbol{\beta}$. Then, a link function $g(\cdot)$ is specified which converts the expected value μ_i of the outcome variable Y_i (i.e., $\mu_i = E[Y_i]$) to the linear predictor η_i

$$g(\mu_i) = \eta_i. \quad (2)$$

Finally, a specification for the form of the variance in terms of the mean μ_i is made. The latter two specifications usually depend on the distribution of the outcome Y_i , which is assumed to fall within the exponential family of distributions.

Fixed effects models, which assume that all observations are independent of each other, are not appropriate for analysis of several types of correlated data structures, in particular, for clustered and/or longitudinal data (see **Clustered Data**). In clustered designs, subjects are observed nested within larger units, for example, schools, hospitals, neighborhoods, workplaces, and so on. In longitudinal designs, repeated observations are nested within subjects (see **Longitudinal Data Analysis; Repeated Measures Analysis of Variance**). These are often referred to as multi-level [16] or hierarchical [41] data (see **Linear Multilevel Models**), in which the level-1 observations (subjects or repeated observations) are nested within the higher level-2 observations (clusters or subjects).

Higher levels are also possible, for example, a three-level design could have repeated observations (level-1) nested within subjects (level-2) who are nested within clusters (level-3).

For analysis of such multilevel data, random cluster and/or subject effects can be added into the regression model to account for the correlation of the data. The resulting model is a mixed model including the usual fixed effects for the regressors plus the random effects. Mixed models for continuous normal outcomes have been extensively developed since the seminal paper by Laird and Ware [28]. For nonnormal data, there have also been many developments, some of which are described below. Many of these developments fall under the rubric of generalized linear mixed models (GLMMs), which extend GLMs by the inclusion of random effects in the predictor. Agresti et al. [1] describe a variety of social science applications of GLMMs; [12], [33], and [11] are recent texts with a wealth of statistical material on GLMMs.

Let i denote the level-2 units (e.g., subjects) and let j denote the level-1 units (e.g., nested observations). The focus will be on longitudinal designs here, but the methods apply to clustered designs as well. Assume there are $i = 1, \dots, N$ subjects (level-2 units) and $j = 1, \dots, n_i$ repeated observations (level-1 units) nested within each subject. A random-intercept model, which is the simplest mixed model, augments the linear predictor with a single random effect for subject i ,

$$\eta_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i, \quad (3)$$

where v_i is the random effect (one for each subject). These random effects represent the influence of subject i on his/her repeated observations that is not captured by the observed covariates. These are treated as random effects because the sampled subjects are thought to represent a population of subjects, and they are usually assumed to be distributed as $\mathcal{N}(0, \sigma_v^2)$. The parameter σ_v^2 indicates the variance in the population distribution, and therefore the degree of heterogeneity of subjects.

Including the random effects, the expected value of the outcome variable, which is related to the linear predictor via the link function, is given as

$$\mu_{ij} = E[Y_{ij} | v_i, \mathbf{x}_{ij}]. \quad (4)$$

This is the expectation of the conditional distribution of the outcome given the random effects. As a

2 Generalized Linear Mixed Models

result, GLMMs are often referred to as conditional models in contrast to the marginal generalized estimating equations (GEE) models (*see Generalized Estimating Equations (GEE)*) [29], which represent an alternative generalization of GLMs for correlated data (*see Marginal Models for Clustered Data*).

The model can be easily extended to include multiple random effects. For example, in longitudinal problems, it is common to have a random subject intercept and a random linear time-trend. For this, denote $\mathbf{z}_{ij}$ as the $r \times 1$ vector of variables having random effects (a column of ones is usually included for the random intercept). The vector of random effects $\mathbf{v}_i$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and variance–covariance matrix Σ_v (*see Catalogue of Probability Density Functions*). The model is now written as

$$\eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{v}_i. \quad (5)$$

Note that the conditional mean μ_{ij} is now specified as $E[Y_{ij}|\mathbf{v}_i, \mathbf{x}_{ij}]$, namely, in terms of the vector of random effects.

Dichotomous Outcomes

Development of GLMMs for dichotomous data has been an active area of statistical research. Several approaches, usually adopting a logistic or probit regression model (*see Probits*) and various methods for incorporating and estimating the influence of the random effects, have been developed. A review article by Pendergast et al. [37] discusses and compares many of these developments.

The mixed-effects logistic regression model is a common choice for analysis of multilevel dichotomous data and is arguably the most popular GLMM. In the GLMM context, this model utilizes the logit link, namely

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \log \left[\frac{\mu_{ij}}{1 - \mu_{ij}} \right] = \eta_{ij}. \quad (6)$$

Here, the conditional expectation $\mu_{ij} = E(Y_{ij}|\mathbf{v}_i, \mathbf{x}_{ij})$ equals $P(Y_{ij} = 1|\mathbf{v}_i, \mathbf{x}_{ij})$, namely, the conditional probability of a response given the random effects (and covariate values).

This model can also be written as

$$P(Y_{ij} = 1|\mathbf{v}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = g^{-1}(\eta_{ij}) = \Psi(\eta_{ij}), \quad (7)$$

where the inverse link function $\Psi(\eta_{ij})$ is the logistic cumulative distribution function (cdf), namely $\Psi(\eta_{ij}) = [1 + \exp(-\eta_{ij})]^{-1}$. A nicety of the logistic distribution, that simplifies parameter estimation, is that the probability density function (pdf) is related to the cdf in a simple way, as $\psi(\eta_{ij}) = \Psi(\eta_{ij})[1 - \Psi(\eta_{ij})]$.

The probit model, which is based on the standard normal distribution, is often proposed as an alternative to the logistic model [13]. For the probit model, the normal cdf and pdf replace their logistic counterparts. A useful feature of the probit model is that it can be used to yield **tetrachoric correlations** for the clustered binary responses, and **polychoric correlations** for ordinal outcomes (discussed below). For this reason, in some areas, for example familial studies, the probit formulation is often preferred to its logistic counterpart.

Example

Gruder et al. [20] describe a smoking-cessation study in which 489 subjects were randomized to either a control, discussion, or social support conditions. Control subjects received a self-help manual and were encouraged to watch twenty segments of a daily TV program on smoking cessation, while subjects in the two experimental conditions additionally participated in group meetings and received training in support and relapse prevention. Here, for simplicity, these two experimental conditions will be combined. Data were collected at four telephone interviews: postintervention, and 6, 12, and 24 months later. Smoking abstinence rates (and sample sizes) at these four timepoints were 17.4% (109), 7.2% (97), 18.5% (92), and 18.2% (77) for the placebo condition. Similarly, for the combined experimental condition it was 34.5% (380), 18.2% (357), 19.6% (337), and 21.7% (295) for these timepoints.

Two logistic GLMM were fit to these data: a random intercept and a random intercept and linear trend of time model (*see Growth Curve Modeling*). These models were estimated using SAS PROC NLMIXED with adaptive quadrature. For these, it is the probability of smoking abstinence, rather than smoking, that is being modeled. Fixed effects included a condition term (0 = control, 1 = experimental), time (coded 0, 1, 2, and 4 for the four timepoints), and the condition by time interaction. Results for both models are presented in Table 1. Based on a likelihood-ratio

Table 1 Smoking cessation study: smoking status (0 = smoking, 1 = not smoking) across time (N = 489), GLMM logistic parameter estimates (Est.), standard errors (SE), and *P* values

Parameter	Random intercept model			Random int and trend model		
	Est.	SE	<i>P</i> value	Est.	SE	<i>P</i> value
Intercept	−2.867	.362	.001	−2.807	.432	.001
Time	.113	.122	.36	−.502	.274	.07
Condition (0 = control; 1 = experimental)	1.399	.379	.001	1.495	.415	.001
Condition by Time	−.322	.136	.02	−.331	.249	.184
Intercept variance	3.587	.600		3.979	1.233	
Intercept Time covariance				.048	.371	
Time variance				1.428	.468	
−2 log likelihood		1631.0			1594.7	

Note: *P* values not given for variance and covariance parameters (see [41]).

test, the model with random intercept and linear time trend is preferred over the simpler random intercept model ($\chi^2_2 = 36.3$). Thus, there is considerable evidence for subjects varying in both their intercepts and time trends. It should be noted that the test statistic does not have a chi-square distribution when testing variance parameters because the null hypothesis is on the border of the parameter space, making the *P* value conservative. Snijders and Bosker [46] elaborate on this issue and point out that a simple remedy, that has been shown to be reasonable in simulation studies, is to divide the *P* value based on the likelihood-ratio chi-square test statistic by two. In the present case, it doesn't matter because the *P* value is $<.001$ for $\chi^2_2 = 36.3$ even without dividing by two.

In terms of the fixed effects, both models indicate a nonsignificant time effect for the control condition, and a highly significant condition effect at time 0 (e.g., $z = 1.495/.415 = 3.6$ in the second model). This indicates a positive effect of the experimental conditions on smoking abstinence relative to control at postintervention. There is also some evidence of a negative condition by time interaction, suggesting that the beneficial condition effect diminishes across time. Note that this interaction is not significant ($P < .18$) in the random intercept and trend model, but it is significant in the random intercept model ($P < .02$). Since the former is preferred by the likelihood-ratio test, we would conclude that the interaction is not significant.

This example shows that the significance of model terms can depend on the structure of the random effects. Thus, one must decide upon a reasonable model for the random effects as well as for the fixed effects. A commonly recommended approach

for this is to perform a sequential procedure for model selection. First, one includes all possible covariates of interest into the model and selects between the possible models of random effects using likelihood-ratio tests and model fit criteria. Then, once a reasonable random effects structure is selected, one trims model covariates in the usual way.

IRT Models

Because the logistic model is based on the logistic response function, and the random effects are assumed normally distributed, this model and models closely related to it are often referred to as logistic/normal models, especially in the latent trait model literature [4]. Similarly, the probit model is sometimes referred to as a normal/normal model. In many respects, latent trait or **item response theory** (IRT) models, developed in the educational testing and psychometric literatures, represent some of the earliest GLMMs. Here, item responses ($j = 1, 2, \dots, n$) are nested within subjects ($i = 1, 2, \dots, N$). The simplest IRT model is the **Rasch model** [40] which posits the probability of a correct response to the dichotomous item j ($Y_{ij} = 1$) conditional on the random effect or 'ability' of subject i (θ_i) in terms of the logistic cdf as

$$P(Y_{ij} = 1|\theta_i) = \Psi(\theta_i - b_j), \quad (8)$$

where b_j is the threshold or difficulty parameter for item j (i.e., item difficulty). Subject's ability is commonly denoted as θ in the IRT literature (i.e., instead of ν). Note that the Rasch model is simply a random-intercepts model that includes

4 Generalized Linear Mixed Models

item dummies as fixed regressors. Because there is only one parameter per item, the Rasch model is also called the *one-parameter IRT model*. A more general IRT model, the two-parameter model [5], also includes a parameter for the discrimination of the item in terms of ability.

Though IRT models were not originally cast as GLMMs, formulating them in this way easily allows covariates to enter the model at either level (i.e., items or subjects). This and other advantages of casting IRT models as mixed models are described by Rijmen et al. [43], who provide a comprehensive overview and bridge between IRT models, mixed models, and GLMMs. As they point out, the Rasch model, and variants of it, belong to the class of GLMMs. However, the more extended two-parameter model is not within the class of GLMMs because the predictor is no longer linear, but includes a product of parameters.

Ordinal Outcomes

Extending the methods for dichotomous responses to ordinal response data has also been actively pursued; Agresti and Natarajan [2] review many of these developments. Because the proportional odds model described by McCullagh [31], which is based on the logistic regression formulation, is a common choice for analysis of ordinal data, many of the GLMMs for ordinal data are generalizations of this model, though models relaxing this assumption have also been described [27]. The proportional odds model expresses the ordinal responses in C categories ($c = 1, 2, \dots, C$) in terms of $C - 1$ cumulative category comparisons, specifically, $C - 1$ cumulative logits (i.e., log odds). Here, denote the conditional cumulative probabilities for the C categories of the outcome Y_{ij} as $P_{ijc} = P(Y_{ij} \leq c | \mathbf{v}_i, \mathbf{x}_{ij}) = \sum_{c=1}^C p_{ijc}$, where p_{ijc} represents the conditional probability of response in category c . The logistic GLMM for the conditional cumulative probabilities $\mu_{ijc} = P_{ijc}$ is given in terms of the cumulative logits as

$$\log \left[\frac{\mu_{ijc}}{1 - \mu_{ijc}} \right] = \eta_{ijc} \quad (c = 1, \dots, C - 1), \quad (9)$$

where the linear predictor is now

$$\eta_{ijc} = \gamma_c - [\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{v}_i], \quad (10)$$

with $C - 1$ strictly increasing model thresholds γ_c (i.e., $\gamma_1 < \gamma_2 < \dots < \gamma_{C-1}$). The thresholds allow the cumulative response probabilities to differ. For identification, either the first threshold γ_1 or the model intercept β_0 is typically set to zero. As the regression coefficients $\boldsymbol{\beta}$ do not carry the c subscript, the effects of the regressors do not vary across categories. McCullagh [31] calls this assumption of identical odds ratios across the $C - 1$ cutoffs the proportional odds assumption.

Because the ordinal model is defined in terms of the cumulative probabilities, the conditional probability of a response in category c is obtained as the difference of two conditional cumulative probabilities:

$$P(Y_{ij} = c | \mathbf{v}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \Psi(\eta_{ijc}) - \Psi(\eta_{ij,c-1}). \quad (11)$$

Here, $\gamma_0 = -\infty$ and $\gamma_C = \infty$, and so $\Psi(\eta_{ij0}) = 0$ and $\Psi(\eta_{ijC}) = 1$ (see **Ordinal Regression Models**).

Example

Hedeker and Gibbons [25] described a random-effects ordinal probit regression model, examining longitudinal data collected in the NIMH Schizophrenia Collaborative Study on treatment related changes in overall severity. The dependent variable was item 79 of the Inpatient Multidimensional Psychiatric Scale (IMPS; [30]), scored as: (a) normal or borderline mentally ill, (b) mildly or moderately ill, (c) markedly ill, and (d) severely or among the most extremely ill. In this study, patients were randomly assigned to receive one of four medications: placebo, chlorpromazine, fluphenazine, or thioridazine. Since previous analyses revealed similar effects for the three antipsychotic drug groups, they were combined in the analysis. The experimental design and corresponding sample sizes are listed in Table 2.

As can be seen from Table 2, most of the measurement occurred at weeks 0, 1, 3, and 6, with some scattered measurements at the remaining timepoints.

Table 2 Experimental design and weekly sample sizes

Group	Sample size at week						
	0	1	2	3	4	5	6
Placebo (n = 108)	107	105	5	87	2	2	70
Drug (n = 329)	327	321	9	287	9	7	265

Note: Drug = Chlorpromazine, Fluphenazine, or Thioridazine.

Table 3 NIMH Schizophrenia Collaborative Study: severity of illness (IMPS79) across time ($N = 437$), GLMM logistic parameter estimates (Est.), standard errors (SE), and P values

Parameter	Est.	SE	P value
Intercept	7.283	.467	.001
Time (sqrt week)	-.879	.216	.001
Drug (0 = placebo; 1 = drug)	.056	.388	.88
Drug by Time	-1.684	.250	.001
Threshold 2	3.884	.209	.001
Threshold 3	6.478	.290	.001
Intercept variance	6.847	1.282	
Intercept-time covariance	-1.447	.515	
Time variance	1.949	.404	
-2 log likelihood		3326.5	

Note: Threshold 1 set to zero for identification. P values not given for variance and covariance parameters (see [41]). NIMH = National Institute of Mental Health; IMPS79 = Inpatient Multidimensional Psychiatric Scale, Item 79.

Here, a logistic GLMM with random intercept and trend was fit to these data using SAS PROC NLMIXED with adaptive quadrature. Fixed effects included a dummy-coded drug effect (placebo = 0 and drug = 1), a time effect (square root of week; this was used to linearize the relationship between the cumulative logits and week) and a drug by time interaction. Results from this analyses are given in Table 3.

The results indicate that the treatment groups do not significantly differ at baseline (drug effect), the placebo group does improve over time (significant negative time effect), and the drug group has greater improvement over time relative to the placebo group (significant negative drug by time interaction). Thus, the analysis supports use of the drug, relative to placebo, in the treatment of schizophrenia.

Comparing this model to a simpler random-intercepts model (not shown) yields clear evidence of significant variation in both the individual intercept and time-trends (likelihood-ratio $\chi^2_2 = 77.7$). Also, a moderate negative association between the intercept and linear time terms is indicated, expressed as a correlation it equals $-.40$, suggesting that those patients with the highest initial severity show the greatest improvement across time (e.g., largest negative time-trends). This latter finding could be a result of a ‘floor effect’, in that patients with low initial severity scores cannot exhibit large negative time-trends due to the limited range in the ordinal outcome variable. Finally, comparing this model to one that allows

nonproportional odds for all model covariates (not shown) supports the proportional odds assumption ($\chi^2_6 = 3.63$). Thus, the three covariates (drug, time, and drug by time) have similar effects on the three cumulative logits.

Survival Analysis Models

Connections between ordinal regression and survival analysis models (see **Survival Analysis**) have led to developments of discrete and grouped-time survival analysis GLMMs [49]. The basic notion is that the time to the event can be considered as an ordinal variable with C possible event times, albeit with right-censoring accommodated. Vermunt [50] also describes related log-linear mixed models for survival analysis or **event history analysis**.

Nominal Outcomes

Nominal responses occur when the categories of the response variable are not ordered. General regression models for multilevel nominal data have been considered, and Hartzel et al. [22] synthesizes much of the work in this area, describing a general mixed-effects model for both clustered ordinal and nominal responses.

In the nominal GLMM, the probability that $Y_{ij} = c$ (a response occurs in category c) for a given individual i , conditional on the random effects $\mathbf{v}$, is given by:

$$p_{ijc} = P(Y_{ij} = c | \mathbf{v}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \frac{\exp(\eta_{ijc})}{1 + \sum_{h=1}^C \exp(\eta_{ijh})} \text{ for } c = 2, 3, \dots, C, \quad (12)$$

$$p_{ij1} = P(Y_{ij} = 1 | \mathbf{v}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \frac{1}{1 + \sum_{h=1}^C \exp(\eta_{ijh})}, \quad (13)$$

with the linear predictor $\eta_{ijc} = \mathbf{x}_{ij}'\boldsymbol{\beta}_c + \mathbf{z}_{ij}'\mathbf{v}_{ic}$. Both the regression coefficients $\boldsymbol{\beta}_c$ and the random-effects carry the c subscript; the latter allows the variance-covariance matrix $\boldsymbol{\Sigma}_{v_c}$ to vary across categories. In the model above, these parameters represent differences relative to the first category. The nominal model can also be written to allow for any possible set of $C - 1$ contrasts, see [24] for an example of this.

Ranks

In ranking data, individuals are asked to rank C distinct options with respect to some criterion. If the individuals are only asked to provide the option with the highest (or lowest) rank of the C categories, then the resulting data consist of either an ordinal outcome (if the C options are ordered) or a nominal outcome (if the C options are not ordered), and analysis can proceed using the models described above. In the more general case, individuals are asked for, say, the top three options, or to fully rank the C options from the ‘best’ to the ‘worst’ (i.e., all options receive a rank from 1 to C). The former case consists of partial ranking data, while the latter case represents full ranking data. As these data types are generalizations of nominal and ordinal data types, it is not surprising that statistical models for ranking data are generalizations of the models for ordinal and nominal models described above. In particular, since the C options are usually not ordered options, models for ranking data have close connections with models for nominal outcomes. GLMMs for ranking data are described in [6] and [45]. These articles show the connections between models for multilevel nominal and ranking data, as well as develop several extensions for the latter.

Counts

For count data, various types of Poisson mixed models have been proposed. A review of some of these methods applied to longitudinal Poisson data is given in [47]. For computational purposes, it is convenient for the univariate random effects to have a gamma distribution in the population of subjects [3]. However, as described in [11], adding multiple normally distributed random effects on the same scale as the fixed effects of the Poisson regression model provides a more general and flexible model.

Let Y_{ij} be the value of the count variable (where Y_{ij} can equal $0, 1, \dots$) associated with individual i and timepoint j . If this count is assumed to be drawn from a Poisson distribution, then the mixed Poisson regression model indicates the expected number of counts as

$$\log \mu_{ij} = \eta_{ij}, \quad (14)$$

with the linear predictor $\eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{v}_i$. In some cases the size of the time interval over which the events are counted varies. For example, McKnight and Van Den Eeden [34] describe a study in which the number of headaches in a week is recorded, however, not all individuals are measured for all seven days. For this, let t_{ij} represent the follow-up time associated with units i and j . The linear predictor is now augmented as

$$\eta_{ij} = \log t_{ij} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{v}_i, \quad (15)$$

which can also be expressed as

$$\mu_{ij} = t_{ij} \exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{v}_i) \quad (16)$$

or $\mu_{ij}/t_{ij} = \exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{v}_i)$ to reflect that it is the number of counts per follow-up period that is being modeled. The term $\log t_{ij}$ is often called an *offset*.

Assuming the Poisson process for the count Y_{ij} , the probability that $Y_{ij} = y$, conditional on the random effects $\mathbf{v}$, is given as

$$P(Y_{ij} = y | \mathbf{v}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \exp(-\mu_{ij}) \frac{(\mu_{ij})^y}{y!}. \quad (17)$$

It is often the case that count data exhibit more zero counts than what is consistent with the Poisson distribution. For such situations, zero-inflated Poisson (ZIP) mixed models, which contain a logistic (or probit) regression for the probability of a nonzero response and a Poisson regression for the zero

and nonzero counts, have been developed [21]. A somewhat related model is described by Olsen and Schafer [36] who propose a two-part model that includes a logistic model for the probability of a nonzero response and a conditional linear model for the mean response given that it is nonzero.

Estimation

Parameter estimation in GLMMs typically involves **maximum likelihood** (ML) or variants of ML. Additionally, the solutions are usually iterative ones that can be numerically quite intensive. Here, the solution is merely sketched; further details can be found in [33] and [12].

For the models presented, (7), (11), (12)–(13), and (17), indicate the probability of a level-1 response Y_{ij} for a given subject i at timepoint j , conditional on the random effects $\mathbf{v}_i$. While the form of this probability depends on the form of the response variable, let $P(Y_{ij}|\mathbf{v}_i)$ represent the conditional probability for any of these forms. Here, for simplicity, we omit conditioning on the covariates $\mathbf{x}_{ij}$. Let $\mathbf{Y}_i$ denote the vector of responses from subject i . The probability of any response pattern $\mathbf{Y}_i$ (of size n_i), conditional on $\mathbf{v}_i$, is equal to the product of the probabilities of the level-1 responses:

$$\ell(\mathbf{Y}_i|\mathbf{v}_i) = \prod_{i=1}^{n_i} P(Y_{ij}|\mathbf{v}_i). \quad (18)$$

The assumption that a subject's responses are independent given the random effects (and therefore can be multiplied to yield the conditional probability of the response vector) is known as the **conditional independence** assumption. The marginal density of $\mathbf{Y}_i$ in the population is expressed as the following integral of the conditional likelihood $\ell(\cdot)$

$$h(\mathbf{Y}_i) = \int_{\mathbf{v}_i} \ell(\mathbf{Y}_i|\mathbf{v}_i) f(\mathbf{v}_i) d\mathbf{v}_i, \quad (19)$$

where $f(\mathbf{v}_i)$ represents the distribution of the random effects, often assumed to be a multivariate normal density. Whereas (18) represents the conditional probability, (19) indicates the unconditional probability for the response vector of subject i . The marginal log-likelihood from the sample of N subjects is then obtained as $\log L = \sum_i^N \log h(\mathbf{Y}_i)$. Maximizing this

log-likelihood yields ML estimates (which are sometimes referred to as maximum marginal likelihood estimates) of the regression coefficients $\boldsymbol{\beta}$ and the variance-covariance matrix of the random effects $\boldsymbol{\Sigma}_{\mathbf{v}_i}$.

Integration over the random-effects distribution

In order to solve the likelihood solution, integration over the random-effects distribution must be performed. As a result, estimation is much more complicated than in models for continuous normally distributed outcomes where the solution can be expressed in closed form. Various approximations for evaluating the integral over the random-effects distribution have been proposed in the literature; many of these are reviewed in [44]. Perhaps the most frequently used methods are based on first- or second-order Taylor expansions. Marginal quasi-likelihood (MQL) involves expansion around the fixed part of the model, whereas penalized or predictive quasi-likelihood (PQL) additionally includes the random part in its expansion [17]. Unfortunately, these procedures yield estimates of the regression coefficients and random effects variances that are biased towards zero in certain situations, especially for the first-order expansions [7].

More recently, Raudenbush et al. [42] proposed an approach that uses a combination of a fully multivariate Taylor expansion and a Laplace approximation. This method yields accurate results and is computationally fast. Also, as opposed to the MQL and PQL approximations, the deviance obtained from this approximation can be used for likelihood-ratio tests.

Numerical integration can also be used to perform the integration over the random-effects distribution. Specifically, if the assumed distribution is normal, Gauss–Hermite quadrature can approximate the above integral to any practical degree of accuracy. Additionally, like the Laplace approximation, the numerical quadrature approach yields a deviance that can be readily used for likelihood-ratio tests. The integration is approximated by a summation on a specified number of quadrature points for each dimension of the integration. An issue with the quadrature approach is that it can involve summation over a large number of points, especially as the number of random-effects is increased. To address this, methods of adaptive quadrature have been developed which use a few points per dimension that are adapted

to the location and dispersion of the distribution to be integrated [39].

More computer-intensive methods, involving iterative simulations, can also be used to approximate the integration over the random effects distribution. Such methods fall under the rubric of **Markov chain Monte Carlo** (MCMC; [15]) algorithms. Use of MCMC for estimation of a wide variety of models has exploded in the last 10 years or so; MCMC solutions for GLMMs are described in [9].

Estimation of random effects

In many cases, it is useful to obtain estimates of the random effects. The random effects v_i can be estimated using empirical Bayes methods (*see Random Effects in Multivariate Linear Models: Prediction*). For the univariate case, this estimator $\hat{v}_i$ is given by:

$$\hat{v}_i = E(v_i | Y_i) = h_i^{-1} \int_{v_i} v_i \ell_i f(v_i) dv_i \quad (20)$$

where ℓ_i is the conditional probability for subject i under the particular model and h_i is the analogous marginal probability. This is simply the mean of the posterior distribution. Similarly, the variance of the posterior distribution is obtained as

$$V(\hat{v}_i | Y_i) = h_i^{-1} \int_{v_i} (v_i - \hat{v}_i)^2 \ell_i f(v_i) dv_i. \quad (21)$$

These quantities may then be used, for example, to evaluate the response probabilities for particular subjects (e.g., person-specific trend estimates). Also, Ten Have [48] suggests how these empirical Bayes estimates can be used in performing residual diagnostics.

Discussion

Though the focus here has been on two-level GLMMs for nonnormal data, three-level (and higher) generalizations have also been considered in the literature [14]. Also, software for fitting GLMMs is readily available in the major statistical packages (i.e., SAS PROC NL MIXED, STATA) and in several independent programs (HLM, [8]; EGRET, [10]; MLwiN, [18]; LIMDEP, [19]; MIXOR, [26]; MIXNO, [23]; GLLAMM, [38]). Not all of these programs fit all of the GLMMs described here; some only allow

random-intercepts models or two-level models, for example, and several vary in terms of how the integration over the random effects is performed. However, though the availability of these software programs is relatively recent, they have definitely facilitated application of GLMMs in psychology and elsewhere. The continued development of these models and their software implementations should only lead to greater use and understanding of GLMMs for analysis of correlated nonnormal data.

Acknowledgments

Thanks are due to Dr. Robin Mermelstein for use of the smoking-cessation study data, and to Drs. Nina Schooler and John Davis for use of the schizophrenia study data. This work was supported by National Institutes of Mental Health Grant MH56146.

References

- [1] Agresti, A., Booth, J.G., Hobart, J.P. & Caffo, B. (2000). Random-effects modeling of categorical response data, *Sociological Methodology* **30**, 27–80.
- [2] Agresti, A. & Natarajan, R. (2001). Modeling clustered ordered categorical data: a survey, *International Statistical Review* **69**, 345–371.
- [3] Albert, J. (1992). A Bayesian analysis of a Poisson random effects model for home run hitters, *The American Statistician* **46**, 246–253.
- [4] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd Edition, Oxford University Press, New York.
- [5] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading.
- [6] Böckenholt, U. (2001). Mixed-effects analyses of rank-ordered data, *Psychometrika* **66**, 45–62.
- [7] Breslow, N.E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika* **82**, 81–91.
- [8] Bryk, A.S., Raudenbush, S.W. & Congdon, R. (2000). *HLM Version 5*, Scientific Software International, Chicago.
- [9] Clayton, D. (1996). Generalized linear mixed models, in *Markov Chain Monte Carlo Methods in Practice*, W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds, Chapman & Hall, New York, pp. 275–303.
- [10] Corcoran, C., Coull, B. & Patel, A. (1999). *EGRET for Windows User Manual*, CYTEL Software Corporation, Cambridge.
- [11] Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Oxford University Press, New York.

-
- [12] Fahrmeir, L. & Tutz, G.T. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd Edition, Springer-Verlag, New York.
- [13] Gibbons, R.D. & Bock, R.D. (1987). Trend in correlated proportions, *Psychometrika* **52**, 113–124.
- [14] Gibbons, R.D. & Hedeker, D. (1997). Random-effects probit and logistic regression models for three-level data, *Biometrics* **53**, 1527–1537.
- [15] Gilks, W., Richardson, S. & Spiegelhalter, D.J. (1997). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, New York.
- [16] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Edition, Halstead Press, New York.
- [17] Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series B* **159**, 505–513.
- [18] Goldstein, H. Rasbash, J. Plewis, I. Draper, D. Browne, W. & Wang, M. (1998). *A User's Guide to MLwiN*, University of London, Institute of Education, London.
- [19] Greene, W.H. (1998). *LIMDEP Version 7.0 User's Manual*, (revised edition), Econometric Software, Plainview.
- [20] Gruder, C.L., Mermelstein, R.J., Kirkendol, S., Hedeker, D., Wong, S.C., Schreckengost, J., Warnecke, R.B., Burzette, R. & Miller, T.Q. (1993). Effects of social support and relapse prevention training as adjuncts to a televised smoking cessation intervention, *Journal of Consulting and Clinical Psychology* **61**, 113–120.
- [21] Hall, D.B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics* **56**, 1030–1039.
- [22] Hartzel, J., Agresti, A. & Caffo, B. (2001). Multinomial logit random effects models, *Statistical Modelling* **1**, 81–102.
- [23] Hedeker, D. (1999). MIXNO: a computer program for mixed-effects nominal logistic regression, *Journal of Statistical Software* **4**(5), 1–92.
- [24] Hedeker, D. (2003). A mixed-effects multinomial logistic regression model, *Statistics in Medicine*, **22** 1433–1446.
- [25] Hedeker, D. & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis, *Biometrics* **50**, 933–944.
- [26] Hedeker, D. & Gibbons, R.D. (1996). MIXOR: a computer program for mixed-effects ordinal probit and logistic regression analysis, *Computer Methods and Programs in Biomedicine* **49**, 157–176.
- [27] Hedeker, D. & Mermelstein, R.J. (1998). A multilevel thresholds of change model for analysis of stages of change data, *Multivariate Behavioral Research* **33**, 427–455.
- [28] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.
- [29] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.
- [30] Lorr, M. & Klett, C.J. (1966). *Inpatient Multidimensional Psychiatric Scale: Manual*, Consulting Psychologists Press, Palo Alto.
- [31] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.
- [32] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, New York.
- [33] McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*, Wiley, New York.
- [34] McKnight, B. & Van Den Eeden, S.K. (1993). A conditional analysis for two-treatment multiple period crossover designs with binomial or Poisson outcomes and subjects who drop out, *Statistics in Medicine* **12**, 825–834.
- [35] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
- [36] Olsen, M.K. & Schafer, J.L. (2001). A two-part random effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association* **96**, 730–745.
- [37] Pendergast, J.F., Gange, S.J., Newton, M.A., Lindstrom, M.J., Palta, M. & Fisher, M.R. (1996). A survey of methods for analyzing clustered binary response data, *International Statistical Review* **64**, 89–118.
- [38] Rabe-Hesketh, S. Pickles, A. & Skrondal, A. (2001). GLLAMM Manual, Technical Report 2001/01, Institute of Psychiatry, King's College, University of London, Department of Biostatistics and Computing.
- [39] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal* **2**, 1–21.
- [40] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute of Educational Research, Copenhagen.
- [41] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data-Analysis Methods*, 2nd Edition, Sage Publications, Thousand Oaks.
- [42] Raudenbush, S.W., Yang, M.-L. & Yosef, M. (2000). Maximum likelihood for generalized linear mixed models with nested random effects via high-order, multivariate Laplace approximation, *Journal of Computational and Graphical Statistics* **9**, 141–157.
- [43] Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory, *Psychological Methods* **8**, 185–205.
- [44] Rodríguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses, *Journal of the Royal Statistical Society, Series A* **158**, 73–89.

- [45] Skrondal, A. & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings, *Psychometrika* **68**, 267–287.
- [46] Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, Thousand Oaks.
- [47] Stukel, T.A. (1993). Comparison of methods for the analysis of longitudinal interval count data, *Statistics in Medicine* **12**, 1339–1351.
- [48] Ten Have, T.R. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses, *Biometrics* **52**, 473–491.
- [49] Ten Have, T.R. & Uttal, D.H. (1994). Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles, *Applied Statistics* **43**, 371–384.
- [50] Vermunt, J.K. (1997). *Log-linear Models for Event Histories*, Sage Publications, Thousand Oaks.

DONALD HEDEKER

Generalized Linear Models (GLM)

BRIAN S. EVERITT

Volume 2, pp. 739–743

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Generalized Linear Models (GLM)

The generalized linear model (GLM) is essentially a unified framework for regression models introduced in a landmark paper by Nelder and Wedderburn [7] over 30 years ago. A wide range of statistical models including **analysis of variance**, **analysis of covariance**, **multiple linear regression**, and **logistic regression** are included in the GLM framework. A comprehensive technical account of the model is given in [6], with a more concise description appearing in [2] and [1].

Regression

The term ‘regression’ was first introduced by **Francis Galton** in the nineteenth century to characterize a tendency to mediocrity, that is, towards the average, observed in the offspring of parent seeds, and used by **Karl Pearson** in a study of the heights of fathers and sons. The sons’ heights tended, on average, to be less extreme than the fathers (*see Regression to the Mean*). In essence, all forms of regression have as their aim the development and assessment of a mathematical model for the relationship between a response variable, y , and a set of q explanatory variables (sometimes confusingly referred to as independent variables), $x_1, x_2, \dots, x_q$. Multiple linear regression, for example, involves the following model for y :

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q + \varepsilon, \quad (1)$$

where $\beta_0, \beta_1, \dots, \beta_q$ are regression coefficients that have to be estimated from sample data and ε is an error term assumed normally distributed with zero mean and variance σ^2 .

An equivalent way of writing the multiple regression model is:

$$y \sim N(\mu, \sigma^2),$$

where $\mu = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q$. This makes it clear that this model is only suitable for continuous response variables with, conditional on the values of the explanatory variables, a normal distribution with constant variance. Analysis of variance is essentially exactly the same model, with $x_1, x_2, \dots, x_q$ being

dummy variables coding factor levels and interactions between factors; analysis of covariance is also the same model with a mixture of continuous and categorical explanatory variables. (The equivalence of multiple regression to analysis of variance and so on is sometimes referred to as the *general linear model* – see, for example [3]).

The assumption of the conditional normality of a continuous response variable is one that is probably made more often than it is warranted. And there are many situations where such an assumption is clearly not justified. One example is where the response is a binary variable (e.g., improved, not improved), another is where it is a count (e.g., number of correct answers in some testing situation). The question then arises as to how the multiple regression model can be modified to allow such responses to be related to the explanatory variables of interest. In the GLM approach, the generalization of the multiple regression model consists of allowing the following three assumptions associated with this model to be modified.

- The response variable is normally distributed with a mean determined by the model.
- The mean can be modeled as a linear function of (possibly nonlinear transformations) the explanatory variables, that is, the effects of the explanatory variable on the mean are additive.
- The variance of the response variable given the (predicted) mean is constant.

In a GLM, some transformation of the mean is modeled by a linear function of the explanatory variables, and the distribution of the response around its mean (often referred to as the *error distribution*) is generalized usually in a way that fits naturally with a particular transformation. The result is a very wide class of regression models, but before detailing the unifying features of GLMs, it will be helpful to look at how a particular type of model, logistic regression, fits into the general framework.

Logistic Regression

Logistic regression is a technique widely used to study the relationship between a binary response and a set of explanatory variables. The expected value (μ) of a binary response is simply the probability, π , that the response variable takes the value one (usually

2 Generalized Linear Models (GLM)

used as the coding for the occurrence of the event of interest, say ‘improved’). Modeling this expected value directly as a linear function of explanatory variables, as is done in multiple linear regression, is now clearly not sensible since it could result in fitted values of the response variable outside the range (0, 1). And, in addition, the error distribution of the response, given the explanatory variables, will clearly not be normal. Consequently, the multiple regression model is adapted by first introducing a transformation of the expected value of the response, $g(\mu)$, and then using a more suitable error distribution. The transformation g is called a *link function* in GLM, and a suitable link function for a binary response is the logistic or logit giving the model

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q. \quad (2)$$

As π varies from 0 to 1, the logit of π can vary from $-\infty$ to ∞ , so overcoming the first problem noted above. Now, we need to consider the appropriate error distribution. In linear regression, the observed value of the response variable is expressed as its expected value, given the explanatory variables plus an error term. With a binary response, we can express an observed value in the same way, that is:

$$y = \pi + \varepsilon, \quad (3)$$

but here, ε can only assume one of two possible values; if $y = 1$, then $\varepsilon = 1 - \pi$ with probability π , and if $y = 0$, then $\varepsilon = -\pi$ with probability $1 - \pi$. Consequently, ε has a distribution with mean zero and variance equal to $\pi(1 - \pi)$, that is, a binomial distribution for a single trial (also known as a Bernoulli distribution – see **Catalogue of Probability Density Functions**).

The Generalized Linear Model

Having seen the changes needed to the basic multiple linear regression model needed to accommodate a binary response variable, we can now see how the model is generalized in a GLM to accommodate a wide range of possible response variables with differing link functions and error distributions. The three essential components of a GLM are:

- A linear predictor, η , formed from the explanatory variables

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q. \quad (4)$$

- A transformation of the mean, μ , of the response variable called the link function, $g(\mu)$. In a GLM, it is $g(\mu)$ which is modeled by the linear predictor

$$g(\mu) = \eta. \quad (5)$$

In multiple linear regression and analysis of variance, the link function is the identity function. Other link functions include the log, logit, probit, inverse, and power transformations, although the log and logit are those most commonly met in practice. The logit link, for example, is the basis of logistic regression.

- The distribution of the response variable given its mean μ is assumed to be a distribution from the *exponential family*; this has the form

$$f(y; \theta, \phi) = \exp \left\{ \frac{(y\theta - b(\theta))}{a(\phi) + c(y, \phi)} \right\}. \quad (6)$$

For some specific functions, a , b , and c , and parameters θ and ϕ .

- For example, in linear regression, a normal distribution is assumed with mean μ and constant variance σ^2 . This can be expressed via the exponential family as follows:

$$\begin{aligned} f(y; \theta, \phi) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp \left\{ \frac{-(y - \mu)^2}{2\sigma^2} \right\} \\ &= \exp \left\{ \frac{(y\mu - \mu^2/2)}{\sigma^2} \right. \\ &\quad \left. - \frac{1}{2} \left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\} \end{aligned} \quad (7)$$

so that $\theta = \mu$, $b(\theta) = \theta^2/2$, $\phi = \sigma^2$ and $a(\phi) = \phi$. Other distributions in the exponential family include the binomial distribution, Poisson distribution, gamma distribution, and exponential distribution (see **Catalogue of Probability Density Functions**).

- Particular link function in GLMs are naturally associated with particular error distributions, for example, the identity link with the Gaussian distribution, the logit with the binomial, and the log with the Poisson. In these cases, the term *canonical link* is used.

The choice of probability distribution determines the relationships between the variance of the response variable (conditional on the explanatory variables) and its mean. This relationship is known as the *variance function*, denoted $\phi V(\mu)$. We shall say more about the variance function later.

Estimation of the parameters in a GLM is usually carried out through **maximum likelihood**. Details are given in [2, 6]. Having estimated the parameters, the question of the fit of the model for the sample data will need to be addressed. Clearly, a researcher needs to be satisfied that the chosen model describes the data adequately before drawing conclusions and making interpretations about the parameters themselves. In practice, most interest will lie in comparing the fit of competing models, particularly in the context of selecting subsets of explanatory variables so as to achieve a more parsimonious model. In GLMs, a measure of fit is provided by a quantity known as the *deviance*. This is essentially a statistic that measures how closely the model-based fitted values of the response approximate the observed values; the deviance quoted in most examples of GLM fitting is actually -2 times the maximized log-likelihood for a model, so that differences in deviances of competing models give a likelihood ratio test for comparing the models. A more detailed account of the assessment of fit for GLMs is given in [1].

An Example of Fitting a GLM

The data shown in Table 1 are given in [5] and also in [9]. They arise from asking randomly chosen household members from a probability sample of a town in the United States where stressful events had occurred within the last 18 months, and to report the month of occurrence of these events. A scatterplot of the data (see Figure 1) indicates a decline in the number of events as these lay further in the past, the result perhaps of the fallibility of human memory.

Since the response variable here is a count that can only take zero or positive values, it would not be appropriate to use multiple linear regression here to investigate the relationship of recalls to time. Instead, we shall apply a GLM with a log link function so that fitted values are constrained to be positive, and, as error distribution, use the Poisson distribution that is suitable for count data. These two assumptions lead to what is usually labelled *Poisson regression*.

Table 1 Distribution by months prior to interview of stressful events reported from subjects; 147 subjects reporting exactly one stressful event in the period from 1 to 18 months prior to interview. (Taken with permission from Haberman, 1978)

time	y
1	15
2	11
3	14
4	17
5	5
6	11
7	10
8	4
9	8
10	10
11	7
12	9
13	11
14	3
15	6
16	1
17	1
18	4

Explicitly, the model to be fitted to the mean number of recalls, μ , is:

$$\log(\mu) = \beta_0 + \beta_1 \text{time}. \quad (8)$$

The results of the fitting procedure are shown in Table 2.

The estimated regression coefficient for time is -0.084 with an estimated standard error of 0.017 . Exponentiating the equation above and inserting the

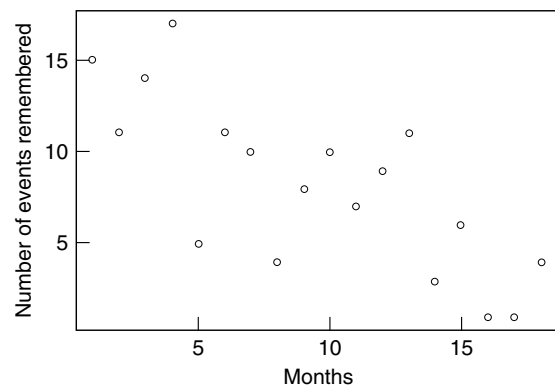


Figure 1 Plot of recalled memories data

4 Generalized Linear Models (GLM)

Table 2 Results of a Poisson regression on the data in Table 1.8

Covariates	Estimated regression coefficient	Standard error	Estimate/SE
(Intercept)	2.803	0.148	18.920
Time	−0.084	0.017	−4.987

(Dispersion Parameter for Poisson family taken to be 1).
Null Deviance: 50.84 on 17 degrees of freedom.
Residual Deviance: 24.57 on 16 degrees of freedom.

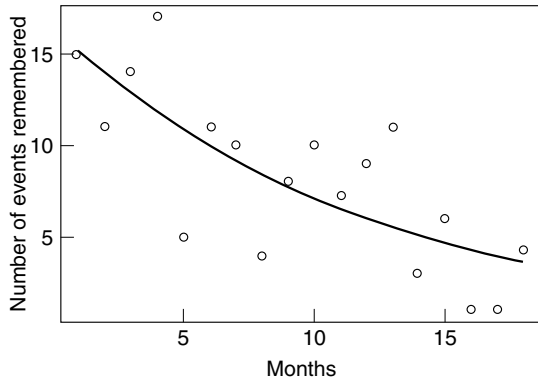


Figure 2 Recalled memories data showing fitted Poisson regression model

estimated parameter values gives the model in terms of the fitted counts rather than their logs, that is,

$$\hat{\mu} = 16.5 \times 0.920^{\text{time}}. \quad (9)$$

The scatterplot of the original data, now also showing the fitted model, is given in Figure 2. The difference in deviance of the null model and one including time as an explanatory variable is large and clearly indicates that the regression coefficient for time is not zero.

Overdispersion

An important aspect of generalized linear models that thus far we have largely ignored is the variance function, $V(\mu)$, that captures how the variance of a response variable depends upon its mean. The general form of the relationship is:

$$\text{Var}(\text{response}) = \phi V(\mu), \quad (10)$$

where ϕ is a constant and $V(\mu)$ specifies how the variance depends on the mean μ . For the error distributions considered previously, this general form becomes:

- (1) Normal: $V(\mu) = 1, \phi = \sigma^2$; here the variance does not depend on the mean and so can be freely estimated
- (2) Binomial: $V(\mu) = \mu(1 - \mu), \phi = 1$
- (3) Poisson: $V(\mu) = \mu; \phi = 1$

In the case of a Poisson variable, we see that the mean and variance are equal, and in the case of a binomial variable, where the mean is the probability of the occurrence of the event of interest, π , the variance is $\pi(1 - \pi)$. Both the Poisson and binomial distributions have variance functions that are completely determined by the mean. There is no free parameter for the variance since in applications of the generalized linear model with binomial or Poisson error distributions the dispersion parameter, ϕ , is defined to be one (see previous results for Poisson regression). But in some applications, this becomes too restrictive to fully account for the empirical variance in the data; in such cases, it is common to describe the phenomenon as *overdispersion*. For example, if the response variable is the proportion of family members who have been ill in the past year, observed in a large number of families, then the individual binary observations that make up the observed proportions are likely to be correlated rather than independent. This nonindependence can lead to a variance that is greater (less) than that on the assumption of binomial variability. And observed counts often exhibit larger variance than would be expected from the Poisson assumption, a fact noted by Greenwood and Yule over 80 years ago [4]. Greenwood and Yule's suggested solution to the problem was a model in which μ was a random variable with a γ distribution leading to a negative binomial distribution for the count (see **Catalogue of Probability Density Functions**).

There are a number of strategies for accommodating overdispersion but a relatively simple approach is one that retains the use of the binomial or Poisson error distributions as appropriate, but allows the estimation of a value of ϕ from the data rather than defining it to be unity for these distributions. The estimate is usually the residual deviance divided by its degrees of freedom, exactly

the method used with Gaussian models. Parameter estimates remain the same but parameter standard errors are increased by multiplying them by the square root of the estimated dispersion parameter. This process can be carried out manually, or almost equivalently, the overdispersed model can be formally fitted using a procedure known as *quasi-likelihood*; this allows estimation of model parameters without fully knowing the error distribution of the response variable – see [6] for full technical details of the approach.

When fitting generalized linear models with binomial or Poisson error distributions, overdispersion can often be spotted by comparing the residual deviance with its degrees of freedom. For a well-fitting model, the two quantities should be approximately equal. If the deviance is far greater than the degrees of freedom, overdispersion may be indicated.

An example of the occurrence of overdispersion when fitting a GLM with a log link and Poisson errors is reported in [8], for data consisting of the observation of number of days absent from school during the school-year amongst Australian Aboriginal and white children. The explanatory variables of interest in this study were gender, age, type (average or slow learner), and ethnic group (Aboriginal or White). Fitting the usual Poisson regression model resulted in a deviance of 1768 with 141 degrees of freedom, a clear indication of overdispersion. In this model, both gender and type were indicated as being highly significant predictors of number of days absent. But when overdispersion was allowed for in the way described above, both these variables became nonsignificant. A possible reason for overdispersion in these data is the substantial variability in children's tendency to miss days of school that cannot be fully explained by the variables that have been included in the model.

Summary

Generalized linear models provide a very powerful and flexible framework for the application of

regression models to medical data. Some familiarity with the basis of such models might allow medical researchers to consider more realistic models for their data rather than to rely solely on linear and logistic regression.

References

- [1] Cook, R.J. (1998). Generalized linear models, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester.
- [2] Dobson, A.J. (2001). *An Introduction to Generalized Linear Models*, 2nd Edition, Chapman & Hall/CRC Press, Boca Raton.
- [3] Everitt, B.S. (2001). *Statistics for Psychologists*, LEA, Mahwah.
- [4] Greenwood, M. & Yule, G.U. (1920). An inquiry into the nature of frequency-distributions of multiple happenings, *Journal of the Royal Statistical Society* **83**, 255.
- [5] Haberman, S. (1978). *Analysis of Qualitative Data*, Vol. I, Academic Press, New York.
- [6] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, London.
- [7] Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.
- [8] Rabe-Hesketh, S. & Everitt, B.S. (2003). *A Handbook of Statistical Analyses Using Stata*, Chapman & Hall/CRC Press, Boca Raton.
- [9] Seeber, G.U.H. (1998). Poisson regression, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester.

(See also **Generalized Additive Model**; **Generalized Estimating Equations (GEE)**)

BRIAN S. EVERITT

Genotype

ROBERT C. ELSTON

Volume 2, pp. 743–744

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Genotype

The genotype of an organism is its genes, or genetic make-up, as opposed to its phenotype, or outward appearance. The physical basis of the human genotype lies in 23 pairs of chromosomes – microscopic bodies present in every cell nucleus. One pair is the sex chromosomes, and the other 22 pairs are known as autosomes, or autosomal chromosomes. The two alleles at each locus on the autosomes comprise the genotype for that locus. If the two alleles are the same, the genotype is called *homozygous*; if they are different, the genotype is called *heterozygous*. Persons with homozygous and heterozygous genotypes are called *homozygotes* and *heterozygotes* respectively.

If the phenotype, or phenotypic distribution, associated with a particular heterozygote is the same as that associated with one of the two corresponding homozygotes, then the allele in that homozygote is **dominant**, and the allele in the other corresponding homozygote is recessive, with respect to the phenotype; the locus is said to exhibit dominance (*see Allelic Association*). If the heterozygote expresses a phenotype that has features of both corresponding homozygotes, – for example, persons with AB blood type have both A and B antigens on their red

cells, determined by A and B alleles at the ABO locus – then there is said to be codominance. At the DNA level, that is, if the phenotype associated with a genotype is the DNA constitution itself, then all loci exhibit codominance.

The genotype being considered may involve the alleles at more than one locus. However, a distinction should be made between the genotype at multiple loci and the multilocus genotype. Whereas the former is specified by all the alleles at the loci involved, the latter is specified by the two haplotypes a person inherited, that is, the separate sets of maternal alleles and paternal alleles at the various loci.

In the case of a quantitative trait, there is a dominance component to the variance if the heterozygote phenotype is not half-way between the two corresponding homozygote phenotypes, that is, if the phenotypic effects of the alleles at a locus are not additive. Similarly, if the phenotypic effect of a multilocus genotype is not the sum of the constituent one-locus genotypes, then there is **epistasis**. Dominance can be thought of as intralocus interaction and epistasis as interlocus interaction. Thus, in the case of a quantitative phenotype, the presence or absence of dominance and/or epistasis depends on the scale of measurement of the phenotype.

ROBERT C. ELSTON

Geometric Mean

DAVID CLARK-CARTER

Volume 2, pp. 744–745

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Geometric Mean

The geometric mean $\bar{X}_g$ of a set of n numbers $X_1, X_2, \dots, X_n (i = 1, 2, \dots, n)$ is defined as

$$\bar{X}_g = (X_1 \times X_2 \times \dots \times X_n)^{\frac{1}{n}} \quad (1)$$

The geometric mean is only defined for a set of positive numbers.

As an example, we see that the geometric mean of 10 and 15 is

$$\begin{aligned} \bar{X}_g &= (10 \times 15)^{\frac{1}{2}} = \sqrt{150} \\ &= 12.25 \text{ (to 2 decimal places).} \end{aligned} \quad (2)$$

The geometric mean has an advantage over the arithmetic mean in that it is less affected by very small or very large values in skewed data. For instance, the arithmetic mean of the scores 10, 15, and 150 is 58.33, whereas the geometric mean for the same figures is 28.23.

An additional use of the geometric mean is when dealing with numbers that are ratios of other numbers, such as percentage increase in weight. Imagine that we want the mean percentage weight gain for the data in Table 1, which shows a person's weight at the end of each year, expressed as a percentage of the previous year.

Table 1 Weight at the end of each year, expressed as a percentage of the previous year

Year	Percentage
1	104
2	107
3	109
4	113

The arithmetic mean of the percentages is 108.25, while the geometric mean is 108.20. Now, if we had been told that a person started year 1 with a weight of 50 kg, then that person's weight after each of the four years would be 52, 55.64, 60.65, and 68.53 kg, respectively. If the arithmetic mean were used to calculate the weights, it would give the person's weight for the four years as 54.13, 58.59, 63.42, and 68.66 kg, while if the geometric mean were used it would give the person's weight for the four years as 54.10, 58.54, 63.34, and 68.53 kg. Thus, although the geometric mean has produced figures for the intermediate years that are not totally accurate, unlike the arithmetic mean it produces the correct figure for the final value. Accordingly, it would be more precise to say that the person had an average weight gain of 8.2% than to say it was 8.25%.

DAVID CLARK-CARTER

Goodness of Fit

STANLEY A. MULAİK

Volume 2, pp. 745–749

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Goodness of Fit

Goodness-of-fit indices (GFI) are indices used in **structural equation modeling** and with other mathematical models to assess the fit of a model to data. They should be interpreted in terms of the overidentifying constraints placed on the model, for only these produce lack of fit. They are of two kinds: (a) exact fit indices and (b) approximate fit indices.

Exact fit indices are used with a sampling distribution to test the hypothesis that the model fits the population data exactly. Thus, before using an exact fit test, the researcher should have good reasons that the sampling distribution to be used in statistical inference applies to the data in question. Exact fit indices only test whether or not the hypothesized model fits the data to within sampling error. Fitting to within sampling error does not mean that the model is true. Other ‘equivalent models’ may be equally formulated to fit the data to within sampling error. Failure to fit to within sampling error only means that the data is so different and so improbable under the assumption of the truth of the hypothesis as to make the hypothesis difficult to believe. With failure to fit, one should seek to determine the sources of lack of fit, using program diagnostics and a review of theory and the content of the variables. However, failure to fit does not imply that every feature of the model is incorrect. It merely indicates that at least one feature *may* be incorrect, but it does not indicate which.

When the model fits to within sampling error, one may act as if the model is provisionally correct, but should be prepared for the possibility that at a later date further studies may overturn the model. Exact fit tests also increase in power to reject the hypothesized model with ever larger samples. Most multivariate models require large samples to justify the theoretical sampling distributions used with exact fit indices. So, just as it becomes reasonable to use these distributions in probabilistic inference, the power of these tests may make it possible to detect miniscule, if not practically negligible, discrepancies between the mathematical model hypothesized and the data, leading to rejection of the model. This, and the fact that many data sets do not satisfy exactly the distributional assumptions of the exact fit test, has been the basis for many researchers’ use of approximate fit indices to judge the quality of their models by their degree of approximation to the data.

Approximate fit indices indicate the degree to which the model fits the data. Some approximate fit indices use indices that range between 0 (total lack of fit) and 1 (perfect fit). Other approximate fit indices measure the distance between the data and the model, with 0 perfect fit, and infinity indicating total lack of fit. Usually, approximate fit indices are used with some criterion value close to perfect fit to indicate whether the model is a good approximation or not. However, it must always be recognized that an approximation is always *that* and not indicative of the incorrigible correctness of the model. Getting good approximations is indicative that a preponderance of the evidence supports the essential features of the model. But, further research with the model should also include a search for the reasons for lack of exact fit.

Let $\mathbf{U}$ and $\mathbf{V}$ be two $p \times p$ covariance matrices, with $\mathbf{U}$ a less restricted variance–covariance matrix representing the data and $\mathbf{V}$ a more constrained variance–covariance matrix hypothesized for the data. Browne [2] noted that most structural equation models are fit to data using discrepancy functions, which are scalar valued-functions $F(\mathbf{U}; \mathbf{V})$ with the following properties: (a) $F(\mathbf{U}; \mathbf{V}) \geq 0$. (b) $F(\mathbf{U}; \mathbf{V}) = 0$ if and only if $\mathbf{U}$ and $\mathbf{V}$ are equal. (c) $F(\mathbf{U}; \mathbf{V})$ is continuous in both $\mathbf{U}$ and $\mathbf{V}$.

There are a number of standard discrepancy functions, like unweighted least squares, weighted least squares (*see Least Squares Estimation*) and maximum likelihood’s fit function (*see Maximum Likelihood Estimation*) that are used in estimating unknown parameters of models conditional on constraints on specified model parameters. For example, the least-squares discrepancy function is

$$F_{LS} = \text{tr}[(\mathbf{U} - \mathbf{V})(\mathbf{U} - \mathbf{V})], \quad (1)$$

and the maximum likelihood’s fit function

$$F_{ML} = \ln|\mathbf{V}| - \ln|\mathbf{U}| - \text{tr}(\mathbf{U}\mathbf{V}^{-1}) - p \quad (2)$$

is also a discrepancy function. In fact, the chi-squared statistic $\chi_{df}^2 = (N - 1)F_{ML}$ is also a discrepancy function.

Cudeck and Henley [4] defined three sources of error in model fitting. First, let $\mathbf{S}$ be a $p \times p$ sample variance–covariance matrix for p observed variables to be used as the data to which a structural equation model is fit. Let Σ be the variance/covariance matrix of the population from which the sample

2 Goodness of Fit

variance/covariance matrix $\mathbf{S}$ has been drawn. Next, define $\hat{\Sigma}_0$ to be the maximum likelihood (ML) estimate of a model with free, fixed, and constrained parameters fit to $\mathbf{S}$. The error of fit given by the discrepancy function value $F(\mathbf{S}; \hat{\Sigma}_0)$ contains both sampling error and error of approximation of the model to the population covariance matrix. Analogously, let $\tilde{\Sigma}_0$ be the estimate of the same model fit to the population variance–covariance matrix Σ . The error of fit in this case is $F(\Sigma; \tilde{\Sigma}_0)$ and is known as the *error of approximation*. It contains no sampling error. It is a population parameter of lack of fit of the model to the data. It is never measured directly and is only inferred from the data. $F(\tilde{\Sigma}_0; \hat{\Sigma}_0)$ is known as the *error of estimation*, the discrepancy between sample estimate and population estimate for the model.

The chi-squared statistic of an exact fit test is given as

$$\chi_{df}^2 = (N-1)F(\mathbf{S}; \hat{\Sigma}_0) = (N-1) \times \left[\ln|\hat{\Sigma}_0| - \ln|\mathbf{S}| + \text{tr}(\hat{\Sigma}_0^{-1}\mathbf{S}) - p \right]. \quad (3)$$

$F(\mathbf{S}; \hat{\Sigma}_0)$ is the minimum value of the discrepancy function when the maximum likelihood estimates are optimized.

Assuming the variables have a joint multivariate normal distribution (see **Catalogue of Probability Density Functions**), this statistic is used to test the hypothesis that the constrained model covariance matrix is the population covariance matrix that generated the sample covariance matrix $\mathbf{S}$. One rejects the null hypothesis that the model under the constraints generated $\mathbf{S}$, when $\chi_{df}^2 > c$, where c is some constant such that $P(\chi_{df}^2 > c | H_0 \text{ is true}) \leq \alpha$. Here α is the probability one accepts of making a Type I error.

In many cases, the model fails to fit the data. In this case, chi squared does not have an approximate chi-squared distribution, but a noncentral chi-squared distribution, whose expectation is

$$E(\chi_{df}^2) = df + \delta^{*2}, \quad (4)$$

where df are the degrees of freedom of the model and δ^{*2} , the noncentrality parameter for the model. The noncentrality parameter is a measure of the lack of fit of the model for samples of size N . Thus, an unbiased estimate of the noncentrality parameter is given by

$$\hat{\delta}^{*2} = \chi_{df}^2 - df. \quad (5)$$

McDonald [8] recommended that instead of using this as a fundamental index of lack of fit, one should normalize this estimate by dividing it by $(N-1)$ to control for the effect of sample size. Thus, the normalized noncentrality estimate is given by

$$\hat{\delta} = \frac{\hat{\delta}^{*2}}{(N-1)} = \frac{\chi_{df}^2 - df}{(N-1)} = F_{ML} - \frac{df}{(N-1)}. \quad (6)$$

The population raw noncentrality δ^{*2} and the population normalized noncentrality δ are related as $\delta^{*2} = (N-1)\delta$. As N increases without bound, and $\delta > 0$, the noncentrality parameter is undefined in the limit.

Browne and Cudeck [3] argue that $\hat{\delta}$ is a less biased estimator of the normalized population discrepancy $\delta = F(\Sigma; \tilde{\Sigma}_0)$ than is the raw $F_{ML} = F(\mathbf{S}; \hat{\Sigma}_0) = \chi_{df}^2/(N-1)$, which has for its expectation

$$E(F_{ML}) = F(\Sigma; \tilde{\Sigma}_0) + \frac{df}{(N-1)}. \quad (7)$$

In fact,

$$\begin{aligned} E(\hat{\delta}) &= E(F_{ML}) - \frac{df}{(N-1)} = E(F_{ML}) \\ &\quad + \frac{df - df}{(N-1)} = F(\Sigma; \tilde{\Sigma}_0) = \delta. \end{aligned} \quad (8)$$

So, the estimated normalized noncentrality parameter is an unbiased estimate of the population normalized discrepancy.

Several indices of approximation are based on the noncentrality and normalized noncentrality parameters.

Bentler [1] and McDonald and Marsh [9] simultaneously defined an index given the name FI by Bentler:

$$\begin{aligned} FI &= \frac{(\hat{\delta}_{\text{null}}^* - \hat{\delta}_k^*)}{\hat{\delta}_{\text{null}}^*} = \frac{[(\chi_{\text{null}}^2 - df_{\text{null}}) - (\chi_k^2 - df_k)]}{(\chi_{\text{null}}^2 - df_{\text{null}})} \\ &= \frac{\hat{\delta}_{\text{null}} - \hat{\delta}_k}{\hat{\delta}_{\text{null}}}. \end{aligned} \quad (9)$$

χ_{null}^2 is the chi squared of a ‘null’ model in which one hypothesizes that the population covariance matrix is a diagonal matrix with zero off-diagonal elements and free diagonal elements. Each nonzero covariance between any pair of variables in the data will produce a lack of fit for the corresponding zero covariance in this model. Hence, lack of fit δ_{null} of this model

can serve as an extreme norm against which to compare the lack of fit of model k , which is actually hypothesized. The difference in lack of fit between the null model and model k is compared to the lack of fit of the null model itself. The result on the right is obtained by dividing the numerator and the denominator by $(N - 1)$. The index depends on unbiased estimates and is itself relatively free of bias at different sample sizes. Bentler [1] further corrected the FI index to be 0 when it became occasionally negative and to be 1 when it occasionally exceeded 1. He called the resulting index the CFI (comparative fit index). A common rule of thumb is that models with $CFI \geq .95$ are ‘acceptable’ approximations.

Another approximation index first recommended by Steiger and Lind [10] but popularized by Browne and Cudeck [3] is the RMSEA (root mean squared error of approximation) index, given by

$$\begin{aligned} RMSEA &= \sqrt{\text{Max.} \left\{ \left(\frac{\chi^2_{df_k} - df_k}{(N - 1)df_k} \right), 0 \right\}} \\ &= \sqrt{\text{Max.} \left\{ \left(\frac{\hat{\delta}_k}{df_k} \right), 0 \right\}}. \end{aligned} \quad (10)$$

This represents the square root of the estimated normalized noncentrality of the model divided by the model’s degrees of freedom. In other words, it is the average normalized noncentrality per degree of freedom. Although some have asserted that this represents the noncentrality adjusted for model parsimony, this is not the case. A model may introduce constraints and be more parsimonious with more degrees of freedom, and the average discrepancy per additional degree of freedom may not change. The RMSEA index ranges between 0 (perfect fit) and infinity. A value of $RMSEA \leq .05$ is considered to be ‘acceptable’ approximate fit. Browne and Cudeck [3] indicate that a confidence interval estimate for the RMSEA is available to indicate the precision of the RMSEA estimate.

Another popular index first popularized by Jöreskog and Sörbom’s LISREL program [7] (see **Structural Equation Modeling: Software**) is inspired by Fisher’s intraclass correlation [5]:

$$R^2 = 1 - \frac{\text{Error Variance}}{\text{Total Variance}}. \quad (11)$$

The GFI index computes ‘error’ as the sum of (weighted and possibly transformed) squared differences between the elements of the observed variance/covariance matrix $\mathbf{S}$ and those of the estimated model variance/covariance matrix $\hat{\Sigma}_0$ and compares this sum to the total sum of squares of the elements in $\mathbf{S}$. The matrix $(\mathbf{S} - \hat{\Sigma}_0)$ is symmetric and produces the element-by-element differences between $\mathbf{S}$ and $\hat{\Sigma}_0$. $\mathbf{W}$ is a transformation matrix that weights and combines the elements of these matrices, depending on the method of estimation. Thus, we have

$$GFI = 1 - \frac{\text{tr}[\mathbf{W}^{-1/2}(\mathbf{S} - \hat{\Sigma}_0)\mathbf{W}^{-1/2}]^2}{\text{tr}[\mathbf{W}^{-1/2}(\mathbf{S})\mathbf{W}^{-1/2}][\mathbf{W}^{-1/2}(\mathbf{S})\mathbf{W}^{-1/2}]}, \quad (12)$$

where

$\hat{\Sigma}_0$ is the model variance/covariance matrix and $\mathbf{S}$ is the unrestricted, sample variance/covariance matrix and

$$\mathbf{W} = \begin{cases} \mathbf{I} - \text{Unweighted Least Squares} \\ \mathbf{S} - \text{Weighted Least Squares} \\ \hat{\Sigma}_0 - \text{Maximum Likelihood} \end{cases}. \quad (13)$$

For maximum likelihood estimation the GFI simplifies to

$$GFI_{ML} = 1 - \frac{\text{tr}[(\mathbf{S}\hat{\Sigma}_0^{-1} - \mathbf{I})(\mathbf{S}\hat{\Sigma}_0^{-1} - \mathbf{I})]}{\text{tr}(\mathbf{S}\hat{\Sigma}_0^{-1}\mathbf{S}\hat{\Sigma}_0^{-1})}. \quad (14)$$

A rule of thumb is again to consider a $GFI > 0.95$ to be an ‘acceptable’ approximation. Hu and Bentler [6] found that the GFI tended to underestimate its asymptotic value in small samples, especially when the latent variables were interdependent. Furthermore, the maximum likelihood (ML) and generalized least squares (GLS) variants of the index performed poorly in samples less than 250.

Steiger [11] has suggested a variant of the GFI such that under a general condition where the model is invariant under a constant scaling function the GFI has a known population parameter

$$\Gamma_1 = \frac{p}{2F_{ML}(\Sigma; \tilde{\Sigma}_0) + p} \quad (15)$$

to estimate. Note that as $F(\Sigma; \tilde{\Sigma}_0)$ becomes close to zero, this index approaches unity, whereas when $F(\Sigma; \tilde{\Sigma}_0)$ is greater than zero and increasing, this

4 Goodness of Fit

parameter declines toward zero, with its becoming zero when $F(\Sigma; \hat{\Sigma}_0)$ is infinitely large. Steiger shows that

$$\hat{\Gamma}_1 = \frac{p}{2F_{ML}(\mathbf{S}; \hat{\Sigma}_0) + p} \quad (16)$$

is equivalent to the GFI_{ML} and an estimate of Γ_1 . But it is a biased estimate, for the expectation of $\hat{\Gamma}_1$ is approximately

$$E(\hat{\Gamma}_1) \approx \frac{p}{2F_{ML}(\Sigma; \hat{\Sigma}_0) + 2df/(N-1) + p}. \quad (17)$$

The bias leads $\hat{\Gamma}_1$ to underestimate Γ_1 , but the bias diminishes as sample size N becomes large relative to the degrees of freedom of the model. Steiger [11] and Browne and Cudeck [3] report a confidence interval estimate using $\hat{\Gamma}_1$ that may be used to test hypotheses about Γ_1 .

There are numerous other indices of approximate fit, but those described here are the most popular.

Goodness of fit should not be the only criterion for evaluating a model. Models with zero degrees of freedom always fit perfectly as a mathematical necessity and, thus, are useless for testing hypotheses. Besides having acceptable fit, the model should be parsimonious in having numerous degrees of freedom relative to the number of nonredundant elements in the variance-covariance matrix, and should be realistic in representing processes in the phenomenon modeled.

References

- [1] Bentler, P.M. (1990). Comparative fit indices in structural models, *Psychological Bulletin* **107**, 238–246.
- [2] Browne, M.W. (1982). Covariance structures, in *Topics in Applied Multivariate Analysis*, D.M. Hawkins, ed., University Press, Cambridge.
- [3] Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage Publications, Newbury Park, pp. 136–162.
- [4] Cudeck, R. & Henley, S.J. (1991). Model selection in covariance structures analysis and the “problem” of sample size: a clarification, *Psychological Bulletin* **109**, 512–519.
- [5] Fisher, R.A. (1925). *Statistical Methods for Research Workers*, Oliver and Boyd, London.
- [6] Hu, L.-T. & Bentler, P.M. (1995). Evaluating model fit, in *Structural Equation Modeling: Concepts, Issues and Applications*, R.H. Hoyle, ed., Sage Publications, Thousand Oaks.
- [7] Jöreskog, K.G. & Sörbom, D. (1981). *LISREL V: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, National Educational Resources, Chicago.
- [8] McDonald, R.P. (1989). An index of goodness of fit based on noncentrality, *Journal of Classification* **6**, 97–103.
- [9] McDonald, R.P. & Marsh, H.W. (1990). Choosing a multivariate model: noncentrality and goodness-of-fit, *Psychological Bulletin* **107**, 247–255.
- [10] Steiger, J.H. (1995). Structural Equation Modeling (SEPATH), in *Statistica/W* (Version 5), Statsoft, Inc., Tulsa, OK, pp. 3539–3688.
- [11] Steiger, J.H. & Lind, J.C. (1980). Statistically based tests for the number of factors, in *Paper Presented at the Annual Spring Meeting of the Psychometric Society*, Iowa City.

STANLEY A. MULAİK

Goodness of Fit for Categorical Variables

ALEXANDER VON EYE, G. ANNE BOGAT AND STEFAN VON WEBER

Volume 2, pp. 749–753

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Goodness of Fit for Categorical Variables

Introduction

The explanation of the frequency distributions of categorical variables often involves the specification of probabilistic models. The appropriateness of these models can be examined using *goodness-of-fit statistics*. A goodness-of-fit statistic assesses the distance between the data and the model. This characteristic is the reason why goodness-of-fit statistics are often also called *badness-of-fit statistics*. Under the null hypothesis that this distance is random, the probability is then estimated that the observed data or data with even larger distances are obtained. This probability is called the *size of the test*.

Goodness-of-fit is examined at two levels. The first is the aggregate level. At this level, one asks whether a probability model describes the data well, overall. The second is the level of individual cells or groups of cells. At this level, one asks whether the distances in individual cells or groups of cells are greater than compatible with the assumption of random deviations. Cells that display particularly strong deviations have been called *outlier cells*, *extreme cells*, *outlandish cells* [14], *rogue cells* [15], *aberrant cells*, *anomalous cells* [21], and *types and antitypes* [10, 16].

A large number of tests has been proposed to examine goodness-of-fit, and performance and characteristics of these tests differ greatly (comparisons were presented, e.g., by [16–18]). In the following paragraphs, we discuss goodness-of-fit tests from two perspectives. First, we present the tests and describe them from the perspective of *overall goodness-of-fit*, that is, from the perspective of whether a probability model as a whole explains the data well. Second, we discuss tests and their performance when applied to individual cells. In both sections, we focus on count data and on multinomial sampling.

Testing Overall Goodness-of-fit

A most general indicator of goodness-of-fit is Cressie and Read's [3, 13] *power divergence statistic*,

$$I(\lambda) = \frac{2}{\lambda(\lambda + 1)} \sum_i n_i \left[\left(\frac{n_i}{m_i} \right)^\lambda - 1 \right], \quad (1)$$

where

λ = real-valued parameter, with $-\infty < \lambda < \infty$,

i = index that goes over all cells of a table,

n_i = observed frequency of Cell i , and

m_i = expected frequency of Cell i .

This statistic is important because it can, by way of selecting particular scores of λ , be shown to be identical to other well-known measures of goodness-of-fit (see [1]). Specifically,

- (1) for $\lambda = 1$, I is equal to Pearson's X^2 ; one obtains

$$I(1) = X^2 = \sum \frac{(n_i - m_i)^2}{m_i}; \quad (2)$$

- (2) as $\lambda \rightarrow 0$, I converges to the likelihood ratio G^2 ,

$$I(\lambda \rightarrow 0) \rightarrow G^2 = 2 \sum_i n_i \log \left(\frac{n_i}{m_i} \right); \quad (3)$$

- (3) as $\lambda \rightarrow -1$, I converges to Kullback's [7] *minimum discrimination information statistic*,

$$I(\lambda \rightarrow -1) \rightarrow GM \rightarrow 2 \sum_i m_i \log \left(\frac{n_i}{m_i} \right); \quad (4)$$

- (4) for $\lambda = -2$, I is equal to Neyman's [12] *modified chi-squared statistic*,

$$I(\lambda = -2) = NM^2 = \sum_i \frac{(n_i - m_i)^2}{n_i}; \quad (5)$$

and

- (5) for $\lambda = -0.5$, I is equal to Freeman and Tukey's [4] statistic

$$I\left(\frac{1}{2}\right) = FT = 4 \sum_i (\sqrt{n_i} - \sqrt{m_i})^2. \quad (6)$$

Under regularity conditions, these six statistics are asymptotically identically distributed. Specifically, these statistics are asymptotically distributed as χ^2 , with $df = I - 1$, where I is the number of cells. However, the value of $\lambda = 2/3$ has shown to be

2 Goodness of Fit for Categorical Variables

superior to other values of λ . It leads to a statistic that keeps the α level better, and has better small sample power characteristics.

Comparisons of the two best known of these six statistics, the Pearson X^2 and the likelihood ratio X^2 (see **Contingency Tables**), have shown that the Pearson statistic is often closer to the χ^2 distribution than G^2 . However, G^2 has better decomposition characteristics than X^2 . Therefore, decompositions of the effects in cross-classifications (see **Chi-Square Decomposition**) and comparisons of hierarchically related **log-linear models** are typically performed using the G^2 statistic.

There exists a number of other goodness-of-fit tests. These include, for instance, the Kolmogorov–Smirnov test, the Cramer-von Mises test, and runs tests.

Testing Cellwise Goodness-of-fit

Testing goodness-of-fit for individual cells has been proposed for at least three reasons. First, the distribution of residuals in cross-classifications that are evaluated using particular probability models can be very uneven such that the residuals are large for a small number of cells and rather small for the remaining cells. Attempts at improving model fit then focus on reducing the discrepancies in the cells with the large residuals. Second, cells with large residuals can be singled out and declared structural frequencies, that is, *fixed*, and not taken into account when the expected frequencies are estimated. Typically, model fit improves considerably when outlandish cells are fixed. Third, cell-specific goodness-of-fit indicators are examined with the goal of finding types and antitypes in **configural frequency analysis** see von Eye in this encyclopedia, [10, 16, 17]. Types and antitypes are then interpreted with respect to the probability model that was used to explain a cross-classification. Different probability models can yield different patterns of types and antitypes. In either context, decisions are made concerning single cells or groups of cells.

The most popular measures of cellwise lack of fit include the *raw residual*, the *Pearson residual*, the *standardized Pearson residual*, and the *adjusted residual*. The raw residual is defined as the difference between the observed and the expected cell frequencies, that is, $n_i - m_i$. This measure indicates

the number of cases by which the observed frequency distribution and the probability model differ in Cell i . The Pearson residual for Cell i is the i th summand of the overall X^2 given above.

For the standardized Pearson residual, one can find different definitions in the literature. According to Agresti [1], for Cell i which has an estimated leverage of $\hat{h}_i$, the standardized Pearson residual is

$$r_i = \frac{n_i - m_i}{\sqrt{m_i \left(1 - \frac{n_i}{N}\right) (1 - \hat{h}_i)}}, \quad (7)$$

where $\hat{h}_i$ is defined as the diagonal element of the *hat* matrix (for more detail on the hat matrix see [11]). The absolute value of the standardized Pearson residual is slightly larger than the square root of the Pearson residual (which is often called the standardized residual; see [8]), and it is approximately normally distributed if the model holds.

The *adjusted residual* [5] is a standardized residual that is divided by its standard deviation. Adjusted residuals are typically closer to the normal distribution than $\sqrt{X^2}$. *Deviance residuals* are the components of the likelihood ratio statistic, G^2 , given above. Exact residual tests can be performed using, for instance, the binomial test and the hypergeometric test. The latter requires product-multinomial sampling (see **Sampling Issues in Categorical Data**).

The characteristics and performance of these and other residual tests have been examined in a number of studies (e.g., [6, 9, 17, 19, 20]). The following list presents a selection of repeatedly reported results of comparison studies.

1. The distribution of the adjusted residual is closer to the normal distribution than that of the standardized residual.
2. Both the approximative and the **exact tests** tend to be conservative when the expected cell frequencies are estimated from the sample marginals.
3. As long as cell frequencies are small (less than about $n_i = 100$), the distribution of residuals tends to be asymmetric such that positive residuals are more likely than negative residuals; for larger cell sizes, this ratio is inverted; this applies to tables of all sizes greater than 2×2 and under multinomial as well as product-multinomial sampling.

4. The α -curves of the residuals suggest conservative decisions as long as cell sizes are small.
5. The β curves of the residuals also suggest that large sample sizes are needed to make sure the β -error is not severely inflated; this applies to tables of varying sizes, to both multinomial and product-multinomial sampling, as well as to the α -levels of 0.05 and 0.01.
6. None of the tests presented here and the other tests that were also included in some of the comparison studies consistently outperformed all others; that is, the tests are differentially sensitive to characteristics of data and tables.

Data Example

The following data example presents a reanalysis of data from a project on the mental health outcomes of women experiencing domestic violence [2]. For the example, we attempt to predict Anxiety (A) from Poverty (Po), Psychological Abuse (Ps), and Physical Abuse (Ph). Anxiety and Poverty were dichotomized at the median, and Psychological and Physical Abuse were dichotomized at the score of 0.7 (to separate the no abuse cases from the abuse cases). For each variable, a 1 indicates a low score, and a 2 indicates a high score.

To test the prediction hypotheses, we estimated the hierarchical log-linear model [A, Po], [A, Ps], [A, Ph], [Po, Ps, Ph]. This model is equivalent to the **logistic regression model** that predicts A from Po, Ps, and Ph. The cross-tabulation of the four variables, including the observed and the estimated expected cell frequencies, appears in Table 1.

Table 1 suggests that the observed cell frequencies are relatively close to the estimated expected cell frequencies. Indeed, the overall goodness-of-fit likelihood ratio $X^2 = 3.50$ ($df = 4$; $p = 0.35$) indicates no significant model-data discrepancies. The Pearson $X^2 = 4.40$ ($df = 4$; $p = 0.48$) leads one to the same conclusion. The values of these two overall goodness-of-fit measures are the same if a model is the true one. In the present case, the values of the test statistics are not exactly the same, but since they are both small and suggest the same statistical decision, there is no reason to alter the model. Substantively, we find that the two abuse variables are significant predictors of Anxiety. In contrast, Poverty fails to make a significant contribution.

Table 1 Prediction of anxiety from poverty, psychological violence, and physical violence

Variable patterns				Frequencies	
A	Po	Ps	Ph	Observed	Expected
1	1	1	1	6	6.93
1	1	1	2	3	2.07
1	1	2	1	13	11.64
1	1	2	2	13	14.36
1	2	1	1	2	1.11
1	2	1	2	0	0.89
1	2	2	1	1	2.32
1	2	2	2	9	7.68
2	1	1	1	4	3.96
2	1	1	2	1	1.04
2	1	2	1	11	11.47
2	1	2	2	13	12.53
2	2	1	1	0	0.00
2	2	1	2	0	0.00
2	2	2	1	5	4.57
2	2	2	2	13	13.43

Table 2 presents the raw, the adjusted, the deviance, the Pearson, and the Freeman–Tukey residuals for the above model.

In the following paragraphs, we discuss four characteristics of the residuals in Table 2. First, the various residual measures indicate that no cell qualifies as extreme. None of the residuals that are distributed either normally or as χ^2 has values that would indicate that a cell deviates from the model (to make this statement, we use the customary thresholds of 2 for normally distributed residuals and 4 for χ^2 -distributed residuals). Before retaining a model, researchers can do worse than inspecting residuals for local model-data deviations.

Second, the Pearson X^2 is the only one that does not indicate the direction of the deviation. From inspecting the Pearson residual alone, one cannot determine whether an observed frequency is larger or smaller than the corresponding estimated expected one.

Third, the correlations among the four arrays of residuals vary within a narrow range, thus indicating that the measures are sensitive largely to the same characteristics of model-data discrepancies. Table 3 displays the correlation matrix.

The correlations in Table 3 are generally very high. Only the correlations with Pearson's measure are low. The reason for this is that the Pearson scores are positive by definition. Selecting only the positive

4 Goodness of Fit for Categorical Variables

Table 2 Residuals from the prediction model in Table 1

Variable patterns				Residuals				
A	Po	Ps	Ph	Raw	Adjusted	Deviance	Pearson	Freeman–Tukey
1	1	1	1	−0.93	−1.14	−1.32	0.13	−0.27
1	1	1	2	0.93	1.14	1.50	0.42	0.69
1	1	2	1	1.36	1.29	1.70	0.16	0.45
1	1	2	2	−1.36	−1.29	−1.61	0.13	−0.30
1	2	1	1	0.89	1.48	1.53	0.71	0.81
1	2	1	2	−0.89	−1.48	0.00	0.89	−1.14
1	2	2	1	−1.32	−1.34	−1.30	0.75	−0.79
1	2	2	2	1.32	1.34	1.69	0.23	0.53
2	1	1	1	0.04	0.06	0.30	0.00	0.13
2	1	1	2	−0.04	−0.06	−0.29	0.002	0.14
2	1	2	1	−0.47	−0.41	−0.96	0.02	−0.07
2	1	2	2	0.47	0.41	0.98	0.02	0.20
2	2	1	1	−0.0004	−0.02	0.00	0.00	−0.001
2	2	1	2	−0.0003	−0.02	0.00	0.00	−0.000
2	2	2	1	0.43	0.43	0.94	0.04	0.29
2	2	2	2	−0.43	−0.43	−0.92	0.01	−0.05

Table 3 Intercorrelations of the residuals in Table 2

	Raw	Adjusted	Deviance	Pearson
Adjusted	0.992	1.000		
Deviance	0.978	0.975	1.000	
Pearson	0.249	0.314	0.297	1.000
Freeman–Tukey	0.946	0.971	0.940	0.344

less than 1. The deviance residual has a standard deviation greater than one. This is an unusual result and may be specific to the data used for the present example. In general, the deviance residual is less variable than $N(0, 1)$, but it can be standardized.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.
- [2] Bogat, G.A., Levendosky, A.A., De Jonghe, E., Davidson, W.S. & von Eye, A. (2004). Pathways of suffering: The temporal effects of domestic violence. *Maltrattamento e abuso all'infanzia* 6(2), 97–112.
- [3] Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society, Series B* 46, 440–464.

residuals, the correlations with the Pearson residuals would be high also.

Fourth, the standard deviations of the residual scores are different than 1. Table 4 displays descriptive measures for the variables in Table 2.

It is a well-known result that the standard deviations of residuals can be less than 1 when a model fits. The Freeman–Tukey standard deviation is clearly

Table 4 Description of the variables in Table 2

	Raw	Adjusted	Deviance	Pearson	Freeman–Tukey
Minimum	−1.360	−1.340	−1.610	0.000	−0.790
Maximum	1.360	1.480	1.700	0.750	0.810
Mean	0.076	0.083	0.112	0.116	0.120
Standard Dev	0.945	0.951	1.303	0.146	0.326
Variance	0.892	0.904	1.697	0.021	0.106
Skewness(G1)	−0.027	−0.003	−0.012	2.551	0.037
SE Skewness	0.249	0.249	0.249	0.249	0.249
Kurtosis(G2)	−1.272	−1.316	−1.745	7.977	−0.628
SE Kurtosis	0.493	0.493	0.493	0.493	0.493

-
- [4] Freeman, M.F. & Tukey, J.W. (1950). Transformations related to the angular and the square root, *Annals of Mathematical Statistics* **77**, 607–611.
- [5] Haberman, S.J. (1973). The analysis of residuals in cross-classifications, *Biometrics* **29**, 205–220.
- [6] Koehler, K.J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials, *Journal of the American Statistical Association* **75**, 336–344.
- [7] Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.
- [8] Lawal, B. (2003). *Categorical data analysis with SAS and SPSS applications*, NJ: Lawrence Erlbaum, Mahwah.
- [9] Lawal, H.B. (1984). Comparisons of χ^2 , G^2 , Y^2 , Freeman-Tukey, and Williams improved G^2 test statistics in small samples of one-way multinomials, *Biometrika* **71**, 415–458.
- [10] Lienert, G.A. & Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types, *Educational and Psychological Measurement* **35**, 231–238.
- [11] Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th Edition, Irwin, Chicago.
- [12] Neyman, J. (1949). Contribution to the theory of the χ^2 test, in *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed., University of California Press, Berkeley, pp. 239–273.
- [13] Read, T.R.C. & Cressie, N. (1988). *Goodness-of-Fit for Discrete Multivariate Data*, Springer-Verlag, New York.
- [14] SYSTAT (2002). *SYSTAT 10.2.*, SYSTAT software, Richmond.
- [15] Upton, G.J.G. (1978). *The Analysis of Cross-Tabulated Data*, Wiley, Chichester.
- [16] von Eye, A. (2002a). *Configural Frequency Analysis – Methods, Models, and Applications*, Lawrence Erlbaum, Mahwah.
- [17] von Eye, A. (2002b). The odds favor antitypes – A comparison of tests for the identification of configural types and antitypes, *Methods of Psychological Research – Online* **7**, 1–29.
- [18] von Weber, S., Lautsch, E. & von Eye, A. (2003). Table-specific continuity corrections for configural frequency analysis, *Psychology Science* **45**, 355–368.
- [19] von Weber, S., von Eye, A. & Lautsch, E. (2004). The Type II error of measures for the analysis of 2×2 tables, *Understanding Statistics* **3**, 259–282.
- [20] West, E.N. & Kempthorne, O. (1972). A comparison of the χ^2 and likelihood ratio tests for composite alternatives, *Journal of Statistical Computation and Simulation* **1**, 1–33.
- [21] Wickens, T. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Hillsdale.

(See also **Configural Frequency Analysis**)

ALEXANDER VON EYE, G. ANNE BOGAT AND
STEFAN VON WEBER

Gosset, William Sealy

ROGER THOMAS

Volume 2, pp. 753–754

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Gosset, William Sealy

Born: June 13, 1876, in Canterbury, England.

Died: October 16, 1937, in Beaconsfield, England.

William Sealy Gosset studied at Winchester College before enrolling at New College, Oxford, where he earned a first class in mathematical moderations (1887) and another in natural sciences, specializing in chemistry (1899). Also, in 1899, and in conjunction with the company's plan to hire university-trained scientists, Arthur Guinness, Son & Co, brewers in Dublin, hired Gosset. Thus began Gosset's career, one that culminated with his appointment in 1935 as head of Guinness's newly constructed brewery in London, a position he held until his death. '...all his most important statistical work was undertaken in order to throw light on problems which arose in the analysis of data connected in some way with the brewery' [2, p. 212]. Gosset was, perhaps, the first and the most important industrial statistician [4].

After familiarizing himself with the operations of the brewery, where he had access to data bearing on brewing methods as well as the production and combinations of the hops and barley used in brewing, Gosset realized the potential value of applying error theory to such data. His first report (November 3, 1904) titled, 'The Application of the "Law of Error" to the work of the Brewery', presented the case for introducing statistical methods to the industry's work. Gosset also observed, 'We have met with the difficulty that none of our books mentions the odds, which are conveniently accepted as being sufficient to establish any conclusion, and it might be of assistance to us to consult some mathematical physicist on the matter' [2, p. 215].

Instead, Gosset contacted **Karl Pearson** (1905), which led to Gosset's studies (1906–1907) with Pearson and W. F. R. Weldon in the Biometric School of the University College, London. Following Francis Galton's lead, Pearson and Weldon were keen on refining measures of variation and correlation, primarily for agricultural and biological purposes, and to do so, they relied on large statistical samples. Gosset had earlier noted that 'correlation coefficients are usually calculated from large numbers of cases, in fact I have found only one paper in

Biometrika of which the cases are as few in number as those at which I have been working lately' (quoted in [2, p. 217].

Thus, it fell to Gosset, who typically had much smaller samples from the brewery's work available to him, to adapt the large-sample statistical methods to small samples. To develop small-sample methods, he drew small samples from some of Pearson et al.'s large samples, and in so doing Gosset provided '...the first instance in statistical research of the random sampling experiment...' [2, p. 223].

Guinness had a policy of not publishing the results of company research, but Gosset was permitted to publish his research on statistical methods using the pseudonym, 'Student'. Student's article, 'The Probable Error of a Mean' (*Biometrika*, 1908), a classic in statistics, introduced the *t* Test for small-sample statistics, and it laid much of the groundwork for **Fisher's** development of **analysis of variance** ([5, p. 167–168]; and see [3]).

Two informative articles about Gosset are "Student" as a statistician' [2] and "Student" as a man' [1]. Of Student as a statistician, Pearson concluded: '[Gosset's]... investigation published in 1908 has done more than any other single paper to bring [chemical, biological, and agricultural] subjects within the range of statistical inquiry; as it stands it has provided an essential tool for the practical worker, while on the theoretical side it has proved to contain the seed of ideas which have since grown and multiplied in hundredfold' [2, p. 224]. Of Student as a man, McMullen [1], described Gosset as being a golfer and a builder and sailor of boats of unusual design, made by preference using simple tools. He was also an accomplished fly fisherman. 'In fishing he was an efficient performer; he used to hold that only the size and general lightness or darkness of a fly were important; the blue wings, red tail, and so on being only to attract the fisherman to the shop' [1 p. 209].

References

- [1] McMullen, L. (1939). Student as a man, *Biometrika* **30**, 205–210.
- [2] Pearson, E.S. (1939). "Student" as statistician, *Biometrika* **30**, 210–250.
- [3] Pearson, E.S. (1968). Studies in the history of probability and statistics. XX: some early correspondence between

- W. S. Gosset, R. A. Fisher and Karl Pearson with notes and comments, *Biometrika* **55**, 445–457.
- [4] Pearson, E.S. (1973). Some historical reflections on the introduction of statistical methods in industry, *The Statistician* **22**, 165–179.
- [5] Rucci, A.J. & Tweney, R.D. (1980). Analysis of variance and the “second discipline” of scientific psychology: a historical account, *Psychological Bulletin* **87**, 166–184.

ROGER THOMAS

Graphical Chain Models

NANNY WERMUTH

Volume 2, pp. 755–757

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Graphical Chain Models

Graphical Markov models represent relations, most frequently among random variables, by combining simple, yet powerful, concepts: data generating processes, graphs, and conditional independence. The origins can be traced back to independent work in genetics ([30]), in physics ([10]), and in probability theory (**A. A. Markov**, 1912, [20]). Wright used directed graphs to describe processes of how his genetic data could have been generated, and to check the consistency of such hypotheses with observed data. He called his method **path analysis**. Gibbs described total energy of systems of particles by the number of nearest neighbors for nodes in undirected graphs. Markov suggested how some seemingly complex structures can sometimes be explained in terms of a chain of simple dependencies using the notion of **conditional independence** (see **Markov Chains**).

Development of these ideas continued largely independently in mathematics, physics, and engineering. In the social sciences and econometrics, an extension of path analysis was developed, called *simultaneous equation models*, which does not directly utilize the notion of conditional independence, and which does not incorporate nonlinear relations or time-dependent variation. In decision analysis, computer science, and philosophy, extensions of path analysis have been called *influence diagrams*, belief networks, or **Bayesian networks**, and are used, among others, for constructing so-called expert systems and systems with learning mechanisms.

A systematic development of graphical Markov models for representing multivariate statistical dependencies for both discrete and continuous variables started in the 1970s, with work on fully undirected graph models for purely discrete and for Gaussian random variables, and on linear models with graphs that are fully directed and have no cycles. This work was extended to models permitting sequences of response variables to be considered on equal footing, that is, without specifications of a direction of dependence. Joint responses can be modeled in quite different ways, some define independence structures of distinct types of graphical chain model. Properties of corresponding types of graph have been studied intensively, so that, in particular, all independencies,

implied by a given graph, can be derived by so-called separation criteria.

Several books give overviews of theory, analyses, and interpretations of graphical Markov models in statistics, based on developments on this work during the first few decades, see [7], [15], [2], [29], and a wide range of different applications has been reported, see for example, [11], [16]. For some compact descriptions and for references see [26], [27].

Applicability of fully directed graph models to very large systems of units has been emphasized recently, see for example, [6] and is simplified by free-source computational tools within the framework of the R-project, see [19], [18], [1].

Special extensions to **time series** have been developed ([5], [8], [9]), and it has been shown that the independence structure defined with any **structural equation model (SEM)** can be read off a corresponding graph [13]. The result does not extend to the interpretation of SEM parameters. Extensions to point processes and to multilevel models (see **Linear Multilevel Models**) are in progress. Graphical criteria for deciding on the identifiability of special linear models, including hidden variables, have been derived [23], [21], [25], [12], [24].

A new approach to studying properties of graphical Markov models is based on binary matrix forms of graphs [28]. This uses analogies between partial inversion of parameter matrices for linear systems and partial closing of directed and of undirected paths in graphs. The starting point for this are step-wise generating processes, either for systems of linear equations, or for joint distributions.

In both cases the graph consists of a set of nodes, with node i representing random variable Y_i , and a set of directed edges. Each edge is drawn as an arrow outgoing from what is called a *parent node* and pointing to an offspring node. The graph is acyclic, if it is impossible to return to any starting node by following arrows pointing in the same direction. The set of parent nodes of node i is denoted by par_i and the graph is called a *parent graph* if there is a complete ordering of the variables as $(Y_1, \dots, Y_d)$. Either a joint density is given by a recursive sequence of univariate conditional densities, or a covariance matrix is generated by a system of recursive equations.

The joint density f_N , generated over a parent graph with nodes $N = (1, \dots, d)$, and written in a compact notation for conditional densities in terms

2 Graphical Chain Models

of nodes, is

$$f_N = \prod_i f_{i|i+1,\dots,d} = \prod_i f_{i|\text{par}_i}. \quad (1)$$

The conditional independence statement $i \perp j | \text{par}_i$ is equivalent to the factorization $f_{i|\text{par}_i,j} = f_{i|\text{par}_i}$, and it is represented by a missing ij -arrow in the parent graph for $i < j$.

The joint covariance matrix Σ of mean-centered and continuous variables Y_i , generated over a parent graph with nodes $N = (1, \dots, d)$, is given by a system of linear recursive equations with uncorrelated residuals, written as

$$AY = \varepsilon, \quad (2)$$

where A is an upper-triangular matrix with unit diagonal elements, and ε is a residual vector of zero-mean uncorrelated random variables ε . A diagonal form of the residual covariance matrix $\text{cov}(\varepsilon) = \Delta$ is equivalent to specifying that each row of A in (2) defines a linear least squares regression equation ([4], p. 302) for response Y_i regressed on $Y_{i+1}, \dots, Y_d$. For the regression coefficient of Y_j in this regression, it holds for $i < j$:

$$-a_{ij} = \beta_{i|j \cdot \{i+1, \dots, d\} \setminus j} = \beta_{i|j \cdot \text{par}_i \setminus j}. \quad (3)$$

Thus, the vanishing contribution of Y_j to the linear regression of Y_i on $Y_{i+1}, \dots, Y_d$ is represented by zero value in position (i, j) in the upper triangular-part of A .

The types of question that can be answered for these generating processes are: which independencies (either linear or probabilistic not both) are preserved if the order of the variables is modified, or if some of the variables are considered as joint instead of univariate responses, or if some of variables are explicitly omitted, or if a subpopulation is selected? [28]. Joint response models that preserve exactly the independencies of the generating process after omitting some variables and conditioning on others form a slightly extended subclass of SEM models [22], [14].

Sequences of joint responses occur in different types of chain graphs. All these chain graphs have in common that the nodes are arranged in a sequence of say d_{CC} chain components g , each containing one or more nodes. For partially ordered nodes, $N = (1, \dots, g, \dots, d_{CC})$, a joint density is considered in the form

$$f_N = \prod_g f_{g|g+1, \dots, d_{CC}}. \quad (4)$$

Within this broad formulation of chain graphs, one speaks of multivariate-regression chains, whenever, for a given chain component g , variables at nodes i and j are considered conditionally, given all variables in chain components $g+1, \dots, d_{CC}$. Then the univariate and bivariate densities

$$f_{i|g+1, \dots, d_{CC}}, \quad f_{ij|g+1, \dots, d_{CC}} \quad (5)$$

determine the presence or absence of a directed ij -edge, which points to node i in chain component g from a node j in $g+1, \dots, d_{CC}$, or of an undirected ij -edge within g when j itself is in g .

The more traditional form of chain graphs results if, for a given chain component g , variables at nodes i and j are considered conditionally given all other variables in g and the variables in $g+1, \dots, d_{CC}$. Then the univariate and bivariate densities

$$f_{i|g \setminus \{i\}, g+1, \dots, d_{CC}}, \quad f_{ij|g \setminus \{i, j\}, g+1, \dots, d_{CC}} \quad (6)$$

are relevant for a directed ij -edge, which points to node i in chain component g from a node j in $g+1, \dots, d_{CC}$, as well as for an undirected ij -edge within g .

These traditional chain graphs are called *blocked-concentration* Graphs, or, sometimes, LWF (Lauritzen, Wermuth, Frydenberg) graphs. Chain graphs with the undirected components as in blocked-concentration graphs and the directed components as in multivariate regressions graphs are called *concentration-regression graphs*, or, sometimes, AMP (Andersen, Madigan, Perlman) graphs. The statistical models corresponding to the latter for purely discrete variables are the so-called marginal models. These belong to the exponential family of models and have canonical parameters for the undirected components and moment parameters for the directed components.

Stepwise generating processes in univariate responses arise both in observational and in intervention studies. With an intervention, the probability distribution is changed so that the intervening variable is decoupled from all variables in the past that relate directly to it in an observational setting, see [17]. Two main assumptions distinguish ‘causal models with potential outcomes’ (or counterfactual models) from general generating processes in univariate responses. These are (1) unit-treatment additivity, and (2) a notional intervention. These two assumptions, taken together, assure that there are no unobserved

confounders, and that there is no interactive effect on the response by an unobserved variable and the intervening variable. One consequence of these assumptions is for linear models that the effect of the intervening variable on the response averaged over past variables coincides with its conditional effects given past unobserved variables. Some authors have named this a causal effect. For a comparison of different definitions of causality from a statistical viewpoint, including many references, and for the use of graphical Markov models in this context, see [3].

References

- [1] Badsberg, J.H. (2004). *DynamicGraph: Interactive Graphical Tool for Manipulating Graphs*, URL: <http://cran.r-project.org>.
- [2] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis, and Interpretation*, Chapman & Hall, London.
- [3] Cox, D.R. & Wermuth, N. (2004). Causality a statistical view, *International Statistical Review* **72**, 285–305.
- [4] Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton.
- [5] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series, *Metrika* **51**, 157–172.
- [6] Dobra, A. (2003). Markov bases for decomposable graphical models, *Bernoulli* **9**, 1093–1108.
- [7] Edwards, D. (2000). *Introduction to Graphical Modelling*, 2nd Edition, Springer, New York.
- [8] Eichler, M., Dahlhaus, R. & Sandkühler, J. (2003). Partial correlation analysis for the identification of synaptic connections, *Biological Cybernetics* **89**, 289–302.
- [9] Fried, R. & Didelez, V. (2003). Decomposability and selection of graphical models for time series, *Biometrika* **90**, 251–267.
- [10] Gibbs, W. (1902). *Elementary Principles of Statistical Mechanics*, Yale University Press, New Haven.
- [11] Green, P.J., Hjort, N.L. & Richardson, S. (2003). *Highly Structured Stochastic Systems*, University Press, Oxford.
- [12] Grzebyk, M., Wild, P. & Chouanière, D. (2003). On identification of multi-factor models with correlated residuals, *Biometrika* **91**, 141–151.
- [13] Koster, J.T.A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors, *Scandinavian Journal of Statistics* **26**, 413–431.
- [14] Koster, J.T.A. (2002). Marginalizing and conditioning in graphical models, *Bernoulli* **8**, 817–840.
- [15] Lauritzen, S.L. (1996). *Graphical Models*, Oxford University Press, Oxford.
- [16] Lauritzen, S.L. & Sheehan, N.A. (2003). Graphical models for genetic analyses, *Statistical Science* **18**, 489–514.
- [17] Lindley, D.V. (2002). Seeing and doing: the concept of causation, *International Statistical Review* **70**, 191–214.
- [18] Marchetti, G.M. (2004). R functions for computing graphs induced from a DAG after marginalization and conditioning, *Proceedings of the American Statistical Association*, Alexandria, VA.
- [19] Marchetti, G.M. & Drton, M. (2003). *GGM: An R Package for Gaussian Graphical Models*, URL: <http://cran.r-project.org>.
- [20] Markov, A.A. (1912). *Wahrscheinlichkeitsrechnung*, (German translation of 2nd Russian edition: A.A. Markoff, ed., 1908), Teubner, Leipzig.
- [21] Pearl, J. (1998). Graph, causality and structural equation models, *Sociological Methods and Research* **27**, 226–284.
- [22] Richardson, T.S. & Spirtes, P. (2002). Ancestral Markov graphical models, *Annals of Statistics* **30**, 962–1030.
- [23] Stanghellini, E. (1997). Identification of a single-factor model using graphical Gaussian rules, *Biometrika* **84**, 241–254.
- [24] Stanghellini, E. & Wermuth, N. (2004). On the identification of path analysis models with one hidden variable, *Biometrika* To appear.
- [25] Vicard, P. (2000). On the identification of a single-factor model with correlated residuals, *Biometrika* **84**, 241–254.
- [26] Wermuth, N. (1998). Graphical Markov models, in *Encyclopedia of Statistical Sciences*, S. Kotz, C. Read & D. Banks, eds, Wiley, New York, pp. 284–300, Second Update Volume.
- [27] Wermuth, N. & Cox, D.R. (2001). Graphical models: overview, in *International Encyclopedia of the Social and Behavioral Sciences*, Vol. 9, P.B. Baltes & N.J. Smelser, eds, Elsevier, Amsterdam, pp. 6379–6386.
- [28] Wermuth, N. & Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems, *Journal of Royal Statistical Society, Series B* **66**, 687–717.
- [29] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.
- [30] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 162–177.

(See also **Markov Chain Monte Carlo and Bayesian Statistics**)

NANNY WERMUTH

Graphical Methods pre-20th Century

LAURENCE D. SMITH

Volume 2, pp. 758–762

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Graphical Methods pre-20th Century

Statistical graphs are such a prominent feature of science today that it is hard to imagine science without them [17, 20, 37]. Surveys indicate that graphs are nearly ubiquitous in journal articles of the natural sciences (although they remain far less common in the behavioral sciences) [12, 31]. Yet graphs have not always been a fixture of science: they were not used during the Scientific Revolution [14], and even a century ago, fewer than a third of published scientific papers included graphs [12]. The slow and uneven rise of statistical graphics has engaged the interest of historians, who have revealed a story of piecemeal growth in graphical methods against a backdrop of controversy and opposition to their use. As Biderman put it, graphical methods have developed in ‘waves of popularity and of neglect’ [2, p. 3].

Early Graphs

At first blush, the near-total absence of graphs during the Scientific Revolution is puzzling. Descartes’s analytical geometry had introduced the requisite framework of Cartesian coordinates in 1637, and other forms of scientific visualization were well developed by the seventeenth century. Galileo, for example, made conspicuous use of geometrical diagrams in reasoning about the laws of acceleration, and numerical tables, anatomical drawings, and maps of heaven and earth were well established in scientific publishing by the end of that century [12, p. 46]. But despite these favorable developments, the seventeenth century saw only scattered uses of graphs, none of them with lasting influence. Most of these graphs, such as Christopher Wren’s line charts of temperature or Edmund Halley’s of barometric pressure (both from the 1680s), were generated by clock-driven instruments, and were used for converting readings to tables of numbers rather than for graphical analysis [1, 34].

The systematic use of graphs – in the modern sense of figures containing axes, scales, and quantitative data – emerged only in the late eighteenth century in the work of a handful of innovators, chief

among them being Johann Lambert and William Playfair. The German natural philosopher Lambert studied phenomena of heat, producing modern-looking line graphs (*see Index Plots*) that showed, for instance, solar warming at different latitudes as a function of the seasons. He drew smooth curves to average out random errors in data and gave explicit consideration to the graphical treatment of error [14, 34]. Often considered the founder of statistical graphics, Playfair, who worked as a draftsman for a London engineering firm, published line charts and **histograms** of economic data in his *Commercial and Political Atlas* of 1786. Among those graphs was his now-famous time series plot of English exports and imports, with the difference between the curves highlighted to represent the country’s balance of trade. In the *Atlas*, Playfair also introduced the **bar graph**, which was supplemented by the **pie chart** in his 1801 *Statistical Breviary*. In both works, Playfair used graphs for purposes of analysis as well as presentation [2]. Areas below and between curves were shaded in color to draw attention to differences, and dashed lines were used to represent projected or hypothetical values. Playfair also designed graphical icons for representing multivariate data [8, 27]. In one version, he used circle charts to code the land areas of countries, and then constructed vertical tangents representing population on the left and revenues on the right. An array of such icons for a dozen or more nations conveyed at a glance the relations between a country’s relative wealth and its available resources, both human and territorial. Playfair’s clear grasp of the cognitive advantages of graphical displays made him an articulate, if sometimes immodest, promoter of graphical methods [5]. From a graph, he wrote, ‘as much information may be obtained in five minutes as would require whole days to imprint on the memory, in a lasting manner, by a table of figures’ [quoted in 8, p. 165].

Despite the foundations laid by Lambert and Playfair, there was little growth in the use of graphical methods for nearly half a century following their work [34]. The slow reception of Playfair’s innovations has been attributed to various factors: the rationalist bias, dating to Descartes, for using graphs to plot abstract equations rather than empirical data; the relative dearth of empirical data suitable for graphical treatment; the belief that graphs are suited only to the teaching of science or its popularization among nonscientists; a perceived lack of rigor in graphical

2 Graphical Methods pre-20th Century

methods; a preference for the numerical precision of tabled numbers; and a general Platonic distrust of visual images as sources of reliable knowledge. The view that statistical graphics represent a vulgar substitute for rigorous numerical methods may have been abetted by Playfair himself, who touted his graphs as a means of communicating data to politicians and businessmen. It seems likely that the disdain of many scientists for graphical methods also stemmed from the dual roots of science in academic natural philosophy and the less prestigious tradition of mechanical arts, of which Playfair was a part [5, 19]. Only when these two traditions successfully merged in the nineteenth century, combining Baconian hands-on manipulation of data with academic mathematical theory, did graphical methods achieve widespread acceptance in science. In Playfair's time, the more common response to graphs was that of the French statistician Jacques Peuchet, who dismissed graphs as 'mere by-plays of the imagination' that are foreign to the aims of science [quoted in 10, p. 295]. Such resistance to graphical methods – which never waned entirely, even during their rise to popularity during the late nineteenth century – is reflected in the fact that the first published graph of a normal distribution [6] appeared more than a century after **De Moivre** had determined the mathematical properties of the normal curve [10].

The use of graphs spread slowly through the first half of the nineteenth century, but not without significant developments. In 1811, the German polymath Alexander von Humboldt, acknowledging Playfair's precedence, published a variety of graphs in his treatise on the Americas. Displaying copious data on the geography, geology, and climate of the New World, he used line graphs as well as divided bar graphs, the latter his own invention. Humboldt echoed Playfair's judgment of the cognitive efficiency of graphs, praising their ability to 'speak to the senses without fatiguing the mind' [quoted in 3, p. 223] and defending them against the charge of being 'mere trifles foreign to science' [quoted in 10, p. 95]. In 1821, the French mathematician J. B. J. Fourier, known for his method of decomposing waveforms, used data from the 1817 Paris census to produce the first published cumulative frequency distribution, later given the name 'ogive' by **Francis Galton**. The application of graphical analysis to human data was further explored by Fourier's student **Adolphe Quetelet** in a series of publications beginning in the 1820s. These

included line charts of birth rates, mortality curves, and probability distributions fitted to histograms of empirical data [32].

Graphical methods also entered science from a different direction with the application of self-registering instruments to biological phenomena. Notable was Carl Ludwig's 1847 invention of the kymograph, which quickly came into common use as a way to make visible a range of effects that were invisible to the naked eye. Hermann von Helmholtz achieved acclaim in the 1850s by touring Europe with his *Froschcurven*, myographic records of the movements of frog muscles. These records included the graphs by which he had measured the speed of the neural impulse, one of the century's most celebrated scientific achievements and one that, as Helmholtz recognized, depended crucially on graphical methods [4, 15]. By midcentury, graphical methods had also gained the attention of philosophers and methodologists. In his influential *Philosophy of the Inductive Sciences* (1837–60), William Whewell hailed the graphical method – which he called the '*method of curves*' – as a fundamental means of discovering the laws of nature, taking its place alongside the traditional inductive methods of Bacon. Based partly on his own investigations of the tides, Whewell judged the method of curves superior to numerical methods, for 'when curves are drawn, the eye often spontaneously detects the law of recurrence in their sinuosities' [36, p. 405]. For such reasons, he even favored the graphical method over the newly developed method of least squares, which was also treated in his text.

The Golden Age of Graphics

The second half of the nineteenth century saw an unprecedented flourishing of graphical methods, leading to its designation as the Golden Age of graphics. According to Funkhouser [10], this period was marked by enthusiasm for graphs not only among scientists and statisticians but also among engineers (notably the French engineers Cheysson and Minard), government officials, and the public. The standardization imposed by the government bureaucracies of the time produced torrents of data well suited to graphical treatment [26]. Under Quetelet's leadership, a series of International Statistical Congresses from 1853 from 1876 staged massive exhibitions of graphical displays (a partial listing of the charts at the 1876

Congress cited 686 items), as well as attempts to standardize the nomenclature of graphs and the rules for their construction. The Golden Age also saw the first published systems for classifying graphical forms, as well as a proliferation of novel graphical formats. In 1857, **Florence Nightingale** produced ‘coxcomb’ plots for displaying the mortality of British soldiers across the cycle of months, a format that survives as today’s rose plots. In 1878, the Italian statistician Luigi Perozzo devised perspective plots called ‘stereograms’ in which complex relations of variables (such as probability of marriage by age and sex) were shown as three-dimensional surfaces. When the results of the ninth U.S. Census were published in 1874, they included such now-standard formats as population pyramids and bilateral frequency polygons. The 1898 report of the eleventh U.S. Census, published toward the end of the Golden Age, contained over 400 graphs and statistical maps in a wide variety of formats, many of them in color. The widespread acceptance of graphs by the end of this era was also signaled by the attention they drew from leading statisticians. During the 1890s, **Karl Pearson**, then a rising star in the world of statistics, delivered a series of lectures on graphical methods at Gresham College. In them, he treated dozens of graphical formats, challenged the ‘erroneous opinion’ that graphs are but a means of popular presentation, and described the graphical method as ‘a *fundamental method of investigating and analyzing statistical material*’ [23, p. 142, emphasis in original].

The spread of graphs among political and economic statisticians during the Golden Age was paralleled by their growing currency in the natural sciences. Funkhouser reports that graphs became ‘an important adjunct of almost every kind of scientific gathering’ [10, p. 330]. Their use was endorsed by leading scientists such as Ernst Mach and Emil du Bois-Reymond on the Continent and Willard Gibbs in America. For his part, Gibbs saw the use of graphs as central to the breakthroughs he achieved in thermodynamics; in fact, his first paper on the subject concerned the design of optimal graphs for displaying abstract physical quantities [14]. The bible of the burgeoning graphics movement was Etienne-Jules Marey’s 1878 masterwork, *La méthode graphique* [22]. This richly illustrated tome covered both statistical graphs and instrument-generated recordings, and included polemics on the cognitive and epistemological advantages of graphical methods.

It was one of many late nineteenth-century works that hailed graphs as the new *langue universelle* of science – a visual language that, true to the positivist ideals of the era, would enhance communication between scientists while neutralizing national origins, ideological biases, and disciplinary boundaries. In 1879, the young G. Stanley Hall, soon to emerge as a leading figure of American psychology, reported in *The Nation* that the graphic method – a method said to be ‘superior to all other modes of describing many phenomena’ – was ‘fast becoming the international language of science’ [13, p. 238]. Having recently toured Europe’s leading laboratories (including Ludwig’s in Leipzig), Hall also reported on the pedagogical applications of graphs he had witnessed at European universities. In an account foreshadowing today’s instructional uses of computerized graphics, he wrote that the graphical method had ‘converted the lecture room into a sort of theatre, where graphic charts are the scenery, changed daily with the theme’ [13, p. 238]. Hall himself would later make extensive use of graphs, including some sophisticated charts with multiple ordinates, in his classic two-volume work *Adolescence* (1904).

Graphs in Behavioral Science

Hall was not alone among the early behavioral scientists in making effective use of graphic methods during the Golden Age. Hermann Ebbinghaus’s 1885 classic *Memory* [7] contained charts showing the repetitions required for memorizing syllable lists as a function of list length, as well as time series graphs that revealed unanticipated periodicities in memory performance, cycles that Ebbinghaus attributed to oscillations in attention. In America, James McKeen Cattell applied graphical methods to one of the day’s pressing issues – the span of consciousness – by estimating the number of items held in awareness from the asymptotes of speed-reading curves. Cattell also analyzed psychophysical data by fitting them against theoretical curves of Weber’s law and his own square-root law, and later assessed the fit of educational data to normal distributions in the course of arguing for differential tuition fees favoring the academically proficient [24]. Cattell’s Columbia colleague Edward L. Thorndike drew heavily on graphical methods in analyzing and presenting the results of the famous puzzle-box experiments that formed

a cornerstone of later behaviorist research. His 1898 paper ‘Animal Intelligence’ [33] contained more than 70 graphs showing various conditioning phenomena and demonstrating that trial-and-error learning occurs gradually, not suddenly as implied by the theory of animal reason.

Despite such achievements, however, the master of statistical graphics among early behavioral scientists was Francis Galton. Galton gained experience with scientific visualization in the 1860s when he constructed statistical maps to chart weather patterns, work which led directly to his discovery of anticyclones. In the 1870s, he introduced the quincunx – a device that demonstrates normal distributions by funneling lead shot across an array of pins – for purposes of illustrating his lectures on heredity and to facilitate his own reasoning about sources of variation and their partitioning [32]. In the 1880s, he began to make contour plots of bivariate distributions by connecting cells of equal frequencies in tabular displays. From these plots, it was a small step to the **scatter plots** that he used to demonstrate regression and, in 1888, to determine the first numerical correlation coefficient, an achievement attained using wholly graphical means [11]. Galton’s graphical intuition, which often compensated for the algebraic errors to which he was prone, was crucial to his role in the founding of modern statistics [25]. Indeed, the ability of graphical methods to protect against numerical errors was recognized by Galton as one of its advantages. ‘It is always well,’ he wrote, ‘to retain a clear geometric view of the facts when we are dealing with statistical problems, which abound with dangerous pitfalls, easily overlooked by the unwary, while they are cantering gaily along upon their arithmetic’ [quoted in 32, p. 291].

Conclusion

By the end of the nineteenth century, statistical graphics had come a long way. Nearly all of the graphical formats in common use today had been established, the Golden Age of graphs had drawn attention to their fertility, and prominent behavioral scientists had used graphs in creative and sophisticated ways. Yet for all of these developments, the adoption of graphical methods in the behavioral sciences would proceed slowly in the following century. At the time of his lectures on graphical techniques in the 1890s, Pearson

had planned to devote an entire book to the subject. But his introduction of the chi-square test in 1900 drew his interests back to numerical methods, and this shift of interests would become emblematic of ensuing developments in the behavioral sciences. It was the inferential statistics of Pearson and his successors (notably **Gossett** and **Fisher**) that pre-occupied psychologists in the century to come [21, 28]. And while the use of hypothesis-testing statistics became nearly universal in the behavioral research of the twentieth century [16], the use of graphical methods lay fallow [9, 29]. Even with the advent of exploratory data analysis (an approach more often praised than practiced by researchers), graphical methods would continue to endure waves of popularity and of neglect, both among statisticians [8, 18, 35] and among behavioral scientists [30, 31].

References

- [1] Beniger, J.R. & Robyn, D.L. (1978). Quantitative graphics in statistics: a brief history, *American Statistician* **32**, 1–11.
- [2] Biderman, A.D. (1990). The Playfair enigma: the development of the schematic representation of statistics, *Information Design Journal* **6**, 3–25.
- [3] Brain, R.M. (1996). The graphic method: inscription, visualization, and measurement in nineteenth-century science and culture, Ph.D. dissertation, University of California, Los Angeles.
- [4] Chadarevian, S. (1993). Graphical method and discipline: self-recording instruments in nineteenth-century physiology, *Studies in the History and Philosophy of Science* **24**, 267–291.
- [5] Costigan-Eaves, P. & Macdonald-Ross, M. (1990). William Playfair (1759–1823), *Statistical Science* **5**, 318–326.
- [6] De Morgan, A. (1838). *An Essay on Probabilities and on Their Application to Life Contingencies and Insurance Offices*, Longman, Brown, Green & Longman, London.
- [7] Ebbinghaus, H. (1885/1964). *Memory: A Contribution to Experimental Psychology*, Dover Publications, New York.
- [8] Fienberg, S.E. (1979). Graphical methods in statistics, *American Statistician* **33**, 165–178.
- [9] Friendly, M. & Denis, D. (2000). The roots and branches of modern statistical graphics, *Journal de la Société Française de Statistique* **141**, 51–60.
- [10] Funkhouser, H.G. (1937). Historical development of the graphical representation of statistical data, *Osiris* **3**, 269–404.
- [11] Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data, *Proceedings of the Royal Society of London* **45**, 135–145.

- [12] Gross, A.G., Harmon, J.E. & Reidy, M. (2002). *Communicating Science: The Scientific Article from the 17th Century to the Present*, Oxford University Press, Oxford.
- [13] Hall, G.S. (1879). The graphic method, *The Nation* **29**, 238–239.
- [14] Hankins, T.L. (1999). Blood, dirt, and nomograms: a particular history of graphs, *Isis* **90**, 50–80.
- [15] Holmes, F.L. & Olesko, K.M. (1995). The images of precision: Helmholtz and the graphical method in physiology, in *The Values of Precision*, M.N. Wise, ed., Princeton University Press, Princeton, pp. 198–221.
- [16] Hubbard, R. & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects, *Educational and Psychological Measurement* **60**, 661–681.
- [17] Krohn, R. (1991). Why are graphs so central in science? *Biology and Philosophy* **6**, 181–203.
- [18] Kruskal, W. (1978). Taking data seriously, in *Toward a Metric of Science*, Y. Elkana, J. Lederberg, R. Merton, A. Thackray & H. Zuckerman, eds, Wiley, New York, pp. 139–169.
- [19] Kuhn, T.S. (1977). Mathematical versus experimental traditions in the development of physical science, in *The Essential Tension*, T.S. Kuhn, ed., University of Chicago Press, Chicago, pp. 31–65.
- [20] Latour, B. (1990). Drawing things together, in *Representation in Scientific Practice*, M. Lynch & S. Woolgar, eds, MIT Press, Cambridge, pp. 19–68.
- [21] Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934–1945, *British Journal of Mathematical and Statistical Psychology* **32**, 151–178.
- [22] Marey, E.-J. (1878). *La Méthode Graphique Dans les Sciences Expérimentales*, Masson, Paris.
- [23] Pearson, E.S. (1938). *Karl Pearson: An Appreciation of Some Aspects of His Life and Work*, Cambridge University Press, Cambridge.
- [24] Poffenberger, A.T., ed. (1947). *James McKeen Cattell, Man of Science: Vol. 1 Psychological Research*, Science Press, Lancaster.
- [25] Porter, T.M. (1986). *The Rise of Statistical Thinking, 1820–1900*, Princeton University Press, Princeton.
- [26] Porter, T.M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton.
- [27] Royston, E. (1956). Studies in the history of probability and statistics: III. A note on the history of the graphical representation of data, *Biometrika* **43**, 241–247.
- [28] Rucci, A.J. & Tweney, R.D. (1980). Analysis of variance and the “Second Discipline” of scientific psychology: an historical account, *Psychological Bulletin* **87**, 166–184.
- [29] Smith, L.D., Best, L.A., Cylke, V.A. & Stubbs, D.A. (2000). Psychology without *p* values: data analysis at the turn of the 19th century, *American Psychologist* **55**, 260–263.
- [30] Smith, L.D., Best, L.A., Stubbs, D.A., Archibald, A.B. & Roberson-Nay, R. (2002). Constructing knowledge: the role of graphs and tables in hard and soft psychology, *American Psychologist* **57**, 749–761.
- [31] Smith, L.D., Best, L.A., Stubbs, D.A., Johnston, J. & Archibald, A.M. (2000). Scientific graphs and the hierarchy of the sciences: a Latourian survey of inscription practices, *Social Studies of Science* **30**, 73–94.
- [32] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge.
- [33] Thorndike, E.L. (1898). Animal intelligence: an experimental study of the associative processes in animals, *Psychological Review Monograph Supplements* **2**, 1–109.
- [34] Tilling, L. (1975). Early experimental graphs, *British Journal for the History of Science* **8**, 193–213.
- [35] Wainer, H. & Thissen, D. (1981). Graphical data analysis, *Annual Review of Psychology* **32**, 191–241.
- [36] Whewell, W. (1847/1967). *Philosophy of the Inductive Sciences*, 2nd Edition, Frank Cass, London.
- [37] Ziman, J. (1978). *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*, Cambridge University Press, Cambridge.

(See also **Exploratory Data Analysis; Graphical Presentation of Longitudinal Data**)

LAURENCE D. SMITH

Graphical Presentation of Longitudinal Data

HOWARD WAINER AND IAN SPENCE

Volume 2, pp. 762–772

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Graphical Presentation of Longitudinal Data

Introduction

Let us begin with a few kind words about the bubonic plague. In 1538, Thomas Cromwell, the Earl of Essex (1485–1540), issued an injunction (one of 17) in the name of Henry VIII that required the registration of all christenings and burials in every English Parish. The London Company of Parish Clerks compiled weekly *Bills of Mortality* from such registers. This record of burials provided a way to monitor the incidence of plague within the city. Initially, these *Bills* were circulated only to government officials; principal among them, the Lord Mayor and members of the King's Council.

They were first made available to the public in 1594, but were discontinued a year later with the abatement of the plague. However, in 1603, when

the plague again struck London, their publication resumed on a regular basis.

The first serious analysis of the *London Bills* was done by John Graunt in 1662, and in 1710, **Dr. John Arbuthnot**, a physician to Queen Anne, published an article that used the christening data to support an argument (possibly tongue in cheek) for the existence of God. These data also provide supporting evidence for the lack of existence of statistical graphs at that time.

Figure 1 is a simple plot of the annual number of christenings in London from 1630 until 1710. As we will see in a moment, it is quite informative. The preparation of such a plot is straightforward, certainly requiring no more complex apparatus than was available to Dr. Arbuthnot in 1710. Yet, it is highly unlikely that Arbuthnot, or any of his contemporaries, ever made such a plot.

The overall pattern we see in Figure 1 is a trend over 80 years of an increasing number of christenings, almost doubling from 1630 to 1710. A number of fits and starts manifest themselves in

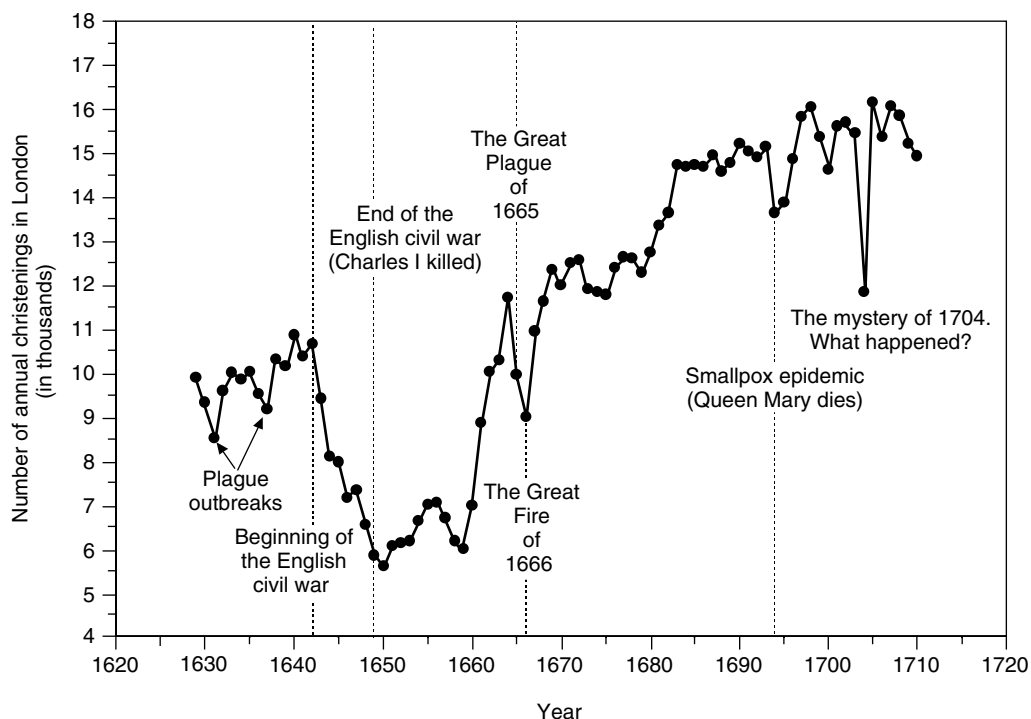


Figure 1 A plot of the annual christenings in London between 1630 and 1710 from the London Bills of Mortality. These data were taken from a table published by John Arbuthnot in 1710

2 Graphical Presentation of Longitudinal Data

substantial jiggles. Yet, each jiggle, save one, can be explained. Some of these explanations are written on the plot. The big dip that began in 1642 can only partially be explained by the onset of the English Civil War. Surely the chaos common to civil war can explain the initial drop, but the war ended in 1649 with the beheading of Charles I at Whitehall, whereas the christenings did not return to their earlier levels until 1660 (1660 marks the end of the protectorate of Oliver and Richard Cromwell and the beginning of the restoration). Graunt offered a more complex explanation that involved the distinction between births and christenings, and the likelihood that Anglican ministers would not enter children born to Catholics or Protestant dissenters into the register.

Many of the other irregularities observed are explained in Figure 1, but what about the mysterious drop in 1704? That year has about 4000 fewer christenings than one might expect from observing the adjacent data points. What happened? There was no sudden outbreak of a war or pestilence, no

great civil uprising, nothing that could explain this enormous drop.

The plot not only reveals the anomaly, it also presents a credible explanation. In Figure 2, we have duplicated the christening data and drawn a horizontal line across the plot through the 1704 data point. In doing so, we immediately see that the line goes through exactly one other point – 1674. If we went back to Arbuthnot's table, we would see that in 1674 the number of christenings of boys and girls were 6113 and 5738, exactly the same number as he had for 1704. Thus, the 1704 anomaly is likely to be a copying error! In fact, the correct figure for that year is 15 895 (8153 boys and 7742 girls), which lies comfortably between the christenings of 1703 and 1705 as expected.

It seems reasonable to assume that if Arbuthnot had noticed such an unusual data point, he would have investigated, and finding a clerical error, would have corrected it. Yet he did not. He did not, despite the fact that when graphed the error stands out, literally, like a sore thumb. Thus, we must conclude

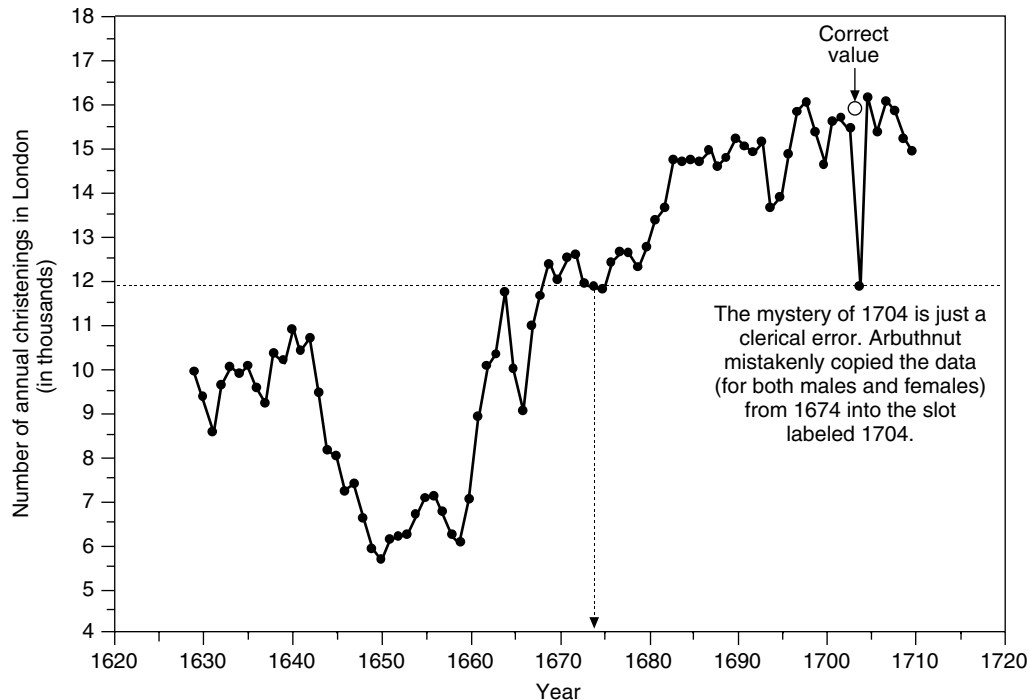


Figure 2 The solution to the mystery of 1704 is suggested by noting that only one other point (1674) had exactly the same values as the 1704 outlier. This coincidence provided the hint that allowed Zabell [11] to trace down Arbuthnot's clerical error. (Data source: Arbuthnot 1710)

that he never graphed his data. Why not? The answer, very simply, is that graphs were not yet part of the statistician's toolbox. (There were a very small number of graphical applications prior to 1710, but they were not widely circulated and Arbuthnot, a very clever and knowledgeable scientist, had likely not been familiar with them.)

The Beginnings of Graphs

Graphs are the most important tool for examining longitudinal data because they convey comparative information in ways that no table or description ever could. Trends, differences, and associations are effortlessly seen in the blink of an eye. The eye perceives immediately what the brain would take much longer to deduce from a table of numbers. This is what makes graphs so appealing – they give numbers a voice, allowing them to speak clearly. Graphs and charts not only show what numbers tell, they also help scientists tease out the critical clues from their data, much as a detective gathers clues at the scene of a crime. Graphs are truly international – a German can read the same graph that an Australian draws. There is no other form of communication that more appropriately deserves the description 'universal language.'

Who invented this versatile device? Have graphs been around for thousands of years, the work of inventors unknown? The truth is that statistical graphs were not invented in the remote past; they were not at all obvious and their creator lived only two centuries ago. He was a man of such unusual skills and experience that had he not devised and published his charts during the Age of Enlightenment we might have waited for another hundred years before the appearance of statistical graphs.

The Scottish engineer and political economist, William Playfair (1759–1823) is the principal inventor of statistical graphs. Although one may point to solitary instances of simple line graphs that precede Playfair's work (see Wainer & Velleman, [10]), such examples generally lack refinement and, without exception, failed to inspire others. In contrast, Playfair's graphs were detailed and well drawn; they appeared regularly over a period of more than 30 years; and they introduced a surprising variety of practices that are still in use today. He invented three of the four basic forms: the statistical line

graph, the **bar chart**, and the **pie chart**. The other important basic form – the **scatterplot** – did not appear until at least a half century later (some credit Herschel [4] with its first use, others believe that Herschel's plot was a time-series plot, no different than Playfair's). Playfair also invented other graphical elements, for example, the circle diagram and statistical Venn diagram; but these innovations are less widely used.

Two Time-series Line Graphs

In 1786, Playfair [5] published his *Commercial and Political Atlas*, which contained 44 charts, but no maps; all of the charts, save one, were variants of the statistical time-series line graph. Playfair acknowledged the influence of the work of Joseph Priestley (1733–1804), who had also conceived of representing time geometrically [6, 7]. The use of a grid with time on the horizontal axis was a revolutionary idea, and the representation of the lengths of reigns of monarchs by bars of different lengths allowed immediate visual comparisons that would otherwise have required significant mental arithmetic. An interesting sidelight to Priestley's plot is that he accompanied the original (1765) version with extensive explanations, which were entirely omitted in the 1769 elaboration when he realized how naturally his audience could comprehend it (Figure 3).

At about the same time that Priestley was drafting his time lines, the French physician Jacques Barbeau-Dubourg (1709–1779) and the Scottish philosopher Adam Ferguson (1723–1816) produced plots that followed a similar principle. In 1753, Dubourg published a scroll that was a complex timeline spanning the 6480 years from The Creation until Dubourg's time. This is demarked as a long thin line at the top of the scroll with the years marked off vertically in small, equal, one-year increments. Below the timeline, Dubourg laid out his record of world history. He includes the names of kings, queens, assassins, sages, and many others, as well as short phrases summarizing events of consequence. These are fixed in their proper place in time horizontally and grouped vertically either by their country of origin or in Dubourg's catch-all category at the bottom of the chart 'événements mémorables.' In 1780, Ferguson published a timeline of the birth and death of civilizations that begins at the time of the Great Flood

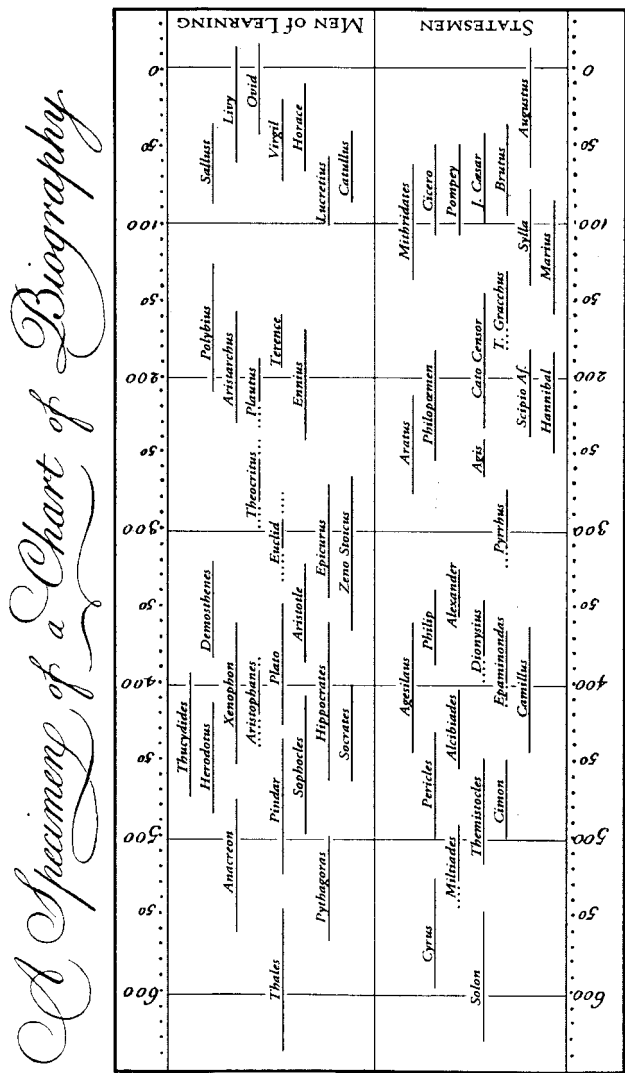


Figure 3 Lifespans of 59 famous people in the six centuries before Christ (Priestley, [6]). Its principal innovation is the use of the horizontal axis to depict time. It also uses dots to show the lack of precise information on the birth and/or death of the individual shown

(2344 BC – indicating clearly, though, that this was 1656 years after The Creation) and continued until 1780. And in 1782, the Scottish minister James Playfair (unrelated to William), published *A System of Chronology*, in the style of Priestley.

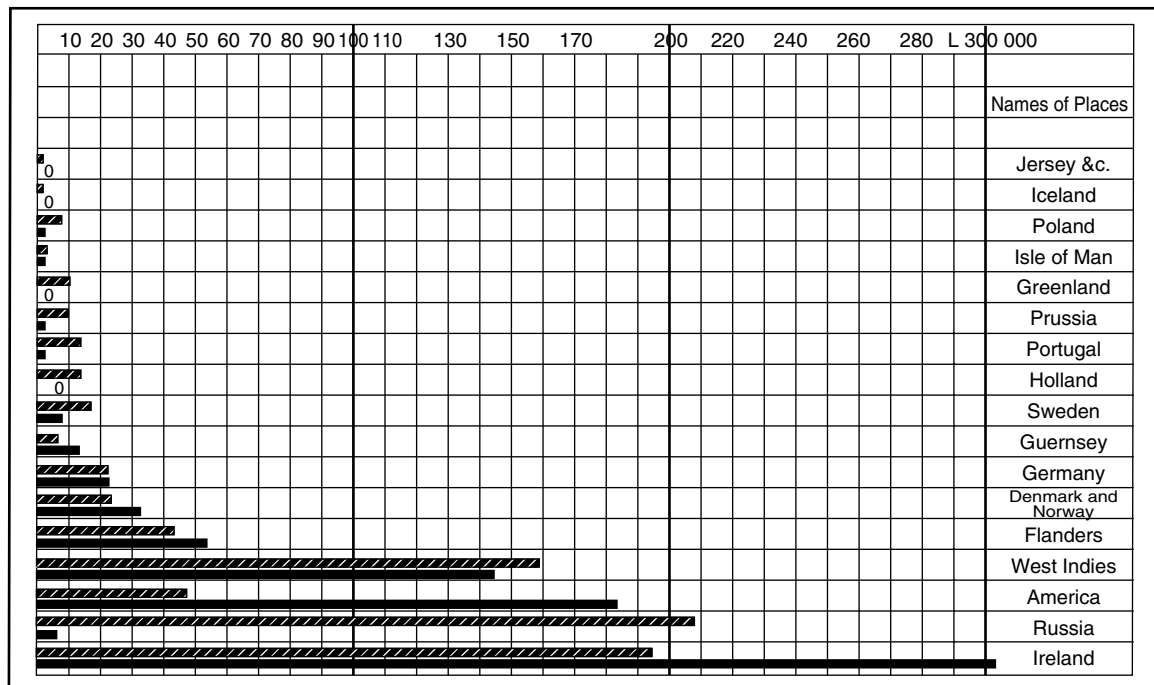
The motivation behind the drafting of graphical representations of longitudinal data remains the same today as it was in eighteenth-century France. Dubourg declared that history has two ancillary fields: geography and chronology. Of the two he believed that geography was the more developed as a means for studying history, calling it ‘lively, convenient, attractive.’ By comparison, he characterizes chronology as ‘dry, laborious, unprofitable, offering the spirit a welter of repulsive dates, a prodigious multitude of numbers which burden the memory.’ He believed that by wedding the methods of geography to the data of chronology he could make the latter as accessible as the former. Dubourg’s name for his invention *chronographie* tells a great deal about what he intended, derived as it is from the Greek *chronos*

(time) and *grapheikos* (writing). Dubourg intended to provide the means for chronology to be a science that, like geography, speaks to the eyes and the imagination, ‘a picture moving and animated.’

Joseph Priestley used his line chart to depict the life spans of famous figures from antiquity; Pythagoras, Socrates, Pericles, Livy, Ovid, and Augustus, all found their way onto Priestley’s plot. Priestley’s use of this new tool was clearly in the classical tradition.

Twenty-one years later, William Playfair used a variant on the same form (See Figure 4) to show the extent of imports and exports of Scotland to 17 other places. Playfair, as has been amply documented (Spence & Wainer, [8]), was an iconoclast and a versatile borrower of ideas who could readily adapt the chronological diagram to show economic data; in doing so, he invented the bar chart. Such unconventional usage did not occur to his more conservative peers in Great Britain, or on the Continent. He had previously done something equally

Exports and imports of Scotland to and from different parts for one year from Christmas 1780 to Christmas 1781



The Upright Divisions are ten thousand pounds each. The black lines are Exports, the ribbed lines, imports.

Published as the Act directs June 7th 1786 by W^m. Playfair

Neele sculp^t 352 strand London

Figure 4 Imports from and exports to Scotland for 17 different places (after Playfair, [5], plate 23)

6 Graphical Presentation of Longitudinal Data

novel when he adapted the line graph, which was becoming popular in the natural sciences, to display economic time series. However, Playfair did not choose to adapt Priestley's chronological diagram because of any special affection for it, but rather of necessity, since he lacked the time-series data he needed to show what he wanted. He would have preferred a line chart similar to the others in his *Atlas*. In his own words,

'The limits of this work do not admit of representing the trade of Scotland for a series of years, which, in order to understand the affairs of that country, would be necessary to do. Yet, though they cannot be represented at full length, it would be highly blameable entirely to omit the concerns of so considerable a portion of this kingdom.'

Playfair's practical subject matter provides a sharp contrast to the classical content chosen by Priestley to illustrate his invention.

In 1787, shortly after publishing the *Atlas*, Playfair moved to Paris. Thomas Jefferson spent five years as ambassador to France (from 1784 until 1789). During that time, he was introduced to Playfair personally [2], and he was certainly familiar with his graphical inventions. One of the most important influences on Jefferson at William and Mary College in Virginia was his tutor, Dr. William Small, a Scots teacher of mathematics and natural philosophy – Small was Jefferson's only teacher

during most of his time as a student. From Small, Jefferson received both friendship and an abiding love of science. Coincidentally, through his friendships with James Watt and John Playfair, Small was responsible for introducing the 17-year-old William Playfair to James Watt, with the former serving for three years as Watt's assistant and draftsman in Watt's steam engine business in Birmingham, England.

Although Jefferson was a philosopher whose vision of democracy helped shape the political structure of the emerging American nation, he was also a farmer, a scientist, and a revolutionary whose feet were firmly planted in the American ethos. So it is not surprising that Jefferson would find uses for graphical displays that were considerably more down to earth than the life spans of heroes from classical antiquity. What is surprising is that he found time, while President of the United States, to keep a keen eye on the availability of 37 varieties of vegetables in the Washington market and compile a chart of his findings (a detail of which is shown in Figure 5).

When Playfair had longitudinal data, he made good use of them, producing some of the most beautiful and informative graphs of such data ever made. Figure 6 is one remarkable example of these. Not only is it the first 'skyrocketing government debt' chart but it also uses the innovation of an irregularly

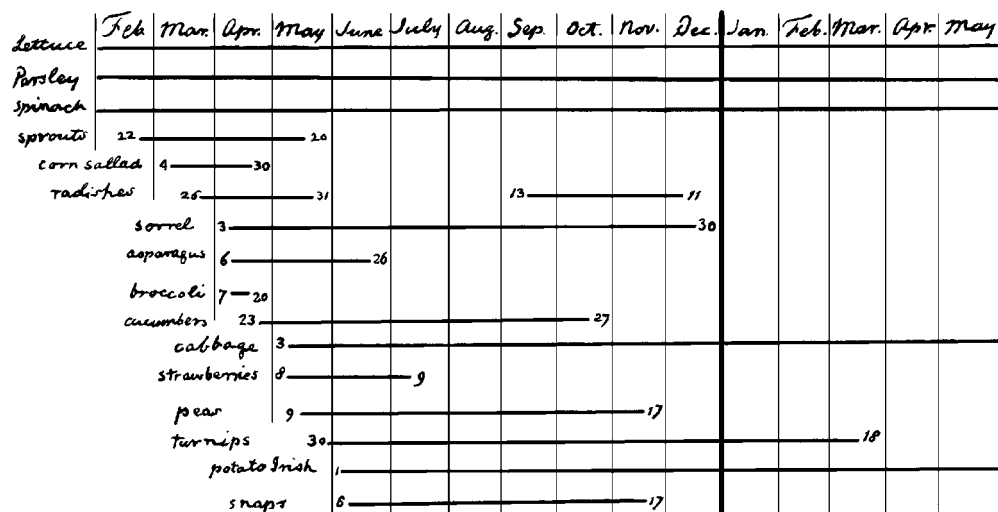


Figure 5 An excerpt from a plot by Thomas Jefferson showing the availability of 16 vegetables in the Washington market during 1802. This figure is reproduced, with permission, from Froncek ([3], p. 101)

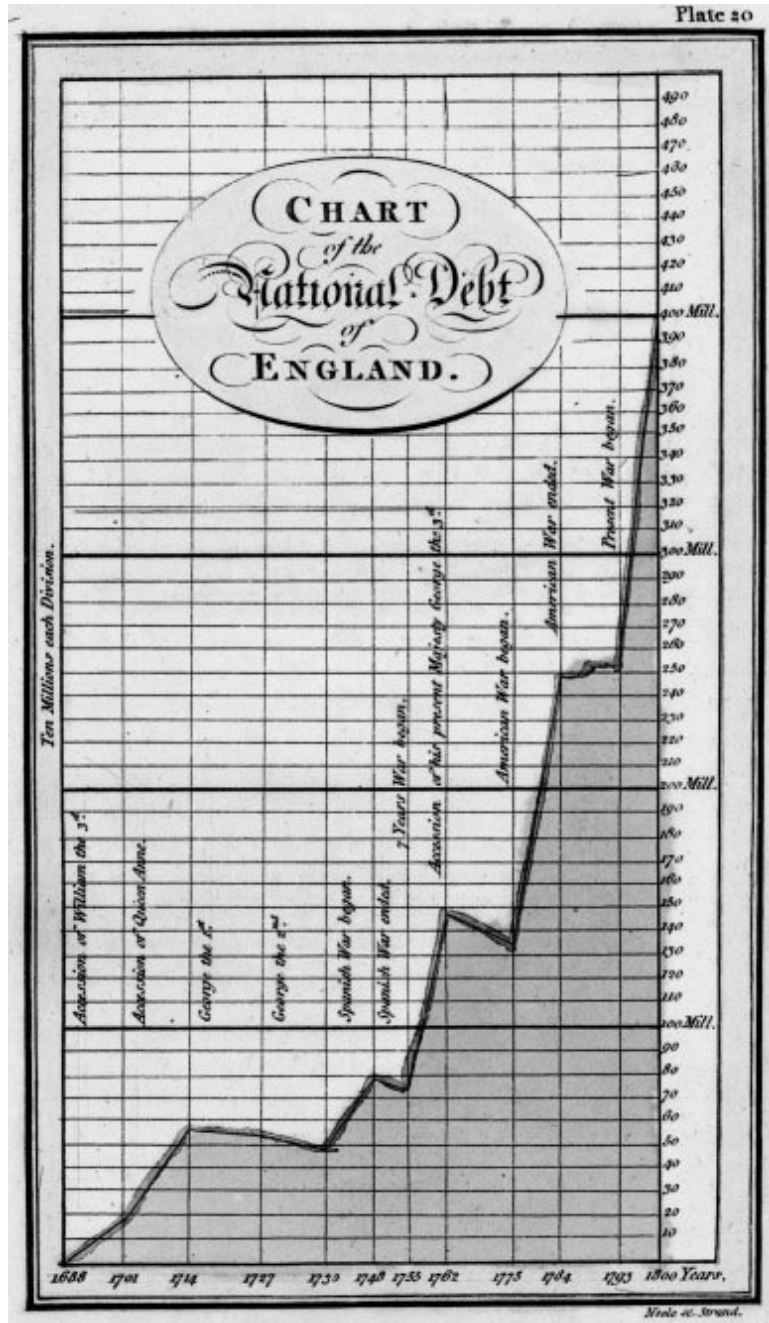


Figure 6 This remarkable 'Chart of the National Debt of England' appeared as plate 20, opposite page 83 in the third edition of Playfair's *Commercial and Political Atlas* in 1801

8 Graphical Presentation of Longitudinal Data

spaced grid along the time axis to demark events of important economic consequence.

Modern Developments

Recent developments in displaying longitudinal data show remarkably few modifications to what was developed more than 200 years ago, fundamentally because Playfair got it right. Modern high-speed computing allows us to make more graphs faster, but they are not, in any important way, different from those Playfair produced. One particularly useful modern example (Figure 7) is taken from Diggle, Heagerty, Liang & Zeger ([1], p. 37–38), which is a hybrid plot combining a scatterplot with a line drawing. The data plotted are the number of CD4+ cells found in HIV positive individuals over time. (CD4+ cells orchestrate the body's immunoresponse to infectious agents. HIV attacks this cell and so keeping track of the number of CD4+ cells allows us to monitor the progress of the disease.) Figure 7 contains the

longitudinal data (*see Longitudinal Data Analysis*) from 100 HIV positive individuals over a period that begins about two years before HIV was detectable (seroconversion) and continues for four more years. If the data were to be displayed as a **scatterplot**, the time trend would not be visible because we have no idea of which points go with which. But (Figure 7(a)) if we connect all the dots together appropriately, the graph is so busy that no pattern is discernable. Diggle et al. [1] propose a compromise solution in which the data from a small, randomly chosen, subset of subjects are connected (Figure 7(b)). This provides a guide to the eye of the general shape of the longitudinal trends. Other similar schemes are obviously possible: for example, fitting a function to the aggregate data and connecting the points for some of the residuals to look for idiosyncratic trends.

A major challenge of data display is how to represent multidimensional data on a two-dimensional surface (*see Multidimensional Scaling; Principal Component Analysis*). When longitudinal data are

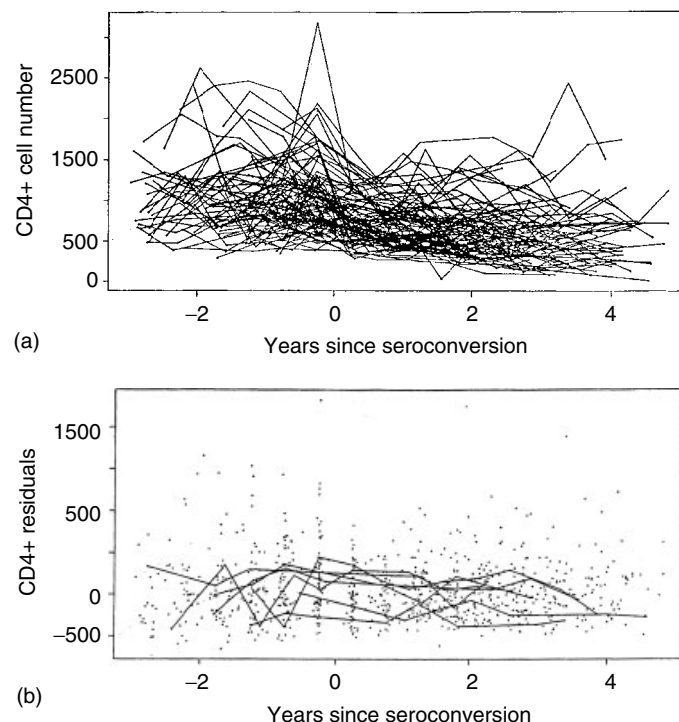


Figure 7 Figures 3.4 and 3.5 from Diggle et al., [1] – reprinted with permission (p. 37–38), showing CD4+ counts against time since seroconversion, with sequences of data on each subject connected (a) or connecting only a randomly selected subset of subjects (b)

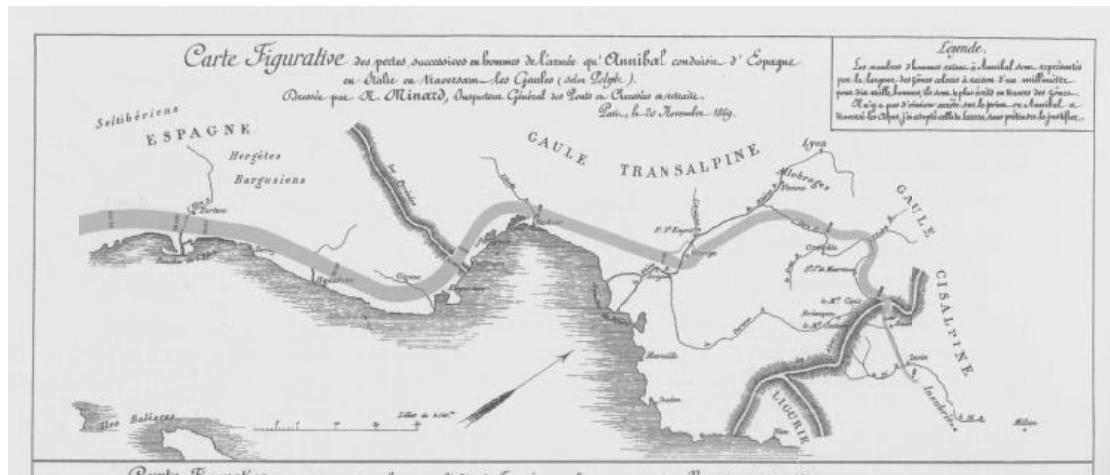


Figure 8 An 1869 plot by Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard, 1845–1869* depicting the size of Hannibal's Army as it crossed from Spain to Italy in his ill-fated campaign in the Second Punic War (218–202 bc). A portfolio of Minard's work is held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris. This figure was reproduced from Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut © 1983, 2001), p. 176. with permission

themselves multivariate (see **Multivariate Analysis: Overview**), this is a problem that has few completely satisfying solutions. Interestingly, we must look back more than a century for the best of these. In 1846, the French civil engineer Charles Joseph Minard (1781–1870) developed a format to show longitudinal data on a geographic background. He used a metaphorical data river flowing across the landscape tied to a timescale. The river's width was proportional to the amount of materials being depicted (e.g., freight, immigrants), flowing from one geographic region to another. He used this almost exclusively to portray the transport of goods by water or land. This metaphor was employed to perfection in his 1869 graphic (Figure 8), in which, through the substitution of soldiers for merchandise, he was able to show the catastrophic loss of life in Napoleon's ill-fated Russian campaign. The rushing river of 422 000 men that crossed into Russia when compared with the returning trickle of 10 000 'seemed to defy the pen of the historian by its brutal eloquence.' This now-famous display has been called (Tufte, [9]) 'the best graph ever produced.' Minard paired his Napoleon plot with a parallel one depicting the loss of life in the Carthaginian general Hannibal's ill-fated crossing of the Alps in the Second Punic War. He began his campaign in 218 BC

in Spain with more than 97 000 men. His bold plan was to traverse the Alps with elephants and surprise the Romans with an attack from the north, but the rigors of the voyage reduced his army to only 6000 men. Minard's beautiful depiction shows the Carthaginian river that flowed across Gaul being reduced to a trickle by the time they crossed the Alps. This chart has been less often reproduced than Napoleon's march and so we prefer to include it here.

Note

1. This exposition is heavily indebted to the scholarly work of Sandy Zabell, to whose writings the interested reader is referred for a much fuller description (Zabell, [11, 12]). It was Zabell who first uncovered Arbuthnot's clerical error.

References

- [1] Diggle, P.J., Heagerty, P.J., Liang, K.Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Clarendon Press, Oxford, pp. 37–38.
- [2] Donnan, D.F. (1805). *Statistical account of the United States of America*. Messrs Greenland and Norris, London.
- [3] Froncek, T. (1985). *An Illustrated History of the City of Washington*, Knopf, New York.

- [4] Herschel, J.F.W. (1833). On the investigation of the orbits of revolving double stars, *Memoirs of the Royal Astronomical Society* **5**, 171–222.
- [5] Playfair, W. (1786). *The Commercial and Political Atlas*, Corry, London.
- [6] Priestley, J. (1765). *A Chart of History*, London.
- [7] Priestley, J. (1769). *A New Chart of History*, London. Reprinted: 1792, Amos Doolittle, New Haven.
- [8] Spence, I. & Wainer, H. (1997). William Playfair: a daring worthless fellow, *Chance* **10**(1), 31–34.
- [9] Tufte, E.R. (2001). *The Visual Display of Quantitative Information*, 2nd Edition, Graphics Press, Cheshire.
- [10] Wainer, H. & Velleman, P. (2001). Statistical graphics: mapping the pathways of science, *The Annual Review of Psychology* **52**, 305–335.
- [11] Zabell, S. (1976). Arbuthnot, Heberden and the Bills of Mortality, Technical Report #40, Department of Statistics, The University of Chicago, Chicago, Illinois.
- [12] Zabell, S. & Wainer, H. (2002). A small hurrah for the black death, *Chance* **15**(4), 58–60.

Further Reading

- Ferguson, S. (1991). The 1753 *carte chronographique* of Jacques Barbeu-Dubourg, *Princeton University Library Chronicle* **52**(2), 190–230.
- Playfair, W. (1801). *The Commercial and Political Atlas*, 3rd Edition, John Stockdale, London.
- Spence, I. & Wainer, H. (2004). William Playfair and the invention of statistical graphs, in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, ed., Academic Press, San Diego.
- Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*, University Press, Princeton.

HOWARD WAINER AND IAN SPENCE

Growth Curve Modeling

JUDITH D. SINGER AND JOHN B. WILLETT

Volume 2, pp. 772–779

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Growth Curve Modeling

‘Time is the one immaterial object which we cannot influence – neither speed up nor slow down, add to nor diminish.’ Maya Angelou

A growth curve model – also known as an *individual growth model* or a *multilevel model for change* – describes how a continuous outcome changes systematically over time. Generations of behavioral scientists have been interested in modeling change, but for decades, the prevailing view was that it was impossible to model change well [2]. During the 1980s, however, methodologists working within a variety of different disciplines developed a class of appropriate methods – known variously as random coefficient models, multilevel models, mixed models, linear mixed effects models and hierarchical linear models – that permit the modeling of change (see **Linear Multilevel Models**). Today we know that it is possible to model change, and do it well, as long as you have longitudinal data [7, 11] (see **Longitudinal Data Analysis**).

Growth curve models can be fit to many different types of longitudinal data sets. The research design can be experimental or observational, prospective or retrospective. Time can be measured in whatever units make sense – from seconds to years, sessions to semesters. The data collection schedule can be fixed (everyone has the same periodicity) or flexible (each person has a unique schedule); the number of waves of data per person can be identical or varying. And do not let the term ‘growth model’ fool you – these models are appropriate for outcomes that *decrease* over time (e.g., weight loss among dieters) or exhibit complex trajectories (including plateaus and reversals).

Perhaps the most intuitively appealing way of postulating a growth curve model is to link it to two distinct questions about change, each arising from a specific level in a natural hierarchy:

- At level-1 – the within-person level – we ask about each person’s individual change trajectory. How does the outcome change over time? Does it increase, decrease, or remain steady? Is change linear or nonlinear?
- At level-2 – the between-person (or interindividual) level – we ask about predictors that might

explain differences among individual’s change trajectories. On average, do men’s and women’s trajectories begin at the same level? Do they have the same rates of change?

These two types of questions – the former within-person and the latter between-persons – lead naturally to the specification of two sets of statistical models that together form the overall multilevel model for change (see **Linear Multilevel Models**).

We illustrate ideas using three waves of data collected by researchers tracking the cognitive performance of 103 African-American infants born into low-income families in the United States [1]. When the children were 6 months old, approximately half ($n = 58$) were randomly assigned to participate in an intensive early intervention program designed to enhance their cognitive performance; the other half ($n = 45$) received no intervention and constituted a control group. Here, we examine the effects of program participation on changes in cognitive performance as measured by a nationally normed test administered three times, at ages 12, 18 and 24 months.

In the left-hand panel of Figure 1, we plot the cognitive performance (COG) of one child in the control group versus his age (rescaled here in years). Notice the downward trend, which we summarize – rather effectively – using an ordinary least squares (OLS) linear regression line (see **Multiple Linear Regression**). Especially when you have few waves of data, it is difficult to argue for anything except a linear within-person model. When examining Figure 1, also note that the hope that we would be assessing whether program participants experience a faster rate of *growth* is confronted with the reality that we may instead be assessing whether they experience a slower rate of *decline*.

The Level-1 Model

The level-1 model represents the change we expect each member of the population to experience during the time period under study (here, the second year of life). Assuming that change is a linear function of age, a reasonable level-1 model is:

$$Y_{ij} = [\pi_{0i} + \pi_{1i}(AGE_{ij} - 1)] + \varepsilon_{ij} \quad (1)$$

This model asserts that, in the population from which this sample was drawn, Y_{ij} , the value of COG

2 Growth Curve Modeling

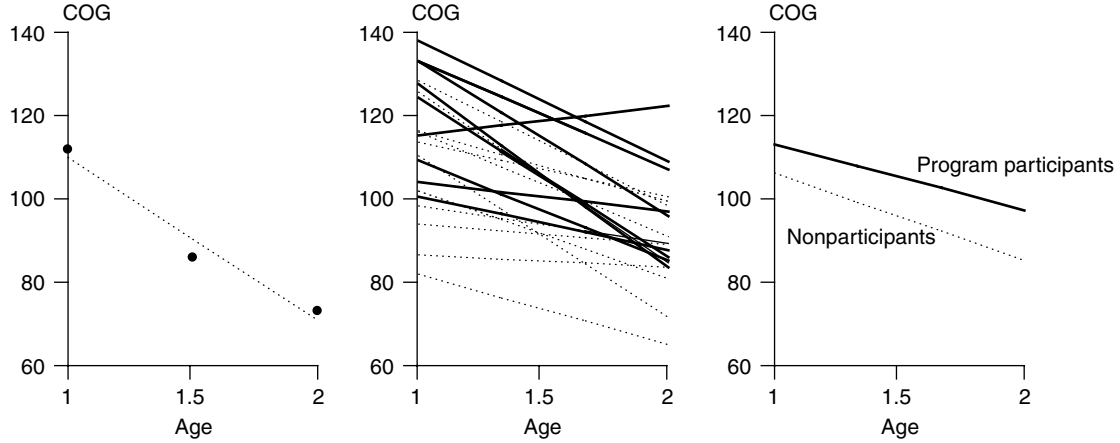


Figure 1 Developing a growth curve model using data on cognitive performance over time. The left-hand panel plots the cognitive performance (COG) of one child in the control group versus his age (rescaled here in years). The middle panel presents fitted OLS trajectories for a random subset of 28 children (coded using solid lines for program participants and dashed lines for nonparticipants). The right-hand panel presents fitted change trajectories for program participants and nonparticipants

for child i at time j , is a linear function of his (or her) age on that occasion (AGE_{ij}). The model assumes that a straight line adequately represents each person's true change trajectory and that any deviations from linearity in sample data result from random error (ε_{ij}). Although everyone in this dataset was assessed on the same three occasions (ages 1.0, 1.5, and 2.0), this basic level-1 model can be used in a wide variety of datasets, even those in which the timing and spacing of waves varies across people.

The brackets in (1) identify the model's structural component, which represents our hypotheses about each person's true trajectory of change over time. The model stipulates that this trajectory is linear in age and has individual growth parameters, π_{0i} and π_{1i} , which characterize its shape for the i th child in the population. If the model is appropriate, these parameters represent fundamental features of each child's true growth trajectory, and as such, become the objects of prediction when specifying the linked level-2 model.

An important feature of the level-1 specification is that the researcher controls the substantive meaning of these parameters by choosing an appropriate metric for the temporal predictor. For example, in this level-1 model, the intercept, π_{0i} , represents child i 's true cognitive performance at age 1. This interpretation accrues because we *centered* AGE in the level-1 model using the predictor $(AGE - 1)$. Had we not centered

age, π_{0i} would represent child i 's true value of Y at age 0, which is meaningless and predates the onset of data collection. Centering time on the first wave of data collection is a popular approach because it allows us to interpret π_{0i} using simple nomenclature: it is child i 's true initial status, his or her true status at the beginning of the study.

The more important individual growth parameter is the slope, π_{1i} , which represents the rate at which individual i changes over time. By clocking age in years (instead of the original metric of months), we can adopt the simple interpretation that π_{1i} represents child i 's true annual rate of change. During the single year under study – as child i goes from age 1 to 2 – his trajectory rises by π_{1i} . Because we hypothesize that each individual in the population has his (or her) own rate of change, this growth parameter has the subscript i .

In specifying a level-1 model, we implicitly assume that all the true individual change trajectories in the population have a common algebraic form. But because each person has his or her own individual growth parameters, we do not assume that everyone follows the same trajectory. The level-1 model allows us to distinguish the trajectories of different people using just their individual growth parameters. This leap is the cornerstone of growth curve modeling because it means that we can study interindividual

differences in growth *curves* by studying interindividual variation in growth *parameters*. This allows us to recast vague questions about the relationship between ‘change’ and predictors as specific questions about the relationship between the individual growth parameters and predictors.

The Level-2 Model

The level-2 model codifies the relationship between interindividual differences in the change trajectories (embodied by the individual growth parameters) and time-invariant characteristics of individuals. To develop an intuition for this model, examine the middle panel of Figure 1, which presents fitted OLS trajectories for a random subset of 28 children in the study (coded using solid lines for program participants and dashed lines for nonparticipants). As noted for the one child in the left panel, cognitive performance (on this age-standardized scale) tends to decline over time. In addition, program participants have generally higher scores at age 1 and decline less precipitously over time. This suggests that their intercepts are higher but their slopes are shallower. Also note the substantial interindividual heterogeneity *within* groups. Not all participants have higher intercepts than nonparticipants; not all nonparticipants have steeper slopes. Our level-2 model must simultaneously account for both the general patterns (the between-group differences in intercepts and slopes) *and* interindividual heterogeneity in patterns within groups.

This suggests that an appropriate level-2 model will have four specific features. First, the level-2 outcomes will be the level-1 individual growth parameters (here, π_{0i} and π_{1i} from (1)). Second, the level-2 model must be written in separate parts, one distinct model for each level-1 growth parameter; (here, π_{0i} and π_{1i}). Third, each part must specify a relationship between the individual growth parameter and the predictor (here, PROGRAM, which takes on only two values, 0 and 1). Fourth, each model must allow individuals who share common predictor values to vary in their individual change trajectories. This means that each level-2 model must allow for stochastic variation (also known as random variation) in the individual growth parameters.

These considerations lead us to postulate the following level-2 model:

$$\begin{aligned}\pi_{0i} &= \gamma_{00} + \gamma_{01} \text{PROGRAM} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} \text{PROGRAM} + \zeta_{1i}\end{aligned}\quad (2)$$

Like all level-2 models, (2) has more than one component; taken together, they treat the intercept (π_{0i}) and the slope (π_{1i}) of an individual’s growth trajectory as level-2 outcomes that may be associated with identified predictors (here, PROGRAM). As in regular regression, we can modify the level-2 model to include other predictors, adding, for example, maternal education or family size. Each component of the level-2 model also has its own residual – here, ζ_{0i} and ζ_{1i} – that permits the level-1 parameters (the π ’s) to differ across individuals.

The structural parts of the level-2 model contain four level-2 parameters – γ_{00} , γ_{01} , γ_{10} , and γ_{11} – known collectively as the *fixed effects*. The fixed effects capture systematic interindividual differences in change trajectories according to values of the level-2 predictor(s). In (2), γ_{00} and γ_{10} are known as level-2 intercepts; γ_{01} and γ_{11} are known as level-2 slopes. As in regular regression, the slopes are of greater interest because they represent the effect of predictors (here, the effect of PROGRAM) on the individual growth parameters. You interpret the level-2 parameters much like regular regression coefficients, except that they describe variation in ‘outcomes’ that are level-1 individual growth parameters. For example, γ_{00} represents the average true initial status (cognitive score at age 1) for nonparticipants, while γ_{01} represents the hypothesized *difference* in average true initial status between groups. Similarly, γ_{10} represents the average true annual rate of change for nonparticipants, while γ_{11} represents the hypothesized difference in average true annual rate of change between groups. The level-2 slopes, γ_{01} and γ_{11} , capture the effects of PROGRAM. If γ_{01} and γ_{11} are nonzero, the average population trajectories in the two groups differ; if they are both 0, they are the same. The two level-2 slope parameters therefore address the question: What is the difference in the average trajectory of true change associated with program participation?

An important feature of both the level-1 and level-2 models is the presence of stochastic terms – ε_{ij} at level-1, ζ_{0i} and ζ_{1i} at level-2 – also known as residuals. In the level-1 model, ε_{ij} accounts for the difference between an individual’s true and observed trajectory. For these data, each level-1 residual represents that part of child i ’s value of COG at time j not predicted by his (or her) age. The level-2 residuals, ζ_{0i} and ζ_{1i} , allow each person’s individual growth parameters to be scattered around their relevant population averages. They represent those

4 Growth Curve Modeling

portions of the outcomes – the individual growth parameters – that remain ‘unexplained’ by the level-2 predictor(s). As is true of most residuals, we are usually less interested in their specific values than in their variance. The level-1 residual variance, σ_ε^2 , summarizes the scatter of the level-1 residuals around each person’s true change trajectory. The level-2 residual variances, σ_0^2 and σ_1^2 , summarize the variation in true individual intercept and slope around the average trajectories *left over* after accounting for the effect(s) of the model’s predictor(s). As a result, these level-2 residual variances are *conditional* residual variances. Conditional on the model’s predictors, σ_0^2 represents the population residual variance in true initial status and σ_1^2 represents the population residual variance in true annual rate of change. The level-2 variance components allow us to address the question: How much heterogeneity in true change remains after accounting for the effects of program participation?

But there is another complication at level-2: might there be an association between individual initial status and individual rates of change? Children who begin at a higher level may have higher (or lower) rates of change. To account for this possibility, we permit the level-2 residuals to be correlated. Since ζ_{0i} and ζ_{1i} represent the deviations of the individual growth parameters from their population averages, their population covariance, σ_{01} , summarizes the association between true individual intercepts and slopes. Again because of their conditional nature, the population covariance of the level-2 residuals, σ_{01} , summarizes the magnitude and direction of the association between true initial status and true annual rate of change, controlling for program participation. This parameter allows us to address the question: Controlling for program participation, are true initial status and true rate of change related?

To fit the model to data, we must make some distributional assumptions about the residuals. At level-1, the situation is relatively simple. In the absence of evidence suggesting otherwise, we usually invoke the classical normality assumption, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. At level-2, the presence of two (or sometimes more) residuals necessitates that we describe their underlying behavior using a **bivariate (or multivariate) distribution**:

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right) \quad (3)$$

The complete set of residual variances and covariances – both the level-2 error variance-covariance matrix and the level-1 residual variance, σ_ε^2 – is known as the model’s **variance components**.

The Composite Growth Curve Model

The level-1/level-2 representation is not the only specification of a growth curve model. A more parsimonious representation results if you collapse the level-1 and level-2 models together into a single *composite* model. The composite representation, while identical to the level-1/level-2 specification mathematically, provides an alternative way of codifying hypotheses and is the specification required by many multilevel statistical software programs.

To derive the composite specification – also known as the *reduced form* growth curve model – notice that any pair of linked level-1 and level-2 models share some common terms. Specifically, the individual growth parameters of the level-1 model are the outcomes of the level-2 model. We can therefore collapse the submodels together by substituting for π_{0i} and π_{1i} from the level-2 model in (2) into the level-1 model (in (1)). Substituting the more generic temporal predictor $TIME_{ij}$ for the specific predictor ($AGE_{ij}-1$), we write:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\ &= (\gamma_{00} + \gamma_{01} PROGRAM_i + \zeta_{0i}) \\ &\quad + (\gamma_{10} + \gamma_{11} PROGRAM_i + \zeta_{1i})TIME_{ij} + \varepsilon_{ij} \end{aligned} \quad (4)$$

Multiplying out and rearranging terms yields the *composite model*:

$$\begin{aligned} Y_{ij} &= [\gamma_{00} + \gamma_{10} TIME_{ij} + \gamma_{01} PROGRAM_i \\ &\quad + \gamma_{11}(PROGRAM_i \times TIME_{ij})] \\ &\quad + [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}] \end{aligned} \quad (5)$$

where we once again use brackets to distinguish the model’s structural and stochastic components.

Even though the composite specification in (5) appears more complex than the level-1/level-2 specification, the two forms are logically and mathematically equivalent. The level-1/level-2 specification is often more substantively appealing; the composite specification is algebraically more parsimonious. In addition, the γ ’s in the composite model describe patterns of change in a different way. Rather than

postulating first how COG is related to TIME and the individual growth parameters, and second how the individual growth parameters are related to PROGRAM, the composite specification postulates that COG depends *simultaneously* on: (a) the level-1 predictor, TIME; (b) the level-2 predictor, PROGRAM, and (c) the *cross-level* interaction, PROGRAM $\times$ TIME. From this perspective, the composite model's structural portion strongly resembles a regular regression model with predictors, TIME and PROGRAM, appearing as main-effects (associated with γ_{10} and γ_{01} respectively) and in a *cross-level* interaction (associated with γ_{11}).

How did this cross-level interaction arise, when the level-1/level-2 specification appears to have no similar term? Its appearance arises from the 'multiplying-out' procedure used to generate the composite model. When we substitute the level-2 model for π_{1i} into its appropriate position in the level-1 model, the parameter γ_{11} , previously associated only with PROGRAM, gets multiplied by TIME. In the composite model, then, this parameter becomes associated with the interaction term, PROGRAM $\times$ TIME. This association makes sense if you consider the following logic. When γ_{11} is nonzero in the level-1/level-2 specification, the *slopes* of the change trajectories differ according to values of PROGRAM. Stated another way, the effect of TIME (whose effect is represented by the slopes of the change trajectories) differs by levels of PROGRAM. When the effects of one predictor (here, TIME) differ by the levels of another predictor (here, PROGRAM), we say that the two predictors *interact*. The cross-level interaction in the composite specification codifies this effect.

Another distinctive feature of the composite model is its 'composite residual', the three terms in the second set of brackets on the right side of (5) that combine together the one level-1 residual and the two level-2 residuals:

$$\text{Composite residual: } [\zeta_{0i} + \zeta_{1i} \text{ TIME}_{ij} + \varepsilon_{ij}]$$

Although the constituent residuals have the same meaning under both representations, the composite residual provides valuable insight into our assumptions about the behavior of residuals over time. Instead of being a simple sum, the second level-2 residual, ζ_{1i} , is multiplied by the level-1 predictor, TIME. Despite its unusual construction, the interpretation of the composite residual is straightforward: it describes the difference between the

observed and predicted value of Y for individual i on occasion j .

The mathematical form of the composite residual reveals two important properties about the occasion-specific residuals not readily apparent in the level-1/level-2 specification: they can be both *autocorrelated* and *heteroscedastic* within person. These are exactly the kinds of properties that you would expect among residuals for repeated measurements of a changing outcome.

When residuals are heteroscedastic, the unexplained portions of each person's outcome have unequal variances across occasions of measurement. Although heteroscedasticity has many roots, one major cause is the effects of omitted predictors – the consequences of failing to include variables that are, in fact, related to the outcome. Because their effects have nowhere else to go, they bundle together, by default, into the residuals. If their impact differs across occasions, the residual's magnitude may differ as well, creating heteroscedasticity. The composite model allows for heteroscedasticity via the level-2 residual ζ_{1i} . Because ζ_{1i} is multiplied by TIME in the composite residual, its magnitude can differ (linearly, at least, in a linear level-1 sub-model) across occasions. If there are systematic differences in the *magnitudes* of the composite residuals across occasions, there will be accompanying differences in residual *variance*, hence heteroscedasticity.

When residuals are autocorrelated, the unexplained portions of each person's outcome are correlated with each other across repeated occasions. Once again, omitted predictors, whose effects are bundled into the residuals, are a common cause. Because their effects may be present identically in each residual over time, an individual's residuals may become linked across occasions. The presence of the time-invariant ζ_{0i} 's and ζ_{1i} 's in the composite residual of (5) allows the residuals to be autocorrelated. Because they have only an 'i' subscript (and no 'j'), they feature identically in each individual's composite residual on every occasion, allowing for autocorrelation across time.

Fitting Growth Curve Models to Data

Many different software programs can fit growth curve models to data. Some are specialized packages written expressly for this purpose (e.g., HLM,

6 Growth Curve Modeling

Table 1 Results of fitting a growth curve for change to data ($n = 103$). This model predicts cognitive functioning between ages 1 and 2 years as a function of AGE-1 (at level-1) and PROGRAM (at level-2)

		Parameter	Estimate	Age	z
Fixed effects					
Initial status, π_{0i}	Intercept	γ_{00}	107.84***	2.04	52.97
	PROGRAM	γ_{01}	6.85*	2.71	2.53
Rate of change, π_{1i}	Intercept	γ_{10}	-21.13***	1.89	-11.18
	PROGRAM	γ_{11}	5.27*	2.52	2.09
Variance components					
Level-1:	Within-person, ε_{ij}	σ_{ε}^2	74.24***		
Level-2:	In initial status, ζ_{0i}	σ_0^2	124.64***		
	In rate of change, ζ_{1i}	σ_1^2	12.29		
	Covariance between ζ_{0i} and ζ_{1i}	σ_{01}	-36.41		

* $p < .05$, ** $p < .01$, *** $p < .001$

Note: Full ML, HLM

MLwiN, and MIXREG). Others are part of popular multipurpose software packages including SAS (PROC MIXED and PROC NLMIXED), SPSS (MIXED), STATA (xtreg and gllamm) and SPLUS (NLME) (see **Software for Statistical Analyses**). At their core, each program does the same job: it fits the growth model to data and provides parameter estimates, measures of precision, diagnostics, and so on. There is also some evidence that all the different packages produce the same, or similar, answers to a given problem [5]. So, in one sense, it does not matter which program you choose. But the packages do differ in many important ways including the ‘look and feel’ of their interfaces, their ways of entering and preprocessing data, their model specification process (the level-1/level-2 specification or the composite specification), their estimation methods (e.g., full **maximum likelihood** vs restricted maximum likelihood – see **Direct Maximum Likelihood Estimation**), their strategies for hypothesis testing, and provision of diagnostics. It is beyond the scope of this entry to discuss these details. Instead, we turn to the results of fitting the growth curve model to data using one statistical program, HLM. Results are presented in Table 1.

Interpreting the Results of Fitting the Growth Curve Model to Data

The fixed effects parameters – the γ ’s of (2) and (5) – quantify the effects of predictors on the individual change trajectories. In our example, they quantify the relationship between the individual growth

parameters and program participation. We interpret these estimates much as we do any regression coefficient, with one key difference: the level-2 ‘outcomes’ that these fixed effects describe are level-1 individual growth parameters. In addition, you can conduct a hypothesis test for each fixed effect using a single parameter test (most commonly, examining the null hypothesis $H_0: \gamma = 0$). As shown in Table 1, we reject all four null hypotheses, suggesting that each parameter plays a role in the story of the program’s effect on children’s cognitive development.

Substituting the $\hat{\gamma}$ ’s in Table 1 into the level-2 model in (2), we have:

$$\begin{aligned}\hat{\pi}_{0i} &= 107.84 + 6.85 \text{ PROGRAM}_i \\ \hat{\pi}_{1i} &= -21.13 + 5.27 \text{ PROGRAM}_i\end{aligned}\quad (6)$$

The first part of the fitted model describes the effects of PROGRAM on initial status; the second part describes its effects on the annual rates of change.

Begin with the first part of the fitted model, for initial status. In the population from which this sample was drawn, we estimate the true initial status (COG at age 1) for the average nonparticipant to be 107.84; for the average participant, we estimate that it is 6.85 points higher (114.69). In rejecting (at the .001 level) the null hypotheses for the two level-2 intercepts, we conclude that the average nonparticipant had a nonzero cognitive score at age 1 (hardly surprising!) but experienced a statistically significant decline over time. Given that this was a randomized trial, you may be surprised to find that the initial status of program participants is 6.85 points

higher than that of nonparticipants. Before concluding that this differential in initial status casts doubt on the randomization mechanism, remember that the intervention started *before* the first wave of data collection, when the children were already 6 months old. This modest 7-point elevation in initial status may reflect early treatment gains attained between ages 6 months and 1 year.

Next examine the second part of the fitted model, for the annual rate of change. In the population from which this sample was drawn, we estimate the true annual rate of change for the average nonparticipant to be -21.13 ; for the average participant, we estimate it to be 5.27 points higher (-15.86). In rejecting (at the .05 level) the null hypotheses for the two level-2 slopes, we conclude that the differences between program participants and nonparticipants in their mean annual rates of change is statistically significant. The average nonparticipant dropped over 20 points during the second year of life; the average participant dropped just over 15. The cognitive performance of both groups of children declines over time, but program participation slows the rate of decline.

Another way of interpreting fixed effects is to plot fitted trajectories for prototypical individuals. Even in a simple analysis like this, which involves just one dichotomous predictor, we find it invaluable to inspect prototypical trajectories visually. For this particular model, only two prototypes are possible: a program participant ($PROGRAM = 1$) and a nonparticipant ($PROGRAM = 0$). Substituting these values into equation (5) yields the predicted initial status and annual growth rates for each:

When $PROGRAM = 0$:

$$\hat{\pi}_{0i} = 107.84 + 6.85(0) = 107.84$$

$$\hat{\pi}_{1i} = -21.13 + 5.27(0) = -21.13$$

When $PROGRAM = 1$:

$$\hat{\pi}_{0i} = 107.84 + 6.85(1) = 114.69$$

$$\hat{\pi}_{1i} = -21.13 + 5.27(1) = -15.86 \quad (7)$$

We use these estimates to plot the fitted change trajectories in the right-hand panel of Figure 1. These plots reinforce the numeric conclusions just articulated. In comparison to nonparticipants, the average participant has a higher score at age 1 and a slower annual rate of decline.

Estimated variance components assess the amount of outcome variability left – at either level-1 or level-2 – after fitting the multilevel model. Because they are harder to interpret in absolute terms, many researchers use null-hypothesis tests, for at least they provide some benchmark for comparison. Some caution is necessary, however, because the null hypothesis is on the border of the parameter space (by definition, these components cannot be negative) and as a result, the asymptotic distributional properties that hold in simpler settings may not apply [9].

The level-1 residual variance, σ_e^2 , summarizes the population variability in an average person's outcome values around his or her own true change trajectory. Its estimate for these data is 74.24, a number that is difficult to evaluate on its own. Rejection of the associated null-hypothesis test (at the .001 level) suggests the existence of additional outcome variation at level-1 (within-person) that may be predictable. This suggests it might be profitable to add time-varying predictors to the level-1 model (such as the number of books in the home or the amount of parent-child interaction).

The level-2 variance components summarize the variability in change trajectories that remains after controlling for predictors (here, $PROGRAM$). Associated tests for these variance components evaluate whether there is any remaining *residual* outcome variation that could potentially be explained by other predictors. For these data, we reject only one of these null hypotheses (at the 0.001 level), for initial status, σ_0^2 . This again suggests the need for additional predictors, but because this is a level-2 variance component (describing residual variation in true initial status), we would consider both time-varying *and* time-invariant predictors to the model. Failure to reject the null hypothesis for σ_1^2 indicates that $PROGRAM$ explains all the potentially predictable variation between children in their true annual rates of change.

Finally turn to the level-2 covariance component, σ_{01} . Failure to reject this null hypothesis indicates that the intercepts and slopes of the individual true change trajectories are uncorrelated – that there is no association between true initial status and true annual rates of change (once the effects of $PROGRAM$ are removed). Were we to continue with model building, this result might lead us to drop the second level-2 residual, ζ_{1i} , from our model, for neither its variance nor covariance with ζ_{0i} , is significantly different from 0.

Postscript

Growth curve modeling offers empirical researchers a wealth of analytic opportunities. The method can accommodate any number of waves of data, the occasions of measurement need not be equally spaced, and different participants can have different data collection schedules. Individual change can be represented by a variety of substantively interesting trajectories, not only the linear functions presented here but also curvilinear and discontinuous functions. Not only can multiple predictors of change be included in a single analysis, simultaneous change across multiple domains (e.g., change in cognitive function and motor function) can be investigated simultaneously. Readers wishing to learn more about growth curve modeling should consult one of the recent books devoted to the topic [3, 4, 6, 8–10].

References

- [1] Burchinal, M.R., Campbell, F.A., Bryant, D.M., Wasik, B.H. & Ramey, C.T. (1997). Early intervention and mediating processes in cognitive performance of low income African American families, *Child Development* **68**, 935–954.
- [2] Cronbach, L.J. & Furby, L. (1970). How should we measure “change” – or should we? *Psychological Bulletin* **74**, 68–80.
- [3] Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd Edition, Oxford University Press, New York.
- [4] Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley, New York.
- [5] Kreft, I.G.G. & de Leeuw, J. (1998). *Introducing Multilevel Modeling*, Sage Publications, Thousand Oaks.
- [6] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edition, Sage Publications, Thousand Oaks.
- [7] Rogosa, D.R., Brandt, D. & Zimowski, M. (1982). A growth curve approach to the measurement of change, *Psychological Bulletin* **90**, 726–748.
- [8] Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, New York.
- [9] Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, London.
- [10] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer, New York.
- [11] Willett, J.B. (1988). Questions and answers in the measurement of change, in *Review of Research in Education (1988–1989)*, E. Rothkopf, ed., American Education Research Association, Washington, pp. 345–422.

(See also **Heteroscedasticity and Complex Variation; Multilevel and SEM Approaches to Growth Curve Modeling; Structural Equation Modeling; Latent Growth Curve Analysis**)

JUDITH D. SINGER AND JOHN B. WILLETT

Guttman, Louise (Eliyahu)

DAVID CANTER

Volume 2, pp. 780–781

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Guttman, Louise (Eliyahu)

Born: February 10, 1916, in Brooklyn, USA.

Died: October 25, 1987, in Minneapolis, USA.

‘Mathematics is contentless, and hence – by itself – not empirical science’: this encapsulates, in his own words, Guttman’s creed. In one of the paradoxes so prevalent in the social sciences, his striving to reveal the fundamental structures of social and psychological phenomena, their ‘content’, has tended to be eclipsed by his many contributions to statistical methodology, the ‘mathematics’ that Guttman always saw as a servant to the discovery of general laws, never an end in its own right. The approach to research that he called *Facet Theory*, which he spent his life developing, is a set of fundamental postulates that describes the interplay between the substantive ways of describing phenomena and the empirical observations of properties of their structures. He showed how this approach generates universally sound predictions, which he called ‘laws’. These are characterized by his ‘First Law of Attitude’:

If any two items are selected from the universe of attitude items towards a given object, and if the population observed is not selected artificially, then the population regressions between these two items will be monotone and with positive or zero sign.

This law summarizes a vast swathe of social science and makes redundant thousands of studies that have poorly defined items or confused methodologies. A similar ‘First Law of Intelligence’ has been hailed as one of the major contributions not only to our understanding of intelligence and how it varies between people but also how it is most appropriately defined.

Guttman completed his BA in 1936 and his MA in 1939 at the University of Minnesota, where his doctorate in social and psychological measurement was awarded in 1942. He had already published, at the age of 24, ‘A Theory and Method of Scale Construction’, describing an innovative approach to measuring attributes that came to bear his name [1]. At the time, it was not appreciated that this ‘Guttman Scale’ enshrined a radically new approach that bridged

the divide between qualitative and quantitative data, demonstrating how, as he later put it, ‘The form of data analysis is part of the hypothesis’. He spent the next half-century developing the implications of his precocious invention into a full-fledged framework for discovering the multidimensional structures of human phenomena [3].

His postgraduate research at The University of Chicago and later at Cornell University as World War II was breaking out gave him a role in a Research Branch of the US War Department, providing him with the starkest awareness of the practical potential of the social sciences. He took this commitment to make psychology and sociology of real significance when he moved to Jerusalem in 1947, setting up a Research Unit within the then Hagana, the Zionist underground army, making this surely the first illegal military group to have an active social research section.

With the establishment of the State of Israel, he converted his military unit into the highly respected Israel Institute of Applied Social Research, which he directed until his death. While being for most of his professional life a Professor at The Hebrew University of Jerusalem, he also always somehow managed to remain active in the United States, usually through visiting professorships, notably at Minnesota, Cornell, Harvard, and MIT. So that even though his Institute provided a crucial service to Israel (as it established itself through many wars with its neighbors) ranging from rapid opinion surveys on topics in the news to the basis for reforms in the civil service grading system, Guttman still contributed to a remarkably wide range of methodological innovations that enabled statistical procedures to be of a more scientific validity. The most notable of these was his discovery of the ‘radex’ as a generalization of factor analysis. This is a structure that combines both qualitative and quantitative facets revealed in data sets drawn from areas as diverse as personal values, intelligence, interpersonal relationships, and even the actions of serial killers [2].

His harnessing of mathematics, in particular, linear algebra, to many problems in multivariate statistics (see **Multivariate Analysis: Overview**) proved of particular value in the development of the computer programs that evolved along with his career. But it is fair to emphasize that his ideas were always pushing the limits of computing capability and it is only now that widely available systems have the

power to achieve what he was aiming for, a truly integrated relationship between statistical analysis and theory development.

He was awarded many honors, including The Rothschild Prize in the Social Sciences (1963), election as President of the Psychometric Society (1970), an Outstanding Achievement Award from the University of Minnesota (1974), The Israel Prize in the Social Sciences (1978), and The Educational Testing Service Measurement Award from Princeton (1984).

Those who knew him well remember him as a mild-mannered person who would not tolerate fools, of a scientific and mathematical integrity that was so impeccable it was easily interpreted as arrogance. An inspiration to his students and close colleagues, whose – always well-intentioned – incisive criticism left those who had contact with him changed for life (see [4] for more on his work).

References

- [1] Guttman, L. (1941). The quantification of a class of attributes – a theory and method of scale construction, in *The Prediction of Personal Adjustment*, P. Horst, ed., Social Science Research Council, New York.
- [2] Guttman, L. (1954). A new approach to factor analysis: the radex, in *Mathematical Thinking in the Social Sciences*, P.F. Lazarsfeld, ed., The Free Press, Glencoe.
- [3] Levy, S., ed. (1994). *Louis Guttman on Theory and Methodology: Selected Writings*, Aldershot, Dartmouth.
- [4] Shye, S. (1997). Louis Guttman, in *Leading Personalities in Statistical Sciences*, N.L. Johnson & S. Kotz, eds, Wiley, New York.

DAVID CANTER

Harmonic Mean

DAVID CLARK-CARTER

Volume 2, pp. 783–784

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Harmonic Mean

The harmonic mean $\bar{X}_h$ of a set of n numbers $X_1, X_2, \dots, X_n (i = 1, 2, \dots, n)$ is defined as

$$\bar{X}_h = \frac{n}{\sum_i \frac{1}{X_i}}. \quad (1)$$

In other words, the harmonic mean is the reciprocal of the mean of the reciprocals of the numbers. Note that the harmonic mean is only defined for sets of positive numbers.

As a simple illustration, we see that the harmonic mean of 20 and 25 is

$$\begin{aligned} \bar{X}_h &= \frac{2}{1/20 + 1/25} = \frac{2}{0.05 + 0.04} = \frac{2}{0.09} \\ &= 22.22 \text{ (to two decimal places)}. \end{aligned} \quad (2)$$

The harmonic mean is applicable in studies involving a between-subjects design but where there are unequal sample sizes in the different groups (unbalanced designs). One application is working out the statistical **power** of a test for a between-subjects t Test when the samples in the two groups are unequal [2]. The recommended sample size necessary to achieve power of 0.8 with an effect size of $d = 0.5$ (i.e., that the mean of the two groups differs by half a standard

deviation), using a two-tailed probability and an alpha level of 0.05 is 64 in each group, which would produce a total sample size of 128. When the sample sizes are unequal, then the equivalent of the sample size for each group required to attain the same level of statistical power is the harmonic mean. Suppose that one group had double the sample size of the other group, then, to achieve the same level of statistical power, the harmonic mean of the sample sizes would have to be 64. This means that the smaller group would require 48 members, while the larger group would need 96. Thus, the total sample size would be 144, which is an increase of 16 people over the design with equal numbers in each group (a balanced design).

A second example of the use of the harmonic mean is in **analysis of variance (ANOVA)** for unbalanced designs [1]. One way to deal with the problem is to treat all the groups as though they have the same sample size: the harmonic mean of the individual sample sizes.

References

- [1] Clark-Carter, D. (2004). *Quantitative Psychological Research: A Student's Handbook*, Psychology Press, Hove.
- [2] Cohen, J. (1988). *Statistical Power for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.

DAVID CLARK-CARTER

Hawthorne Effect

GARY D. GOTTFREDSON

Volume 2, pp. 784–785

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hawthorne Effect

The term *Hawthorne effect* has come to refer mistakenly to an increase in worker performance resulting merely from research or managerial attention to a work group. References to a Hawthorne effect in describing research have also come to connote flawed research. For example, when researchers obtrusively measure progress in implementing a health, educational, or social program, critics may raise questions about whether a Hawthorne effect may account for outcomes observed. Accordingly, understanding not only how the performance improvements observed in the Hawthorne study were produced but also of the design flaws in the original research are helpful for stimulating better worker performance and for conducting research that allows sound causal inferences.

Preliminary studies conducted at the Western Electric plant in the late 1920s briefly described by Rothlisberger and Dickson ([8, pp. 14–18], citing a brief nontechnical account, [9]) implied that the changes researchers made in illumination levels produced improvements in worker performance whether lighting was increased or decreased. A detailed account of these preliminary studies is not available. The preliminary studies led to a subsequent five-year program of research on industrial fatigue that is well reported. These studies were part of a larger inquiry about social relations in the workplace that broadly influenced the course of subsequent research and practice in industrial relations and industrial organizational psychology. Among the findings of the Hawthorne researchers about workplace social relations was the observation that social influences discouraging excess productivity can limit worker output.

In the well-documented relay-assembly study that was part of the Hawthorne research, the research participants were a group of five women who assembled electrical components (relays). Relay assembly was a repetitive task in which a worker selected parts, discarded flawed parts, and held pieces together while inserting screws. The five women in the study were located in a special room that was partitioned from an area in which a much larger group of about 100 assemblers worked.

The practice in the general relay-assembly room was to guarantee workers an hourly wage, but if the group produced a number of units with a piecework

value greater than the workers' aggregate guaranteed wages, the pay would be increased. For example, if the piecework value of relays assembled was 3% greater than the aggregate guaranteed wages of the 100 workers, then each worker would receive 103% of her guaranteed pay. Thus, the pay received in the general relay room was based on a large-group-contingency plan not expected to have much influence on productivity because reward is only tenuously contingent on individual performance [4]. Three key changes occurred in the experimental relay-assembly experiment. First, the five experimental women were moved to a separate room. Second, a running tally of each woman's production was made by means of an electronic counter, and the tally was available to the workers as they worked. Third, pay became contingent on the productivity of a small group of five people (rather than a large group of 100).

The researchers introduced additional changes over the course of the experiment. They increased scheduled rest periods. They also varied the food or beverages provided to the workers. And they experimented with shortening the workday and length of the workweek. The hourly production of the relay assemblers began to rise when the small group contingency pay plan was put in place, and with only occasional temporary downturns, production continued to rise over the course of the study [2, 3, 8].

Although a number of explanations of the increased productivity have been offered [1, 2, 10, 11], the most persuasive are those suggested by Parsons [7] and Gottfredson [3]. Rothlisberger and Dickson [8] had noted that the introduction near the beginning of the study of the small group contingency for pay seemed to lead to an increase in productivity, but they argued that this could not be responsible for the continuing rise in productivity over subsequent weeks. The researchers had selected experienced assemblers to participate in the experiment to rule out learning effects in what they viewed as a study in industrial fatigue, but they apparently overlooked the possibility that even experienced assemblers could learn to produce more relays per hour.

Parsons [7] persuasively argued for a learning explanation of the Hawthorne effect (*see Carry-over and Sequence Effects*). Basically, the provision of feedback and rewards contingent on performance led to learning and increased speed and accuracy in assembly. In addition, the separation of the workers

from the main group of assemblers and the use of small group contingencies may have stimulated peer influence to favor rather than limit productivity. If, as seems most likely, the learning interpretation is correct, a serious design flaw in the Hawthorne relay-assembler experiment was the failure to establish a performance baseline before varying rest periods and other working conditions presumed to be related to fatigue. Once reward for performance and feedback were both present, productivity began to increase and generally increased thereafter even during periods when other conditions were constant. For details, see accounts by Gottfredson and Parsons [3, 7]. Furthermore, observation logs imply that the workers set personal goals of improving their performance. For example, one assembler said, “I made 421 yesterday, and I’m going to make better today” [8, p. 74].

In contemporary perspective, the Hawthorne effect is understandable in terms of goal-setting theory [5, 6]. According to goal-setting theory, workers attend to feedback on performance when they adopt personal performance goals. Contingent rewards, goals set by workers, attention to information, and the removal of social obstacles to improved productivity led workers to learn to assemble relays faster – and to display their learning by producing more relays per hour. Gottfredson [3] has provided additional examples in which a similar process – a Hawthorne effect according to this understanding – is produced. Understanding the remarkable improvement in worker performance in the Hawthorne relay-assembly study in this way is important because it suggests how one obtains the ‘Hawthorne effect’ when improvements in worker performance are desired: remove obstacles to improvement, set goals, and provide feedback. Then if learning is possible, performance may improve.

What of the Hawthorne effect as research design flaw? There was a flaw in the relay-assembly study – the failure of the design to rule out learning as a rival

hypothesis to working conditions as an explanation for the changes in productivity observed. Designs that rule out this rival hypothesis, such as the establishment of an adequate baseline or the use of a randomly equivalent control group, are therefore often desirable in research.

References

- [1] Carey, A. (1967). The Hawthorne studies: a radical criticism, *American Sociological Review* **32**, 40–416.
- [2] Franke, R.H. & Kaul, J.D. (1978). The Hawthorne experiments: first statistical interpretation, *American Sociological Review* **43**, 623–643.
- [3] Gottfredson, G.D. (1996). The Hawthorne misunderstanding (and how to get the Hawthorne effect in action research), *Journal of Research in Crime and Delinquency* **33**, 28–48.
- [4] Lawler, E.E. (1971). *Pay and Organizational Effectiveness*, McGraw-Hill, New York.
- [5] Locke, E.A. & Latham, G.P. (1990). *A Theory of Goal Setting and Task Performance*, Prentice Hall, Englewood Cliffs.
- [6] Locke, E.A. & Latham, G.P. (2002). Building a practically useful theory of goal setting and task motivation: a 35 year Odyssey, *American Psychologist* **57**, 705–717.
- [7] Parsons, H.M. (1974). What happened at Hawthorne? *Science* **183**, 922–932.
- [8] Rothlisberger, F.J. & Dickson, W.J. (1930). *Management and the Worker*, Harvard University Press, Cambridge.
- [9] Snow, C.E. (1927). A discussion of the relation of illumination intensity to productive efficiency, *Technical Engineering News* 256.
- [10] Stagner, R. (1956). *Psychology of Industrial Conflict*, Wiley, New York.
- [11] Viteles, M.S. (1953). *Motivation and Morale in Industry*, Norton, New York.

GARY D. GOTTFREDSON

Heritability

JOHN L. HOPPER

Volume 2, pp. 786–787

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Heritability

Before discussing what genetic heritability is, it is important to be clear about what it is not. For a binary trait, such as whether an individual has a disease, heritability is not the proportion of disease in the population attributable to, or caused by, genetic factors. For a continuous trait, genetic heritability is not a measure of the proportion of an individual's score attributable to genetic factors. Heritability is not about cause *per se*, but about the causes of variation in a trait across a particular population.

Definitions

Genetic heritability is defined for a quantitative trait. In general terms, it is the proportion of variation attributable to genetic factors. Following a genetic and environmental variance components approach, let Y have a mean μ and variance σ^2 , which can be partitioned into genetic and environmental components of variance, such as additive genetic variance σ_a^2 , dominance genetic variance σ_d^2 , common environmental variance σ_c^2 , individual specific environmental variance σ_e^2 , and so on (*see ACE Model*).

Genetic heritability in the narrow sense is defined as

$$\frac{\sigma_a^2}{\sigma^2}, \quad (1)$$

while genetic heritability in the broad sense is defined as

$$\frac{\sigma_g^2}{\sigma^2}, \quad (2)$$

where σ_g^2 includes all genetic components of variance, including perhaps components due to **epistasis** (gene–gene interactions; *see Genotype*) [3]. In addition to these random genetic effects, the total genetic variation could also include that variation explained when the effects of measured genetic markers are modeled as a fixed effect on the trait mean.

The concept of genetic heritability, which is really only defined in terms of variation in a quantitative trait, has been extended to cover categorical traits by reference to a genetic liability model (*see Liability Threshold Models*). It is assumed that there is an underlying, unmeasured continuous ‘liability’ scale divided into categories by ‘thresholds’. Under

the additional assumption that the liability follows a normal distribution, genetic and environmental components of variance are estimated from the pattern of associations in categorical traits measured in relatives. The genetic heritability of the categorical trait is then often defined as the genetic heritability of the presumed liability (latent variable), according to (1) and (2).

Comments

There is no unique value of the genetic heritability of a characteristic. Heritability varies according to which factors are taken into account in specifying both the mean and the total variance of the population under consideration. That is to say, it is dependent upon modeling of the mean, and of the genetic and environmental variances and covariances. Moreover, the total variance and the variance components themselves may not be constants, even in a given population. For example, even if the genetic variance actually increased with age, the genetic heritability would decrease with age if the variation in nongenetic factors increased with age more rapidly. That is to say, genetic heritability and genetic variance can give conflicting impressions of the ‘strength of genetic factors’.

Genetic heritability will also vary from population to population. For example, even if the heritability of a characteristic in one population is high, it may be quite different in another population in which there is a different distribution of environmental influences.

Measurement error in a trait poses an upper limit on its genetic heritability. Therefore, traits measured with large measurement error cannot have substantial genetic heritabilities, even if variation about the mean is completely independent of environmental factors. By the definitions above, one can increase the genetic heritability of a trait by measuring it more precisely, for example, by taking repeat measurements and averaging, although strictly speaking the definition of the trait has been changed also. A trait that is measured poorly (in the sense of having low reliability) will inevitably have a low heritability because much of the total variance will be due to measurement error (σ_e^2). However, a trait with relatively little measurement error will have a high heritability if all the nongenetic factors are known and taken into account in the modeling of the mean.

2 Heritability

Fisher [1] recognized these problems and noted that

whereas ... the numerator has a simple genetic meaning, the denominator is the total variance due to errors of measurement [including] those due to uncontrolled, but potentially controllable environmental variation. It also, of course contains the genetic variance ... Obviously, the information contained in [the genetic variance] is largely jettisoned when its actual value is forgotten, and it is only reported as a ratio to this hotch-potch of a denominator.

Historically, other quantities have also been termed *heritabilities*, but it is not clear what parameter is being estimated, for example, Holzinger's $H = (r_{MZ} - r_{DZ})$ (the correlation between monozygotic twins minus the correlation between dizygotic twins) [2], Nichol's $HR = 2(r_{MZ} - r_{DZ})/r_{MZ}$ [5], the E of Neel & Schull [4] based on twin data alone, and Vandenberg's $F = 1/[1 - (\sigma_a^2/\sigma^2)]$ [6]. Furthermore, the statistical properties of these estimators do not appear to have been studied.

References

- [1] Fisher, R.A. (1951). Limits to intensive production in animals, *British Agricultural Bulletin* **4**, 217–218.
- [2] Holzinger, K.J. (1929). The relative effect of nature and nurture influences on twin differences, *Journal of Educational Psychology* **20**, 245–248.
- [3] Lush, J.L. (1948). Heritability of quantitative characters in farm animals, *Supplement of Hereditas* **1948**, 256–375.
- [4] Neel, J.V. & Schull, W.J. (1954). *Human Heredity*, University of Chicago Press, Chicago.
- [5] Nichols, R.C. (1965). The national merit twin study, in *Methods and Goals in Human Behaviour Genetics*, S.G. Vandenberg, ed., Academic Press, New York.
- [6] Vandenberg, S.G. (1966). Contributions of twin research to psychology, *Psychological Bulletin* **66**, 327–352.

JOHN L. HOPPER

Heritability: Overview

RICHARD J. ROSE

Volume 2, pp. 787–790

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Heritability: Overview

The contributor to this essay consulted a resource given to him when he entered graduate school for a definition of the term heritability. The resource aspired to include ‘all terms frequently used by psychologists’, and it purported to be a ‘comprehensive dictionary’ of psychological terms [2]. Heritability does not appear in the dictionary. But, under the entry ‘hereditarianism’, that 1958 dictionary asked the question: ‘To what extent do genetic factors influence behavior?’ And, that is the very question addressed by the statistical estimate we now call heritability. Once omitted from a comprehensive dictionary of psychological terms, heritability now occupies a full chapter in most introductory psychology textbooks, appears frequently in news releases on research in behavior and medicine, and yields dozens, perhaps hundreds, of links in an Internet search: heritability has become a central concept in behavioral science.

A statistical measure originating in quantitative genetics, heritability is an estimate of the contribution of genetic differences to the differences observed in a measured variable (e.g., some dimension of behavior) in a given population at a given time. Estimates of heritability have been obtained from many species, for many diverse behaviors, in samples of relatives from many human cultures, and across the human lifespan from infancy to senescence. Together, the accumulated heritability estimates offer compelling evidence of the importance of genetic influences on behavior, and, as a result, the concept of heritability has assumed critical importance in understanding the meaning and development of individual differences in behavioral development. Efforts to obtain heritability estimates have generated volumes of research over the past four decades, and, initially, these estimates aroused controversy and misunderstanding as heritability research within quantitative genetic studies of plants and animals was broadened to widespread application in human behavior genetics. Within the fields of behavioral and psychiatric genetics, the fields whose practitioners have developed the analytic tools to yield estimates of heritability for complex human behaviors, the term assumed new meaning: throughout the 1960s and 1970s, the mere demonstration of nonzero heritability for diverse behaviors was a goal, often *the* research goal; behavioral geneticists sought to show that the name of their discipline was not an

oxymoron. But by the 1980s and 1990s, the cumulative results of heritability research convinced most behavioral scientists that heritability estimates for nearly all behaviors are nonzero, and current research focuses much less on demonstrating heritability, and much more on how it is modulated by changing environments, or gene-environment interaction.

Heritability has two definitions. As a statistical estimate, the term is defined as the proportion of observable or phenotypic variance attributable to underlying genetic variation (*see* **Genotype**). And within that definition, narrow heritability considers **additive genetic variance** only, so the term is defined as the ratio of variance due to additive genes to the total variance observed in the behavior under study. That definition originated in selective breeding studies of animals, and it remains important in those applications, where the question addressed is the extent to which offspring will ‘breed true’. In contrast, a broad-sense statistical definition of heritability considers it to be the ratio of observed differences to *all* sources of genetic variation, additive or nonadditive (*see* **ACE Model**). Behavioral scientists are less interested in breeding coefficients than in the extent to which individual differences in behavior are due to genetic differences, whatever their source. So it is this second definition that captures the usual meaning of the concept for behavioral science: heritability defines the extent to which individual differences in genes contribute to individual differences in observable behavior. Some important caveats: Heritability is an abstract concept. More importantly, it is a population concept. It does not describe individuals, but rather the underlying differences between people. It is an estimate, typically made from the resemblance observed among relatives, and, as is true of any statistical estimate, heritability estimates include errors of estimation as a function of (genetic) effect size, the size and representativeness of the studied sample of relatives, and the precision with which the studied outcome can be measured. For some outcomes, such as adult height, the genetic effect is very large, and the outcome can be measured with great precision. For others, such as prosocial behavior in childhood, the genetic effect may be more modest, and the measurement much more uncertain. Heritability estimates vary, also, with age and circumstance, because the magnitude of genetic variance may dramatically change during development and across environments.

2 Heritability: Overview

Heritability is a relative concept in another sense: it is derived from the comparative similarity of relatives who differ in their shared genes. The most common approach is to compare samples of the two kinds of twins. Monozygotic (MZ) twins derive from a single zygote, and, barring rare events, they share all their genes identical-by-descent. Dizygotic (DZ) twins, like ordinary siblings, arise from two zygotes created by the same parents, and share, on average, one-half of their segregating genes. If, in a large and representative sample of twins, behavioral similarity of DZ twins approaches that found for MZ twins, genetic factors play little role in creating individual differences in that behavior; heritability is negligible, and the observed behavioral differences must be due to differences in environments shared by both kinds of cotwins in their homes, schools, and neighborhoods. Conversely, if the observed correlation of MZ cotwins doubles that found for DZ twins – a difference in resemblance that parallels their differences in genetic similarity – heritability must be nonzero (*see Twin Designs*). We can extend the informational yield found in contrasts of the two kinds of twins by adding additional members of the families of the twins. Consider, for example, children in families of monozygotic twin parents. Children in each of the two nuclear families derive half their genes from a twin parent, and those genes are identical with the genes of the parent's twin sister or brother, the children's 'twin aunt' or 'twin uncle'. Because the children and the twin aunt or uncle do not live in the same household, their resemblance cannot be due to household environment. And because the MZ twin parents have identical sets of nuclear genes, their children are genetically related to one another as half-siblings; socially, they are reared as cousins in separate homes. Thus, MZ twin families yield informative relationships ranging from those who share all their genes (MZ parents) to those sharing one-half (siblings in each nuclear family; parents and their children; children and their twin aunt/uncle), one-quarter (the cousins who are half-siblings), or zero (children and their spousal aunt or uncle).

We studied two measures in families of MZ twin parents: one, a behavioral measure of nonverbal intelligence, the other, the sum of fingerprint ridge counts, a morphological measure known to be highly heritable. For both measures, familial resemblance appeared to be a direct function of shared genes [8]. But, there was a substantial difference in

the magnitude of heritability estimates found for the two measures, ranging from 0.68 to 0.92 for total ridge count, but much less, 0.40 to 0.54 for the measure of nonverbal intelligence. That finding is consistent with research on many species [5]: behavioral traits exhibit moderate levels of heritability, much less than what is found for morphological and physiological traits, but greater than is found for life-history characteristics. The heritability estimates illustrated from families of MZ twins date from a 1979 study. They were derived from coefficients of correlation and regression among different groups of relatives in these families; interpretation of those estimates was confounded by the imprecision of the coefficients on which they were based, and the fact that effects of common environment were ignored. Now, 25 years later, estimates of heritability typically include 95% confidence intervals, and they are derived from robust analytic models. The estimates are derived from models fit to data from sets of relatives, and heritability is documented by showing that models that set it to zero result in a significantly poorer fit of the model to the observed data. Effects of common environments are routinely tested in an analogous manner. Analytic techniques for estimating heritability are now much more rigorous, and allow for tests of differential heritability in males and females. But they remain estimates derived from the relative resemblance of relatives who differ in the proportion of their shared genes.

Why do people differ? Why do brothers and sisters, growing up together, sharing half their genes and many of their formative experiences, turn out differently – in their interests, aptitudes, lifestyles? The classic debate was framed as nature *versus* nurture, as though genetic dispositions and experiential histories were somehow oppositional, and as though a static decomposition of genetic and environmental factors could adequately capture a child's developmental trajectory. But, clearly, this is simplification. If all environmental differences were removed in a population, such that all environments offered the same opportunities and incentives for acquisition of cognitive skills (and if all tests were perfectly reliable), people with the same genes would obtain the same aptitude test scores. If, conversely, the environments people experienced were very different for reasons independent of their genetic differences, heritability would be negligible. High heritability estimates do not elucidate *how* genetic differences effect differences in

behavioral outcomes, and it seems likely that many, perhaps most, gene effects on behavior are largely indirect, influencing the trait-relevant environment to which people are exposed.

Much recent research in the fields of behavioral and psychiatric genetics demonstrate substantial gene by environment interaction. Such research makes it increasingly apparent that the meaning of heritability depends on the circumstances in which it is assessed. Recent data suggest that it is nonsensical to conceptualize 'the heritability' of a complex behavioral trait, as if it were fixed and stable over time and environments. Across different environments, the modulation of genetic effects on adolescent substance use ranges as much as five or six fold, even when those environments are crudely differentiated as rural versus urban residential communities [1, 7], or religious versus secular households [4]. Similarly, heritability estimates for tobacco consumption vary dramatically for younger and older cohorts of twin sisters [3]. Such demonstrations suggest that genetic factors play much more of a role in adolescent alcohol use in environments where alcohol is easily accessed and community surveillance is reduced. And, similarly, as social restrictions on smoking have relaxed across generations, heritable influences have increased. Equally dramatic modulation of genetic effects by environmental variation is evident in effects of differences in socioeconomic status on the heritability of children's IQ [10].

In recent years, **twin studies** have almost monotonously demonstrated that estimates of heritable variance are nonzero across all domains of individual behavioral variation that can be reliably assessed. These estimates are often modest in magnitude, and, perhaps more surprisingly, quite uniform across different behavioral traits. But if heritability is so ubiquitous, what consequences does it have for scientific understanding of behavioral development?

All human behaviors are, to some degree, heritable, but that cannot be taken as evidence that the complexity of human behavior can be reduced to relatively simple genetic mechanisms [9]. Confounding heritability with strong biological determinism is an error. An example is criminality; like nearly all behaviors, criminality is to some degree heritable. There is nothing surprising, and nothing morally repugnant in the notion that not all children have

equal likelihood of becoming a criminal. To suggest that, however, is neither to suggest that specific biological mechanisms for criminality are known, nor even that they exist. All behavior is biological and genetic, but some behaviors are biological in a stronger sense than others, and some behaviors are genetic in a stronger sense than others [9]. Criminality is, in part, heritable, but unnecessary mischief is caused by reference to genes 'for' criminality – or any similar behavioral outcome. We inherit dispositions, not destinies [6]. Heritability is an important concept, but it is important to understand what it is. And what it is not.

References

- [1] Dick, D.M., Rose, R.J., Viken, R.J., Kaprio, J. & Koskenvuo, M. (2001). Exploring gene-environment interactions: socioregional moderation of alcohol use, *Journal of Abnormal Psychology* **110**, 625–632.
- [2] English, H.B. & English, A.C. (1958). *A Comprehensive Dictionary of Psychological and Psychoanalytical Terms: a Guide to Usage*, Longmans, Green & Co..
- [3] Kendler, K.S., Thornton, L.M. & Pedersen, N.L. (2000). Tobacco consumption in Swedish twins reared apart and reared together, *Archives of General Psychiatry* **57**, 886–892.
- [4] Koopmans, J.R., Slutske, W.S., van Baal, C.M. & Boomsma, D.I. (1999). The influence of religion on alcohol use initiation: evidence for a Genotype X environment interaction, *Behavior Genetics* **29**, 445–453.
- [5] Mosseau T.A. & Roff D.A. (1987). Natural selection and the heritability of fitness components, *Heredity* **59**, 181–197.
- [6] Rose, R.J. (1995). Genes and human behavior, *Annual Reviews of Psychology* **46**, 625–654.
- [7] Rose, R.J., Dick, D.M., Viken, R.J. & Kaprio, J. (2001). Gene-environment interaction in patterns of adolescent drinking: regional residency moderates longitudinal influences on alcohol use, *Alcoholism: Clinical & Experimental Research* **25**, 637–643.
- [8] Rose, R.J., Harris, E.L., Christian, J.C. & Nance, W.E. (1979). Genetic variance in nonverbal intelligence: data from the kinships of identical twins, *Science* **205**, 1153–1155.
- [9] Turkheimer, E. (1998). Heritability and biological explanation, *Psychological Review* **105**, 782–791.
- [10] Turkheimer, E.N., Haley, A., Waldron, M., D'Onofrio, B.M. & Gottesman, I.I. (2003). Socioeconomic status modifies heritability of IQ in young children, *Psychological Science* **14**, 623–628.

RICHARD J. ROSE

Heteroscedasticity and Complex Variation

HARVEY GOLDSTEIN

Volume 2, pp. 790–795

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Heteroscedasticity and Complex Variation

Introduction

Consider the simple linear regression model with normally distributed residuals

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad e_i \sim N(0, \sigma_e^2) \quad (1)$$

where β_0 , β_1 are the intercept and slope parameters respectively, i indexes the observation, and e_i is an error term (*see Multiple Linear Regression*). In standard applications, such a model for a data set typically would be elaborated by adding further continuous or categorical explanatory variables and interactions until a suitable model describing the observed data is found (*see Model Selection*). A common diagnostic procedure is to study whether the constant residual variance (homoscedasticity) assumption in (1) is satisfied. If not, a variety of actions have been suggested in the literature, most of them concerned with finding a suitable nonlinear transformation of the response variable so that the homoscedasticity assumption is more closely approximated (*see Transformation*). In some cases, however, this may not be possible, and it will also in general change the nature of any regression relationship. An alternative is to attempt to model the heteroscedasticity explicitly, as a function of explanatory variables. For example, for many kinds of behavioral and social variables males have a larger variance than females, and rather than attempting to find a transformation to equalize these variances, which would in this case be rather difficult, we could fit a model that had separate variance parameters for each gender. This would have the advantage not only of a better fitting model, but also of providing information about variance differences that is potentially of interest in its own right.

This article discusses general procedures for modeling the variance as a function of explanatory variables. It shows how efficient estimates can be obtained and indicates how to extend the case of linear models such as (1) to handle multilevel data (*see Linear Multilevel Models*) [2]. We will first describe, through a data example using a simple linear model, a model fitting separate gender variances and then discuss general procedures.

An Example Data Set of Examination Scores

The data have been selected from a very much larger data set of examination results from six inner London Education Authorities (school boards). A key aim of the original analysis was to establish whether some schools were more ‘effective’ than others in promoting students’ learning and development, taking account of variations in the characteristics of students when they started Secondary school. For a full account of that analysis, see Goldstein et al. [5].

The variables we shall be using are an approximately normally distributed examination score for 16-year-olds as the response variable, with a standardized reading test score for the same students at age 11 and gender as the explanatory variables.

The means and variances for boys and girls are given in Table 1.

We observe, as expected, that the variance for girls is lower than for the boys.

We first fit a simple model which has a separate mean for boys and girls and which we write as

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad e_i \sim N(0, \sigma_e^2)$$

$$x_{1i} = 1 \text{ if a boy, } 0 \text{ if a girl, } x_{2i} = 1 - x_{1i} \quad (2)$$

There is no intercept in this model since we have a dummy variable for both boys and girls. Note that these data in fact have a two-level structure with significant variation between schools. Nevertheless, for illustrative purposes here we ignore that, but see Browne et al. [1] for a full multilevel analysis of this data set.

If we fit this model to the data using ordinary least squares (OLS) regression (*see Least Squares Estimation; Multiple Linear Regression*), we obtain the estimates in Table 2.

Note that the fixed coefficient estimates are the same as the means in Table 1, so that in this simple case the estimates of the means do not depend on the homoscedasticity assumption. We refer to the explanatory variable coefficients as ‘fixed’ since they

Table 1 Exam scores by gender

	Boy	Girl	Total
N	1623	2436	4059
Mean	−0.140	0.093	−0.000114
Variance	1.051	0.940	0.99

2 Heteroscedasticity and Complex Variation

Table 2 OLS estimates from separate gender means model (2)

	Coefficient	Standard error
<i>Fixed</i>		
Boy (β_1)	-0.140	0.024
Girl (β_2)	0.093	0.032
<i>Random</i>		
Residual variance (σ_e^2)	0.99	0.023
-2 log-likelihood	11455.7	

have a fixed underlying population value, and the residual variance is under the heading ‘random’ since it is associated with the random part of the model (residual term).

Modeling Separate Variances

Now let us extend (2) to incorporate separate variances for boys and girls. We write

$$\begin{aligned}
 y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + e_{1i} x_{1i} + e_{2i} x_{2i} \\
 e_{1i} &\sim N(0, \sigma_{e1}^2), \quad e_{2i} \sim N(0, \sigma_{e2}^2) \\
 x_{1i} &= 1 \text{ if a boy, } 0 \text{ if a girl, } x_{2i} = 1 - x_{1i} \quad (3)
 \end{aligned}$$

so that we have separate residuals, with their own variances for boys and girls. Fitting this model, using the software package MLwiN [6], we obtain the results in Table 3.

We obtain, of course, the same values as in Table 1 since this model is just fitting a separate mean and variance for each gender. (Strictly speaking they will not be exactly identical because we have used **maximum likelihood estimation** for our model estimates, whereas Table 1 uses unbiased estimates for the variances; if restricted maximum likelihood (REML) model estimates are used, then they will be

Table 3 Estimates from separate gender means model (3)

	Coefficient	Standard error
<i>Fixed</i>		
Boy (β_1)	-0.140	0.025
Girl (β_2)	0.093	0.020
<i>Random</i>		
Residual variance Boys (σ_{e1}^2)	1.051	0.037
Residual variance Girls (σ_{e2}^2)	0.940	0.027
-2 log-likelihood	11449.5	

identical (*see Maximum Likelihood Estimation*)). Note that the difference in the -2 log-likelihood values is 6.2, which judged against a chi squared distribution on 1 degree of freedom (because we are adding just 1 parameter to the model) is significant at approximately the 1% level.

Now let us rewrite (3) in a form that will allow us to generalize to more complex variance functions.

$$\begin{aligned}
 y_i &= \beta_0 + \beta_1 x_{1i} + e_i \\
 e_i &= e_{0i} + e_{1i} x_{1i} \\
 \text{var}(e_i) &= \sigma_{e0}^2 + 2\sigma_{e01} x_{1i} + \sigma_{e1}^2 x_{1i}^2, \quad \sigma_{e1}^2 \equiv 0 \\
 x_{1i} &= 1 \text{ if a boy, } 0 \text{ if a girl} \quad (4)
 \end{aligned}$$

Model (4) is equivalent to (3) with

$$\begin{aligned}
 \beta_2^* &\equiv \beta_0, \quad \beta_1^* \equiv \beta_0 + \beta_1 \\
 \sigma_{e2}^{2*} &\equiv \sigma_{e0}^2, \quad \sigma_{e1}^{2*} \equiv \sigma_{e0}^2 + 2\sigma_{e01} \quad (5)
 \end{aligned}$$

where the * superscript refers to the parameters in (3).

In (4), for convenience, we have used a standard notation for variances and the term σ_{e01} is written as if it were a covariance term. We have written the residual variance in (4) as $\text{var}(e_i) = \sigma_{e0}^2 + 2\sigma_{e01} x_{1i} + \sigma_{e1}^2 x_{1i}^2$, $\sigma_{e1}^2 \equiv 0$, which implies a covariance matrix with one of the variances equal to zero but a nonzero covariance. Such a formulation is not useful and the variance in (4) should be thought of simply as a reparameterization of the residual variance as a function of gender. The notation in (4) in fact derives from that used in the general multilevel case [2], and in the next section we shall move to a more straightforward notation that avoids any possible confusion with covariance matrices.

Modeling the Variance in General

Suppose now that instead of gender the explanatory variable in (4) is continuous, for example, the reading test score in our data set, which we will now denote by x_{3i} . We can now write a slightly extended form of (4) as

$$\begin{aligned}
 y_i &= \beta_0 + \beta_3 x_{3i} + e_i \\
 e_i &= e_{0i} + e_{3i} x_{3i} \\
 \text{var}(e_i) &= \sigma_{e0}^2 + 2\sigma_{e03} x_{3i} + \sigma_{e3}^2 x_{3i}^2 \quad (6)
 \end{aligned}$$

Table 4 Estimates from fitting reading score as an explanatory variable with a quadratic variance function

	Coefficient	Standard error
<i>Fixed</i>		
Intercept (β_0)	-0.002	
Reading (β_3)	0.596	0.013
<i>Random</i>		
Intercept variance (σ_{e0}^2)	0.638	0.017
Covariance (σ_{e03})	0.002	0.007
Reading variance (σ_{e3}^2)	0.010	0.011
-2 log-likelihood	9759.6	

This time we can allow the variance to be a quadratic function of the reading score; in the case of gender, since there are really only two parameters (variances) one of the parameters in the variance function (σ_{e1}^2) was redundant. If we fit (6), we obtain the results in Table 4.

The deviance (-2 log-likelihood) for a model that assumes a simple residual variance is 9760.5, so that there is no evidence here that complex variation exists in terms of the reading score. This is also indicated by the standard errors for the random parameters, although care should be taken in interpreting these (and more elaborate Wald tests) using Normal theory since the distribution of variance estimates will often be far from Normal.

Model (6) can be extended by introducing several explanatory variables with 'random coefficients' e_{hi} . Thus, we could have a model where the variance is a function of gender (with x_{2i} as the dummy variable for a girl) and reading score, that is,

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

$$\text{var}(e_i) = \sigma_{e_i}^2 = \alpha_0 + \alpha_2 x_{2i} + \alpha_3 x_{3i} \quad (7)$$

We have changed the notation here so that the residual variance is modeled simply as a linear function of explanatory variables (Table 5).

The addition of the gender term in the variance is associated only with a small reduction in deviance (1.6 with 1 degree of freedom), so that including the reading score as an explanatory variable in the model appears to remove the heterogeneous variation associated with gender. Before we come to such a conclusion, however, we look at a more elaborate model where we allow for the variance to depend on the interaction between gender and the reading score,

Table 5 Estimates from fitting reading score and gender (girl = 1) as explanatory variables with linear variance function

	Coefficient	Standard error
<i>Fixed</i>		
Intercept (β_0)	-0.103	
Girl (β_2)	0.170	0.026
Reading (β_3)	0.590	0.013
<i>Random</i>		
Intercept (α_0)	0.665	0.023
Girl (α_2)	-0.038	0.030
Reading (α_3)	0.006	0.014
-2 log-likelihood	9715.3	

that is,

$$y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

$$\text{var}(e_i) = \sigma_{e_i}^2 = \alpha_0 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{2i} x_{3i} \quad (8)$$

Table 6 shows that the fixed effects are effectively unchanged after fitting the interaction term, but that the latter is significant with a reduction in deviance of 6.2 with 1 degree of freedom. The variance function for boys is given by $0.661 - 0.040x_3$ and for girls by $0.627 + 0.032x_3$. In other words, the residual variance decreases with an increasing reading score for boys but increases for girls, and is the same for boys and girls at a reading score of about 0.5 standardized units. Thus, the original finding that boys have more variability than girls needs to be modified: initially low achieving boys (in terms of reading) have higher variance, but the girls have higher variance if they are initially high achievers. It is interesting to note that if we fit an interaction term between reading and gender in the fixed part of the model, we obtain a very small and nonsignificant coefficient whose inclusion does not affect the estimates for the remaining parameters. This term therefore, is omitted from Table 6.

One potential difficulty with linear models for the variance is that they have no constraint that requires them to be positive, and in some data sets the function may become negative within the range of the data or provide negative variance predictions that are unreasonable outside the range. An alternative formulation that avoids this difficulty is to formulate a nonlinear model, for example, for the logarithm of the variance having the general

4 Heteroscedasticity and Complex Variation

Table 6 Estimates from fitting reading score and gender (girl = 1) as explanatory variables with linear variance function including interaction

	Coefficient	Standard error
<i>Fixed</i>		
Intercept (β_0)	-0.103	
Girl (β_2)	0.170	0.026
Reading (β_3)	0.590	0.013
<i>Random</i>		
Intercept (α_0)	0.661	0.023
Girl (α_2)	-0.034	0.030
Reading (α_3)	-0.040	0.022
Interaction (α_4)	0.072	0.028
-2 log-likelihood	9709.1	

form

$$\log[\text{var}(e_i)] = \sum_h \alpha_h x_{hi}, \quad x_{0i} \equiv 1 \quad (9)$$

We shall look at estimation algorithms suitable for either the linear or nonlinear formulations below.

Covariance Modeling and Multilevel Structures

Consider the repeated measures model where the response is, for example, a growth measure at successive occasions on a sample of individuals as a polynomial function of time (t)

$$y_{ij} = \sum_{h=0}^p \beta_h t_{ij}^h + e_{ij}$$

$$\text{cov}(\mathbf{e}_j) = \Omega_e \quad \mathbf{e}_j = \{e_{ij}\} \quad (10)$$

where $\mathbf{e}_j$ is the vector of residuals for the j th individual and i indexes the occasion. The residual covariance matrix between measurements at different occasions (Ω_e) is nondiagonal since the same individuals are measured at each occasion and typically there would be a relatively large between-individual variation. The covariance between the residuals, however, might be expected to vary as a function of their distances apart so that a simple model might be as follows

$$\text{cov}(e_{tj}, e_{t-s,j}) = \sigma_e^2 \exp(-\alpha s) \quad (11)$$

which resolves to a first-order autoregressive structure (see **Time Series Analysis**) where the time intervals are equal.

The standard formulation for a repeated measures model is as a two-level structure where individual random effects are included to account for the covariance structure with correlated residuals. A simple such model with a random intercept u_{0j} and random ‘slope’ u_{1j} can be written as follows

$$y_{ij} = \sum_{h=0}^p \beta_h t_{ij}^h + u_{0j} + u_{1j} t_{ij} + e_{ij}$$

$$\text{cov}(\mathbf{e}_j) = \sigma_e^2 I, \quad \text{cov}(\mathbf{u}_j) = \Omega_u, \quad \mathbf{u}_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \quad (12)$$

This model incorporates the standard assumption that the covariance matrix of the level 1 residuals is diagonal, but we can allow it to have a more complex structure as in (11). In general, we can fit complex variance and covariance structures to the level 1 residual terms in any multilevel model. Furthermore, we can fit such structures at any level of a data hierarchy. A general discussion can be found in Goldstein [2, Chapter 3] and an application modeling the level 2 variance in a multilevel generalized linear model (see **Generalized Linear Mixed Models**) is given by Goldstein and Noden [4]; in the case of generalized linear models, the level 1 variance is heterogeneous by virtue of its dependence on the linear part of the model through the (nonlinear) link function.

Estimation

For normally distributed variables, the likelihood equations can be solved, iteratively, in a variety of ways. Goldstein et al. [3] describe an iterative generalized least squares procedure (see **Least Squares Estimation**) that will handle either linear models such as (7) or nonlinear ones such as (9) for both variances and covariances. Bayesian estimation can be carried out readily using Monte Carlo Markov Chain (MCMC) methods (see **Markov Chain Monte Carlo and Bayesian Statistics**), and a detailed comparison of likelihood and Bayesian estimation for models with complex variance structures is given in Browne et al. [1]. These authors also compare the fitting of linear and loglinear models for the variance.

Conclusions

This article has shown how to specify and fit a model that expresses the residual variance in a linear

model as a function of explanatory variables. These variables may or may not also enter the fixed, regression part of the model. It indicates how this can be extended to the case of multilevel models and to the general modeling of a covariance matrix. The example chosen shows how such models can uncover differences between groups and according to the values of a continuous variable. The finding that an interaction exists in the model for the variance underlines the need to apply considerations of model adequacy and fit for the variance modeling. The relationships exposed by modeling the variance will often be of interest in their own right, as well as better specifying the model under consideration.

References

- [1] Browne, W., Draper, D., Goldstein, H. & Rasbash, J. (2002). Bayesian and likelihood methods for fitting multilevel models with complex level 1 variation, *Computational Statistics and Data Analysis* **39**, 203–225.
- [2] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Edward Arnold, London.
- [3] Goldstein, H., Healy, M.J.R. & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data, *Statistics in Medicine* **13**, 1643–1655.
- [4] Goldstein, H. & Noden, P. (2003). Modelling social segregation, *Oxford Review of Education* **29**, 225–237.
- [5] Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. & Thomas, S. (1993). A multilevel analysis of school examination results, *Oxford Review of Education* **19**, 425–433.
- [6] Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. & Lewis, T. (2000). *A User's Guide to MLwiN*, 2nd Edition, Institute of Education, London.

(See also **Cross-classified and Multiple Membership Models**)

HARVEY GOLDSTEIN

Heuristics

ULRICH HOFFRAGE

Volume 2, pp. 795–795

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Heuristics

Heuristic comes from the Greek *heuriskein* meaning to find, hence *eureka* meaning I found it (out). Since its first use in English during the early 1800s, the term has acquired a range of meanings. For instance, in his Nobel prize-winning paper published in 1905 ('On a heuristic point of view concerning the generation and transformation of light'), Albert Einstein used the term to indicate that his view served to find out or to discover something [2]. Such a heuristic view may yield an incomplete and unconfirmed, eventually even false, but nonetheless useful picture. For the Gestalt psychologists who conceptualized thinking as an interaction between external problem structure and inner processes, heuristics (e.g., inspecting the problem, analyzing the conflict, the situation, the materials, and the goal) served the purpose of guiding the search for information in the environment and of restructuring the problem by internal processes [1]. In the 1950s and 60s, Herbert Simon and Allen Newell, two pioneers of Artificial Intelligence and cognitive psychology, used the term to refer to methods for finding solutions to problems. In formalized computer programs (e.g., the General Problem Solver), they implemented heuristics, for instance, the means-end analysis, which tried to set subgoals and to find

operations that finally reduced the distance between the current state and the desired goal state [5]. With the advent of information theory in cognitive psychology, the term heuristic came to mean a useful shortcut, an approximation, or a rule of thumb for searching through a space of possible solutions. Two prominent research programs in which heuristics play a key role are the *heuristics-and-biases program* [4] and the program of *simple heuristics* (often also referred to as **heuristics**, fast and frugal) [3].

References

- [1] Duncker, K. (1935). *Zur Psychologie des Produktiven Denkens*, [The psychology of productive thinking]. Julius Springer, Berlin.
- [2] Einstein, A. (1905). Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt, *Annalen der Physik* **17**, 132.
- [3] Gigerenzer, G., Todd, P.M. & The ABC Research Group. (1999). *Simple Heuristics that Make Us Smart*, Oxford University Press, New York.
- [4] Kahneman, D., Slovic, P. & Tversky, A. eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York.
- [5] Newell, A. & Simon, H. (1972). *Human Problem Solving*, Prentice Hall, Englewood Cliffs.

ULRICH HOFFRAGE

Heuristics: Fast and Frugal

ULRICH HOFFRAGE

Volume 2, pp. 795–799

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Heuristics: Fast and Frugal

To understand what fast and frugal **heuristics** are, it is helpful first of all to shed some light on the notion of *bounded rationality*, a term that has been coined by Herbert Simon [11]. In contrast to models that aim at finding the optimal solution to a problem at hand, models of bounded rationality take into account that humans often have only limited information, time, and computational capacities when making judgments or decisions. Given these constraints, the optimal solution is often unattainable. Moreover, many problems (e.g., chess) are too complex for the optimal solution to be found within a reasonable amount of time, even if all the relevant knowledge is available (i.e., all the rules and the positions of all the figures on the chess board are known) and the most powerful computers are used. Models of bounded rationality specify the (cognitive) processes that lead to a *satisficing* solution to a given problem, that is, to a solution that is both *satisfying* and *sufficing*.

Fast and frugal heuristics are such models of bounded rationality [4, 6]. They are task-specific, that is, they are designed to solve a particular task (e.g., choice, numerical estimation, categorization), but cannot solve tasks that they are not designed for – just like a hammer, which is designed to hammer nails but is useless for sawing a board. In fact, this task-specificity is key to the notion of the *adaptive toolbox* [5], the collection of heuristics that has evolved and can be used by the human mind.

Although fast and frugal heuristics differ with respect to the tasks they are designed to solve, they share the same guiding construction principles. In particular, they are composed of *building blocks*, which specify how information is searched for (*search rule*), when information search is stopped (*stopping rule*), and how a decision is made based on the information acquired (*decision rule*). These heuristics are fast for two reasons. First, they do not integrate the acquired information in a complex and time-consuming way. In this respect, many heuristics of the adaptive toolbox are as simple as possible because they do not combine pieces of information at all; instead, the decision is based on just one single reason (*one-reason decision making*). Secondly, they are fast as a consequence of being frugal, that

is, they stop searching for further information early in the process of information acquisition.

Fast and frugal heuristics are ecologically rational. In the present context, the notion of *ecological rationality* has two meanings. First, the performance of a heuristic is not evaluated against a norm, be it a norm from probability theory (*see Bayesian Statistics; Decision Making Strategies*) or logic (e.g., the *conjunction rule*, according to which the probability that an object belongs to both classes A and B cannot exceed the probability of belonging to class A). Rather, its performance is evaluated against a criterion that exists in the environment (i.e., in the ecology). This implies that (most) fast and frugal heuristics have been designed to make inferences about objective states of the world rather than to develop subjective preferences that reflect an individual's utilities. For instance, the QuickEst heuristic [8] makes inferences about the numerical values of objects (e.g., number of inhabitants of cities), and is evaluated by comparing estimated and true values. Secondly, a heuristic is ecologically rational to the extent that its building blocks reflect the structure of information in the environment. This fit of a heuristic to the environment in which it is evaluated is an important aspect of fast and frugal heuristics, which gave rise to a series of studies and important insights [13].

Studies on fast and frugal heuristics include (a) computer simulations to explore the performance of the heuristics in a given environment, in particular, in real-world environments (e.g., [1]), (b) the use of mathematical or analytical methods to explore when and why they perform as well as they do (eventually supported by simulations, in particular, in artificially created environments in which information structures are systematically varied) (e.g., [9]), and (c) experimental and observational studies to explore whether and when people actually use these heuristics (e.g., [10]). In the remainder of this entry, two heuristics (including their ecological rationality and the empirical evidence) are briefly introduced.

The Recognition Heuristic. Most people would agree that it is usually better to have more information than to have less. There are, however, situations in which partial ignorance is informative, which the recognition heuristic exploits. Consider the following question: Which city has more inhabitants, San

2 Heuristics: Fast and Frugal

Antonio, or San Diego? If you have grown up in the United States, you probably have a considerable amount of knowledge about both cities, and should do far better than chance when comparing the cities with respect to their populations. Indeed, about two-thirds of University of Chicago undergraduates got this question right [7]. In contrast, German citizens' knowledge of the two cities is negligible. So how much worse will they perform? The amazing answer is that within a German sample of participants, 100% answered the question correctly [7]. How could this be? Most Germans might have heard of San Diego, but do not have any specific knowledge about it. Even worse, most have never even heard of San Antonio. However, this difference with respect to name recognition was sufficient to make an inference, namely that San Diego has more inhabitants. Their lack of knowledge allowed them to use the recognition heuristic, which, in general, says: *If one of two objects is recognized and the other not, then infer that the recognized object has the higher value with respect to the criterion* [7, p. 76]. The Chicago undergraduates could not use this heuristic, because they have heard of both cities – they knew too much.

The ecological rationality of the recognition heuristic lies in the positive correlation between criterion and recognition values of cities (if such a correlation were negative, the inference would have to go in the opposite direction). In the present case, the correlation is positive, because larger cities (as compared to smaller cities) are more likely to be mentioned in mediators such as newspapers, which, in turn, increases the likelihood that their names are recognized by a particular person. It should thus be clear that the recognition heuristic only works if recognition is correlated with the criterion. Examples include size of cities, length of rivers, or productivity of authors; in contrast, the heuristic will probably not work when, for instance, cities have to be compared with respect to their mayor's age or their altitude above sea level. By means of mathematical analysis, it is possible to specify the conditions in which a *less-is-more effect* can be obtained, that is, the maximum percentage of recognized objects in the reference class that would increase the performance in a complete paired comparison task and the point from which recognizing more objects would lead to a decrease in performance [7]. In a series of experiments in which participants had to compare cities

with respect to their number of inhabitants, it could be shown that they chose the object predicted by the recognition heuristic in more than 90% of the cases – this was even true in a study in which participants were taught knowledge contradicting recognition [7]. Moreover, the authors could empirically demonstrate two less-is-more effects, one in which participants performed better in a domain in which they recognized a lower percentage of objects, and another one in which performance decreased through successively working on the same questions (so that recognition of objects increased during the course of the experiment).

Take The Best. If both objects are recognized in a pair-comparison task (see **Bradley–Terry Model**), the recognition heuristic does not discriminate between them. A fast and frugal heuristic that could be used in such a case is Take The Best. For simplicity, it is assumed that all cues (i.e., predictors) are binary (positive or negative), with positive cue values indicating higher criterion values. Take The Best is a simple, lexicographic strategy that consists of the following building blocks:

- (0) *Recognition heuristic:* see above.
- (1) *Search rule:* If both objects are recognized, choose the cue with the highest validity (where validity is defined as the percentage of correct inferences among those pairs of objects in which the cue discriminates) among those that have not yet been considered for this task. Look up the cue values of the two objects.
- (2) *Stopping rule:* If one object has a positive value and the other does not (i.e., has either a negative or unknown value), then stop search and go on to Step 3. Otherwise go back to Step 1 and search for another cue. If no further cue is found, then guess.
- (3) *Decision rule:* Infer that the object with the positive cue value has the higher value on the criterion.

Note that Take The Best's search rule ignores cue dependencies and will therefore most likely not establish the optimal ordering. Further note that the stopping rule does not attempt to compute an optimal stopping point at which the costs of further search exceed its benefits. Rather, the motto of the heuristic is 'Take The Best, ignore the rest'. Finally

note that Take The Best uses ‘one-reason decision making’ because its decision rule does not weight and integrate information, but relies on one cue only.

Another heuristic that also employs one-reason decision making is the Minimalist. It is even simpler than Take The Best because it does not try to order cues by validity, but chooses them in random order. Its stopping rule and its decision rule are the same as those of Take The Best.

What price does one-reason decision making have to pay for being fast and frugal? How much more accurate are strategies that use all cues and combine them? To answer these questions, Czerlinski, Gigerenzer, and Goldstein [1] evaluated the performance of various strategies in 20 data sets containing real-world structures rather than convenient multivariate normal structures; they ranged from having 11 to 395 objects, and from 3 to 19 cues. The predicted criteria included demographic variables, such as mortality rates in US cities and population sizes of German cities; sociological variables, such as drop-out rates in Chicago public high schools; health variables, such as obesity at age 18; economic variables, such as selling prices of houses and professors’ salaries; and environmental variables, such as the amount of rainfall, ozone, and oxidants. In the tests, half of the objects from each environment were randomly drawn. From all possible pairs within this training set, the order of cues according to their validities was determined (Minimalist used the training set only to determine whether a positive cue value indicates a higher or lower criterion). Thereafter, performance was tested both on the training set (fitting) and on the other half of the objects (generalization). Two

linear models were introduced as competitors: **multiple linear regression** and a simple unit-weight linear model [2]. To determine which of two objects has the higher criterion value, multiple regression estimated the criterion of each object, and the unit-weight model simply added up the number of positive cue values.

Table 1 shows the counterintuitive results obtained by averaging across frugality and percentages of correct choices in each of the 20 different prediction problems. The two simple heuristics were most frugal: they looked up fewer than a third of the cues (on average, 2.2 and 2.4 as compared to 7.7). What about the accuracy? Multiple regression was the winner when the strategies were tested on the training set, that is, on the set in which their parameters were fitted. However, when it came to predictive accuracy, that is, to accuracy in the hold-out sample, the picture changed. Here, Take The Best was not only more frugal, but also more accurate than the two linear strategies (and even Minimalist, which looked up the fewest cues, did not perform too far behind the two linear strategies). This result may sound paradoxical because multiple regression processed all the information that Take The Best did and more. However, by being sensitive to many features of the data – for instance, by taking correlations between cues into account – multiple regression suffered from overfitting, especially with small data sets (*see Model Evaluation*). Take The Best, on the other hand, uses few cues. The first cues tend to be highly valid and, in general, they will remain so across different subsets of the same class of objects. The stability of highly valid cues is a main factor for the robustness of Take The Best, that is, its low danger of overfitting in cross-validation as well as in other

Table 1 Performance of two fast and frugal heuristics (Take The Best and Minimalist) and two linear models (multiple regression and a unit-weight linear model) across 20 data sets. *Frugality* denotes the average number of cue values looked up; *Fitting* and *Generalization* refer to the performance in the training set and the test set, respectively (see text). Adapted from Gigerenzer, G. & Todd, P.M., and the ABC Research Group. (1999). *Simple heuristics that make us smart*, Oxford University Press, New York [6]

Strategy	Frugality	Accuracy (% correct)	
		Fitting	Generalization
Take The Best	2.4	75	71
Minimalist	2.2	69	65
Multiple Regression	7.7	77	68
Unit-weight linear model	7.7	73	69

forms of incremental learning. Thus, Take The Best can have an advantage against more savvy strategies that capture more aspects of the data, when the task requires making out-of-sample predictions (for other aspects of Take The Best's ecological rationality, see [9]).

There is meanwhile much empirical evidence that people use fast and frugal heuristics, in particular, when under time pressure or when information is costly (for a review of empirical studies see [3]). For other fast and frugal heuristics beyond the two introduced above, for instance QuickEst (for numerical estimation), Categorization-by-elimination (for categorization), RAFT (Reconstruction After Feedback with Take The Best, for an application to a memory phenomenon, namely *the hindsight bias*), the gaze heuristics (for catching balls on the playground), or various simple rules for terminating search through sequentially presented options, see [3, 5, 6, 12, 13]; for a discussion of this research program, see the commentaries and the authors' reply following [12].

References

- [1] Czerlinski, J., Gigerenzer, G. & Goldstein, D.G. (1999). How good are simple heuristics? in *Simple Heuristics that Make us Smart*, G. Gigerenzer, P.M. Todd, and the ABC Research Group, Oxford University Press, New York, pp. 97–118.
- [2] Dawes, R.M. (1979). The robust beauty of improper linear models in decision making, *American Psychologist* **34**, 571–582.
- [3] Gigerenzer, G. & Dieckmann, A. (in press). Empirical evidence on fast and frugal heuristics.
- [4] Gigerenzer, G. & Goldstein, D.G. (1996). Reasoning the fast and frugal way: models of bounded rationality, *Psychological Review* **103**, 650–669.
- [5] Gigerenzer, G. & Selten, R. (2001). *Bounded Rationality: The Adaptive Toolbox*, The MIT Press, Cambridge.
- [6] Gigerenzer, G., Todd, P.M., and the ABC Research Group. (1999). *Simple Heuristics that Make us Smart*, Oxford University Press, New York.
- [7] Goldstein, D.G. & Gigerenzer, G. (2003). Models of ecological rationality: the recognition heuristic, *Psychological Review* **109**, 75–90.
- [8] Hertwig, R., Hoffrage, U. & Martignon, L. (1999). Quick estimation: letting the environment do the work, in *Simple Heuristics that Make us Smart*, G. Gigerenzer, P.M. Todd, and the ABC Research Group, Oxford University Press, New York, pp. 209–234.
- [9] Martignon, L. & Hoffrage, U. (2002). Fast, frugal and fit: simple heuristics for paired comparison, *Theory and Decision* **52**, 29–71.
- [10] Rieskamp, J. & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? in G. Gigerenzer, P.M. Todd, and the ABC Research Group, *Simple Heuristics that make us Smart*, Oxford University Press, New York, pp. 141–167.
- [11] Simon, H. (1982). *Models of Bounded Rationality*, The MIT Press, Cambridge.
- [12] Todd, P.M. & Gigerenzer, G. (2000). Précis of *Simple heuristics that make us smart*, *Behavioral and Brain Sciences* **23**, 727–780.
- [13] Todd, P.M., Gigerenzer, G. & the ABC Research Group (in press). *Ecological Rationality of Simple Heuristics*.

ULRICH HOFFRAGE

Hierarchical Clustering

MORVEN LEESE

Volume 2, pp. 799–805

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hierarchical Clustering

Cluster analysis is a term for a group of multivariate methods that explore the similarities or differences between cases in order to find subgroups containing relatively homogenous cases (*see Cluster Analysis: Overview*). The cases may be, for example, patients with various symptoms, ideas arising from a focus group, clinics with different types of patient. There are two main types of cluster analysis: *optimization methods* that produce a single partition of the data (*see k-means Analysis*), usually into mutually exclusive (but possible overlapping) clusters, and *hierarchical clustering*, which is the subject of this article. Hierarchical clustering forms a nested series of mutually exclusive partitions or subgroups of the data, rather than a single partition. The process by which the series is formed is often displayed in a diagram known as a *dendrogram*. The investigator generally has to choose one of the partitions in the series as the ‘cluster solution’. Hierarchical clustering is possibly the most commonly applied type of cluster analysis and it is widely available in general statistical software packages.

Clusters are either successively divided, starting with a single cluster containing all individuals (*divisive clustering*) or they are successively merged, starting with n singleton clusters and ending with one large cluster containing all n individuals (*agglomerative clustering*). Both divisive and agglomerative techniques attempt to find the optimal step at each stage in the progressive subdivision or synthesis of the data. Divisions or fusions, once made, are irrevocable so that when an agglomerative algorithm has joined two individuals they cannot subsequently be separated, and when a divisive algorithm has made a split they cannot be reunited.

Proximities

Typically the process starts with a *proximity matrix* of the similarities or dissimilarities between all pairs of cases to be clustered (*see Proximity Measures*). The matrix may be derived from direct judgments: for example, in market research studies a number of subjects might be asked to assess the pairwise similarities of various foods using a visual analogue scale ranging from very dissimilar to very similar, their

average opinion giving a single value for each pair of foods. More commonly, however, the proximities are combined from similarities or differences defined on the basis of several different measurements, for example, saltiness, sweetness, sourness and so on.

The formula for combining similarities based on several variables depends on the type of data and the relative weight to be placed on different variables. For example binary data may be coded as series of 1s and 0s, denoting presence or absence of an attribute. In the case where each category is of equal weight, such as gender or white/nonwhite ethnic group, a *simple matching coefficient* (the proportion of matches between two individuals) could be used. However, if the attributes were the presence of various symptoms, proximity might be more appropriately measured using the asymmetric *Jaccard coefficient*, based on the proportion of matches where there is a positive match (i.e., ignoring joint negative matches). In genetics, binary matches may be assigned different weights depending on the part of the genetic sequence from which they arise. For continuous data, the *Euclidean distance* between individuals i and j , is often used:

$$s_{ij} = \left(\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right)^{1/2}, \quad (1)$$

where p is the number of variables and x_{ik} is the value of the k th variable for case i . Applied to binary data it is the same as the simple matching coefficient. The following is another example of a measure for continuous data, the (range-scaled) *city block* measure:

$$s_{ij} = 1 - \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{R_k}, \quad \text{where} \quad (2)$$

R_k is the range for the k th variable.

For data that contain both continuous and categorical variables, *Gower's coefficient* [6] has been proposed:

$$s_{ij} = \frac{\sum_{k=1}^p w_{ijk} s_{ijk}}{\sum_{k=1}^p w_{ijk}}, \quad (3)$$

where s_{ijk} is the similarity between the i th and j th individual as measured by the k th variable.

2 Hierarchical Clustering

For components of distance derived from binary or categorical data s_{ijk} takes the value of 1 for a complete match and 0 otherwise. For components derived from continuous data the range-scaled city block measure mentioned above is suggested. The value of w_{ijk} can be set to 0 or 1 depending on the whether the comparison is valid (for example, with a binary variable it can be set to 0 to exclude negative matches, as in the Jaccard coefficient); w_{ijk} can also be used to exclude similarity components when one or both values are missing for variable k .

Single Linkage Clustering

Once a proximity matrix has been defined, the next step is to form the clusters in a hierarchical sequence. There are many algorithms for doing this, depending on the way in which clusters are merged or divided. The algorithm usually entails defining proximity between clusters, as well as between individuals as outlined above. In one of the simplest hierarchical clustering methods, *single linkage* [14], also known as the *nearest neighbor technique*, the distance between clusters is defined as that between the closest pair of individuals, where only pairs consisting of one individual from each group are considered. Single linkage can be applied as an agglomerative method, or as a divisive method by initially splitting the data into two clusters with maximum nearest neighbor distance. The fusions made at each stage of agglomerative single linkage are now shown in a numerical example.

Consider the following distance matrix:

$$\mathbf{D}_1 = \begin{array}{ccccc} & 1 & 2 & 3 & 4 & 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{array}{c} 0.0 \\ 2.0 \\ 6.0 \\ 10.0 \\ 9.0 \end{array} & \begin{array}{c} \\ 0.0 \\ 5.0 \\ 9.0 \\ 8.0 \end{array} & \begin{array}{c} \\ \\ 0.0 \\ 4.0 \\ 5.0 \end{array} & \begin{array}{c} \\ \\ \\ 0.0 \\ 3.0 \end{array} & \begin{array}{c} \\ \\ \\ \\ 0.0 \end{array} \end{array} \quad (4)$$

The smallest entry in the matrix is that for individuals 1 and 2, consequently these are joined to form a two-member cluster. Distances between this cluster and the other three individuals are obtained as

$$\begin{aligned} d_{(1,2)3} &= \min[d_{13}, d_{23}] = d_{23} = 5.0 \\ d_{(1,2)4} &= \min[d_{14}, d_{24}] = d_{24} = 9.0 \\ d_{(1,2)5} &= \min[d_{15}, d_{25}] = d_{25} = 8.0 \end{aligned} \quad (5)$$

A new matrix may now be constructed whose entries are inter-individual and cluster-individual distances.

$$\mathbf{D}_2 = \begin{array}{ccccc} & (1, 2) & 3 & 4 & 5 \\ \begin{array}{c} (1, 2) \\ 3 \\ 4 \\ 5 \end{array} & \begin{array}{c} 0.0 \\ 5.0 \\ 9.0 \\ 8.0 \end{array} & \begin{array}{c} \\ 0.0 \\ 4.0 \\ 5.0 \end{array} & \begin{array}{c} \\ \\ 0.0 \\ 3.0 \end{array} & \begin{array}{c} \\ \\ \\ 0.0 \end{array} \end{array} \quad (6)$$

The smallest entry in $\mathbf{D}_2$ is that for individuals 4 and 5, so these now form a second two-member cluster and a new set of distances found:

$$\begin{aligned} d_{(1,2)3} &= 5.0 \text{ as before} \\ d_{(1,2)(4,5)} &= \min[d_{14}, d_{15}, d_{24}, d_{25}] = d_{25} = 8.0 \\ d_{(4,5)3} &= \min[d_{34}, d_{35}] = d_{34} = 4.0 \end{aligned} \quad (7)$$

These may be arranged in a matrix $\mathbf{D}_3$:

$$\mathbf{D}_3 = \begin{array}{ccccc} & (1, 2) & 3 & (4, 5) \\ \begin{array}{c} (1, 2) \\ 3 \\ (4, 5) \end{array} & \begin{array}{c} 0.0 \\ 5.0 \\ 8.0 \end{array} & \begin{array}{c} \\ 0.0 \\ 4.0 \end{array} & \begin{array}{c} \\ \\ 0.0 \end{array} \end{array} \quad (8)$$

The smallest entry is now $d_{(4,5)3}$ and individual 3 is added to the cluster containing individuals 4 and 5. Finally, the groups containing individuals 1,2 and 3,4,5 are combined into a single cluster.

Dendrogram

The process agglomerative process above is illustrated in a dendrogram in Figure 1. The *nodes* of

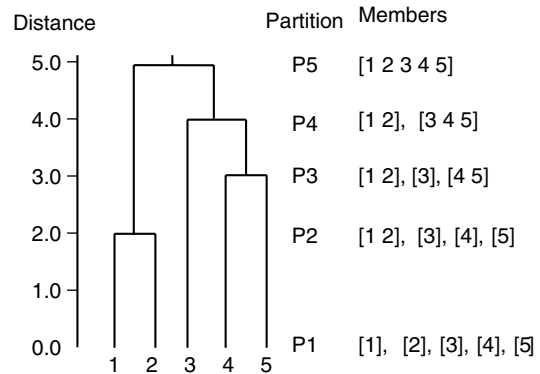


Figure 1 A dendrogram produced by single linkage hierarchical clustering. The process successively merges five individuals based on their pairwise distances (see text) to form a sequence of five partitions P1–P5

the dendrogram represent clusters and the lengths of the stems (*heights*) represent the dissimilarities at which clusters are joined. The same data and clustering procedure can give rise to 2^{n-1} dendrograms with different appearances, and it is usual for the software to choose an order for displaying the nodes that is optimal (in some sense). Drawing a line across the dendrogram at a particular height defines a particular partition or cluster solution (such that clusters below that height are distant from each other by at least that amount). The structure resembles an evolutionary tree and it is in applications where hierarchies are implicit in the subject matter, such as biology and anthropology, where hierarchical clustering is perhaps most relevant. In other areas it can still be used to provide a starting point for other methods, for example, optimization methods such as *k-means*.

While the dendrogram illustrates the hierarchical process by which series of cluster solutions are produced, low dimensional plots of the data (e.g., **principal component** plots) are more useful for interpreting particular solutions. Such plots can show the relationships among clusters, and among individual cases within clusters, which may not be obvious from a dendrogram. Comparisons between the mean levels or frequency distributions of individual variables within clusters, and the identification of representative members of the clusters (*centrotypes* or *exemplars* [19]) can also be useful. The latter are defined as the objects having the maximum within-cluster average similarity (or minimum dissimilarity), for example, the *medoid* (the object with the minimum absolute distance to the other members of the cluster).

The dendrogram can be regarded as representing the original relationships amongst the objects, as implied by their observed proximities. Its success in doing this can be measured using the *cophenetic matrix*, whose elements are the heights where two objects become members of the same cluster in the dendrogram. The product-moment correlation between the entries in the cophenetic matrix and the corresponding ones in the proximity matrix (excluding 1s on the diagonals) is known as the *cophenetic correlation*. Comparisons using the cophenetic correlation can also be made between different dendrograms representing different clusterings of the same data set. Dendrograms can be compared using randomization tests to assess the statistical significance of the cophenetic correlation [11].

Other Agglomerative Clustering Methods

Single linkage operates directly on a proximity matrix. Another type of clustering, *centroid clustering* [15], however, requires access to the original data. To illustrate this type of method, it will be applied to the set of bivariate data shown in Table 1.

Suppose Euclidean distance is chosen as the inter-object distance measure, giving the following distance matrix:

$$\mathbf{D}_1 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{matrix} 0.0 \\ 1.0 & 0.0 \\ 5.39 & 5.10 & 0.0 \\ 7.07 & 7.0 & 2.24 & 0.0 \\ 7.07 & 7.28 & 3.61 & 2.0 & 0.0 \end{matrix} \end{matrix} \quad (9)$$

Examination of $\mathbf{D}_1$ shows that d_{12} is the smallest entry and objects 1 and 2 are fused to form a group. The mean vector (centroid) of the group is calculated (1,1.5) and a new Euclidean distance matrix is calculated.

$$\mathbf{D}_2 = \begin{matrix} & \begin{matrix} (1, 2) & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} (1, 2) \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{matrix} 0.0 \\ 5.22 & 0.0 \\ 7.02 & 2.24 & 0.0 \\ 7.16 & 3.61 & 2.0 & 0.0 \end{matrix} \end{matrix} \quad (10)$$

The smallest entry in $\mathbf{D}_2$ is d_{45} and objects 4 and 5 are therefore fused to form a second group, the mean vector of which is (8.0,1.0). A further distance matrix $\mathbf{D}_3$ is now calculated.

$$\mathbf{D}_3 = \begin{matrix} & \begin{matrix} (1, 2) & 3 & (4, 5) \end{matrix} \\ \begin{matrix} (1, 2) \\ 3 \\ (4, 5) \end{matrix} & \begin{matrix} 0.0 \\ 5.22 & 0.0 \\ 7.02 & 2.83 & 0.0 \end{matrix} \end{matrix} \quad (11)$$

In $\mathbf{D}_3$ the smallest entry is $d_{(45)3}$ and so objects 3,4, and 5 are merged into a three-member cluster. The final stage consists of the fusion of the two remaining groups into one.

Table 1 Data on five objects used in example of centroid clustering

Object	Variable 1	Variable 2
1	1.0	1.0
2	1.0	2.0
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

4 Hierarchical Clustering

Different definitions of intergroup proximity give rise to different agglomerative methods. *Median linkage* [5] is similar to centroid linkage except that the centroids of the clusters to be merged are weighted equally to produce the new centroid of the merged cluster, thus avoiding the more numerous of the pair of clusters dominating. The new centroid is intermediate between the two constituent clusters. In the centroid linkage shown above Euclidean distance was used, as is usual. While other proximity measures are possible with centroid or median linkage, they would lack interpretation in terms of the raw data. *Complete linkage* (or *furthest neighbor*) [16], is opposite to single linkage, in the sense that distance between groups is now defined as that of the most distant pair of individuals (the *diameter* of the cluster). In (*group*) *average linkage* [15], the distance between two clusters is the average of the distance between all pairs of individuals that are made up of one individual from each group. Average, centroid, and median linkage are also known as UPGMA, UPGMC, and WPGMC methods respectively (U: unweighted; W: weighted; PG: pair group; A: average; C: centroid).

Ward introduced a method based on minimising an objective function at each stage in the hierarchical process, the most widely used version of which is known as *Ward's method* [17]. The objective at each stage is to minimize the increase in the total within-cluster error sum of squares. This increase is in fact a function of the weighted Euclidean distance between the centroids of the merged clusters. Lance and William's *flexible* method is defined by values of the parameters of a general recurrence formula [10] and many of the standard methods mentioned above can be defined in terms of the parameters of the Lance and Williams formulation.

Divisive Clustering Methods

As mentioned earlier, divisive methods operate in the other direction from agglomerative methods, starting with one large cluster and successively splitting clusters. *Polythetic divisive* methods are relatively rarely used and are more akin to the agglomerative methods discussed above, since they use all variables simultaneously, and are computationally demanding. For data consisting of binary variables, however,

relatively simple and computationally efficient *monothetic divisive* methods are available. These methods divide clusters according to the presence or absence of each variable, so that at each stage clusters contain members with certain attributes that are either all present or all absent.

Instead of cluster homogeneity, the attribute used at each step in a divisive method can be chosen according to its overall association with other attributes: this is sometimes termed *association analysis* [18]. The split at each stage is made according to the presence or absence of the attribute whose association with the others (i.e., the summed criterion above) is a maximum. For example for one pair of variables V_i and V_j with values 0 and 1 the frequencies observed might be as follows:

	V_i	1	0
V_j			
	1	a	b
	0	c	d

Examples of measures of association based on these frequencies (summed over all pairs of variables) are $|ad - bc|$ and $(ad - bc)^2 n / [(a + b)(a + c)(b + d)(c + d)]$. Hubálek gives a review of 43 such coefficients [7]. A general problem with this method is that the possession of a particular attribute, which is either rare or rarely found in combination with others, may take an individual down the 'wrong' path.

Choice of Number of Clusters

It is often the case that an investigator is not interested in the complete hierarchy but only in one or two partitions obtained from it (or 'cluster solutions'), and this involves deciding on the number of groups present. In standard agglomerative or polythetic divisive clustering, partitions are achieved by selecting one of the solutions in the nested sequence of clusterings that comprise the hierarchy. This is equivalent to cutting a dendrogram at an optimal height (this choice sometimes being termed the *best cut*). The choice of height is generally based on large changes in fusion levels in the dendrogram, and a *scree-plot* of height against number of clusters can be used as an informal guide. A relatively widely used formal test procedure is based on the relative sizes of the different fusion levels [13], and a number of other formal approaches for determining

the number of clusters have been reviewed [12] (*see Number of Clusters*).

Choice of Cluster Method

Apart from the problem of deciding on the number of clusters, the choice of the appropriate method in

any given situation may also be difficult. Hierarchical clustering algorithms are only stepwise optimal, in the sense that at any stage the next step is chosen to be optimal in some sense but that may not guarantee the globally optimal partition, had all possibilities had been examined. Empirical and theoretical studies have rarely been conclusive. For example,

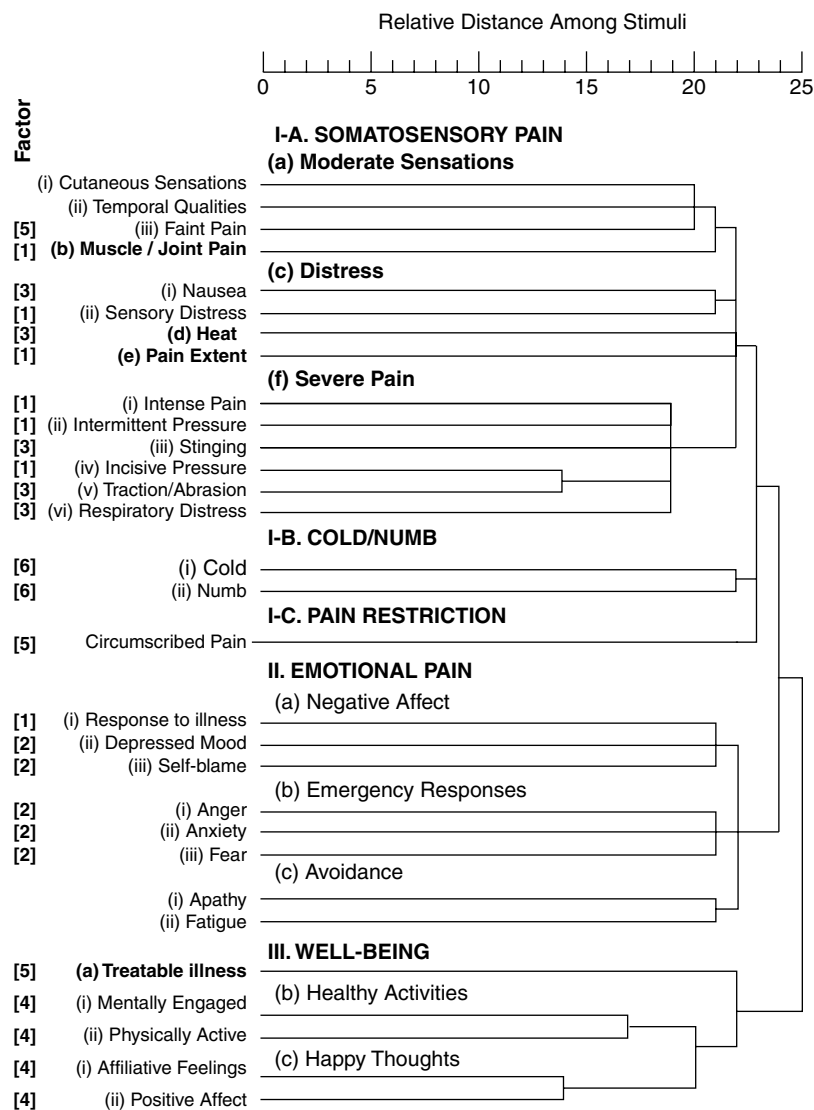


Figure 2 Dendrogram produced by cluster analysis of similarity judgments of pain descriptors obtained from healthy volunteers, using average-linkage cluster analysis. The data are healthy peoples' responses to the descriptors in the MAPS (Multidimensional Affect and Pain Survey). The dendrogram has been cut at 30 clusters and also shows 'superclusters' joining at higher distances. A separate factor analysis obtained from patients' responses of the 30-cluster concepts found six factors, indicated along the left-hand side

single linkage tends to have satisfactory mathematical properties and is also easy to program and apply to large data sets, but suffers from ‘chaining’, in which separated clusters with ‘noise’ points in between them tend to be joined together. Ward’s method often appears to work well but may impose a spherical structure where none exists and can be sensitive to outliers. Relatively recently, however, hierarchical methods based on the *classification likelihood* have been developed [1]. Such model-based methods are more flexible than the standard methods discussed above, in that they assume an underlying mixture of multivariate normal distributions and hence allow the detection of elliptical clusters, possibly with varying sizes and orientations. They also have the advantage of providing an associated test of the weight of evidence for varying numbers of clusters (*see Model Based Cluster Analysis; Finite Mixture Distributions*).

An Example of Clustering

As a practical example, Figure 2 shows the top part of a dendrogram resulting from clustering pain concepts [2]. The similarities between 101 words describing pain (from the Multidimensional Affect and Pain Survey) were directly assessed by a panel of health people using a *pile-sort* procedure, and these similarities were analyzed using average linkage clustering. Thirty pain concept clusters (e.g., ‘cutaneous sensations’) could be further grouped into ‘superclusters’ (e.g., ‘somatosensory pain’) and these clusters and superclusters are shown on the dendrogram. Individual pain descriptors that comprise the clusters are not shown (the dendrogram has been ‘cut’ at the 30-cluster level). The results of a factor analysis of responses by cancer outpatients is shown along the left-hand edge, and this was used to validate the structure of the MAPS pain concepts derived from the cluster analysis.

Further Information

General reviews of cluster analysis are available that include descriptions of hierarchical methods and their properties, and examples of their application [3,4], including one that focuses on robust methods [9]. Recently, specialist techniques have been developed for newly expanding subject areas such as genetics [8].

References

- [1] Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**, 803–821.
- [2] Clark, W.C., Kuhl, J.P., Keohan, M.L., Knotkova, H., Winer, R.T. & Griswold, G.A. (2003). Factor analysis validates the cluster structure of the dendrogram underlying the multidimensional affect and pain survey (MAPS) and challenges the a priori classification of the descriptors in the McGill Pain Questionnaire (MPQ), *Pain* **106**, 357–363.
- [3] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th Edition, Edward Arnold, London.
- [4] Gordon, A.E. (1999). *Classification*, 2nd Edition, Chapman & Hall, New York.
- [5] Gower, J.C. (1967). A comparison of some methods of cluster analysis, *Biometrics* **23**, 623–628.
- [6] Gower, A.E. (1971). A general coefficient of similarity and some of its properties, *Biometrics* **27**, 857–872.
- [7] Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation, *Biological Review* **57**, 669–689.
- [8] Jajuga, K., Sokolowski, A. & Bock, H.-H., eds (2002). *Classification, Clustering and Data Analysis: Recent Advances and Applications*, Springer-Verlag, Berlin.
- [9] Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*, Wiley, New York.
- [10] Lance, G.N. & Williams, W.T. (1967). A general theory of classificatory sorting strategies: 1 hierarchical systems, *Computer Journal* **9**, 373–380.
- [11] Lapointe, F.-J. & Legendre, P. (1995). Comparison tests for dendrograms: a comparative evaluation, *Journal of Classification* **12**, 265–282.
- [12] Milligan, G.W. & Cooper, M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research* **21**, 41–58.
- [13] Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation, *Computer Journal* **20**, 359–363.
- [14] Sneath, P.H.A. (1957). The application of computers to taxonomy, *Journal of General Microbiology* **17**, 201–226.
- [15] Sokal, R.R. & Michener, C.D. (1958). A Statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* **38**, 1409–1438.
- [16] Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, *Biologiske Skrifter* **5**, 34.
- [17] Ward, J.H. (1963). Hierarchical groupings to optimize an objective function, *Journal of the American Statistical Association* **58**, 236–244.

- [18] Williams, W.T. & Lambert, J.M. (1959). Multivariate methods in plant ecology, 1 association analysis in plant communities, *Journal of Ecology* **47**, 83–101.
- [19] Wishart, D. (1999). Clustangraphics 3: interactive graphics for cluster analysis, in *Classification in the Information Age*, W. Gaul & H. Locarek-Junge, eds, Springer-Verlag, Berlin, pp. 268–275.

(See also **Additive Tree; Cluster Analysis: Overview; Fuzzy Cluster Analysis; Overlapping Clusters; Two-mode Clustering**)

MORVEN LEESE

Hierarchical Item Response Theory Modeling

DANIEL BOLT AND JEE-SEON KIM

Volume 2, pp. 805–810

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hierarchical Item Response Theory Modeling

Item response data are frequently collected in settings where the objects of measurement are hierarchically nested (e.g., students within schools, clients within therapists, repeated measures within persons). Such data structures commonly lead to statistical dependence among observations. For example, test scores from students attending the same school are usually more similar than scores from students attending different schools. While this creates problems for statistical analyses that require independence, it also offers opportunities for understanding the nature of contextual influence [20], such as characteristics of effective schools. **Hierarchical models**, referred to elsewhere as *multilevel*, *random coefficients*, or *mixed effects* models, account for this dependence through the use of random effects (see **Linear Multilevel Models**).

For traditional users of hierarchical models, the synthesis of hierarchical models with item response theory (IRT) (see **Item Response Theory (IRT) Models for Polytomous Response Data & Intrinsic Linearity**) allows the practical advantages of the latent trait metric (see **Latent Variable**) in IRT to be realized in hierarchical settings where observed measures (e.g., test scores) can be problematic [17]. For traditional users of IRT, hierarchical extensions allow more complex IRT models that can account for the effects that higher level units (e.g., schools) have on lower level units (e.g., students) [9, 13]. Hierarchical IRT is often portrayed more generally than this, however, as even the simplest of IRT models can be viewed as hierarchical models where multiple item responses are nested within persons [2, 14]. From this perspective, hierarchical modeling offers a very broad framework within which virtually all IRT models and applications (e.g., differential item functioning; equating) can be unified and generalized, as necessary, to accommodate unique sources of dependence that may arise in different settings [1, 19]. In this entry, we consider two classes of hierarchical IRT models, one based on a nesting of item responses within persons, the other a nesting of item responses within both persons and items.

Hierarchical IRT Models with Random Person Effects

In one form of hierarchical IRT modeling, item responses (Level 1 units) are portrayed as nested within persons (Level 2 units) [2]. A Level-1 model, $P(U_{ij} = m | \theta_i, \xi_j)$, also called a *within-person model*, expresses the probability of item score m conditional upon person parameters, denoted θ_i (generally latent trait(s)), and item parameters, denoted ξ_j (e.g., item difficulty). Any of the common dichotomous or polytomous IRT models (see **Item Response Theory (IRT) Models for Polytomous Response Data**) (see [4] for examples) could be considered as within-person models.

At Level 2, variability in the person parameters is modeled through a *between-person* model. In hierarchical IRT, it is common to model the latent trait as a function of other person variables, such as socioeconomic status or gender. For example, a between-person model might be expressed as:

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \tau_i \quad (1)$$

where $X_{i1}, X_{i2}, \dots, X_{ip}$ denote p person variables, β_0 is an intercept parameter, $\beta_1, \dots, \beta_p$ are regression coefficient parameters, and τ_i a residual, commonly assumed to be normally distributed with mean 0 and variance τ_r . It is the presence of this residual that makes the person effect a random effect. The item parameters of the IRT model, usually assumed constant across persons, represent fixed effects.

The combination of a within-person model and a between-person model as described above provides a multilevel representation of Zwinderman's [25] manifest predictors model. Other existing IRT models can also be viewed as special cases. For example, a multigroup IRT model [3] uses group identifier variables as predictors in the between-person model. Traditional IRT models (e.g., Rasch, two-parameter models) have no predictors.

There are several advantages of portraying IRT models within a hierarchical framework [10, 13]. First, it allows covariates of the person traits to be used as collateral information for IRT estimation. This can lead to improved estimation of not only the person traits, but also (indirectly) the item parameters [11, 12]. (However, the latter effects tend to be quite small in most practical applications [2].)

A second advantage is improved estimation of relationships between person variables and the latent

2 Hierarchical Item Response Theory Modeling

traits. A hierarchical IRT model avoids problems of attenuation bias that are introduced when using a two-step estimation procedure, specifically, one that first estimates the person traits using an IRT model, and then in a separate analysis regresses the trait estimates on the person variables. When these analyses are executed simultaneously, as in hierarchical IRT, the regression coefficient estimates are based on the latent traits and thus are not attenuated due to estimation error.

A third advantage of hierarchical IRT is its capacity to include additional levels above persons [9, 13]. To illustrate, we consider a three-level dataset from the 1999 administration of the mathematics section of the Texas Assessment of Academic Skills (TAAS). The dataset contains correct/incorrect item responses to 52 items for a sample of 26,289 fifth-graders from 363 schools. Student variables related to socioeconomic status (FLUNCH=1 implies free or reduced-price lunch, 0=regular-price lunch) and gender (GENDER=1 implies female, 0=male) were also considered. Here we analyze just 20 of the 52 items. The three-level model involves item responses nested with students, and students nested within schools.

For the within-person model, we use a **Rasch model**. In a Rasch model, the probability of correct item response is modeled through an item difficulty parameter b_j and a single person trait parameter θ_{ik} , the latter now double-indexed to identify student i from school k :

$$P(U_{ik,j} = 1 | \theta_{ik}, b_j) = \frac{\exp(\theta_{ik} - b_j)}{1 + \exp(\theta_{ik} - b_j)}. \quad (2)$$

At Level 2, between-student variability is modeled as:

$$\theta_{ik} = \beta_{0k} + \beta_{1k}\text{FLUNCH}_{ik} + \beta_{2k}\text{GENDER}_{ik} + r_{ik}, \quad (3)$$

where β_{0k} , β_{1k} , and β_{2k} are the intercept and regression coefficient parameters for school k ; r_{ik} is a normally distributed residual with mean zero and variance τ_r .

Next, we add a third level associated with school. At the school level, we can account for the possibility that certain effects at the student level (represented by the coefficients β_{0k} , β_{1k} , β_{2k}) may vary across schools. In the current model, we allow for between-school variability in the intercepts

(β_{0k}) and the within-school FLUNCH effects (β_{1k}). We also create a school variable, FLUNCH.SCH $_k$, the mean of FLUNCH across all students within school k , to represent the average socioeconomic status of students within the school. The variable FLUNCH.SCH $_k$ is added as a predictor both of the school intercepts and the within-school FLUNCH effect. This results in the following Level-3 (between-school) model:

$$\beta_{0k} = \gamma_{00} + \gamma_{01}\text{FLUNCH.SCH}_k + u_{0k}, \quad (4)$$

$$\beta_{1k} = \gamma_{10} + \gamma_{11}\text{FLUNCH.SCH}_k + u_{1k}, \quad (5)$$

$$\beta_{2k} = \gamma_{20}. \quad (6)$$

In this representation, each of the γ parameters represents a fixed effect, while the u_{0k} and u_{1k} are random effects associated with the school intercepts and school FLUNCH effects, respectively. Across schools, we assume (u_{0k}, u_{1k}) to be bivariate normally distributed with mean (0,0) and covariance matrix $\mathbf{T}_u$, having diagonal elements $\tau_{u_{00}}$ and $\tau_{u_{11}}$. (Note that by omitting a similar residual for β_{2k} , the effects of GENDER are assumed to be the same, that is, fixed, across schools.) Variations on the above model could be considered by modifying the nature of effects (fixed versus random) and/or predictors of the effects.

To estimate the model above, we follow a procedure described by Kamata [9]. In Kamata's method, a hierarchical IRT model is portrayed as a hierarchical **generalized linear model**. Random person effects are introduced through a random intercept, while item effects are introduced through the fixed (across persons) coefficients of item-identifier dummy variables at Level 1 of the model (see [9] for details). In this way, the model can be estimated using a quasi-likelihood algorithm implemented for generalized linear models in the software program HLM [16].

A portion of the results is shown in Tables 1 and 2. In Table 1, the fixed effect estimates are seen to be statistically significant for FLUNCH, FLUNCH.SCH, and GENDER, implying lower levels of math ability on average for students receiving free or reduced-price lunch within a school ($\hat{\gamma}_{10} = -.53$), and also lower (on average) abilities for students coming from schools that have a larger percentage of students that receive free or reduced-price lunch ($\hat{\gamma}_{01} = -.39$). The effect for gender is significant but weak ($\hat{\gamma}_{20} = .04$), with females having slightly higher ability. No

Table 1 Estimates of fixed and random effect parameters in multilevel Rasch model, Texas assessment of academic skills data

Fixed effect estimates	Coeff	se	t-stat	Approx. df	P value
γ_{00} , intercept	1.53	.04	34.24	361	.000
γ_{01} , FLUNCH.SCH	-.39	.06	-6.15	361	.000
γ_{10} , FLUNCH	-.53	.04	-12.04	361	.000
γ_{11} , FLUNCH $\times$ FLUNCH.SCH	.05	.08	.67	361	.503
γ_{20} , GENDER	.04	.01	3.02	26824	.003
Variance estimates	Variance	χ^2	df	P value	Reliability
τ_r	1.26	48144	23445	.000	.927
$\tau_{u_{00}}$	.16	2667	227	.000	.888
$\tau_{u_{11}}$	.05	383	227	.000	.345

significant interaction was detected for FLUNCH and FLUNCH.SCH.

The variance estimates, also shown in Table 1, suggest significant between-school variability both in the residual intercepts ($\hat{\tau}_{u_{00}} = .16$) and in the residual FLUNCH effects ($\hat{\tau}_{u_{11}} = .05$). This implies that even after accounting for the effects of FLUNCH.SCH, there remains significant between-school variability in both the mean ability levels of non-FLUNCH students, and in the within-school effects of FLUNCH. Likewise, significant between-student variability remains across students within school ($\tau_r = 1.26$) even after accounting for the effects of FLUNCH and GENDER. Recall that Rasch item difficulties are also estimated as fixed effects in the model. These estimates (not shown here) ranged from -2.49 to 2.06.

Table 2 provides empirical Bayes estimates of the residuals for three schools (see **Random Effects in Multivariate Linear Models: Prediction**). Such estimates allow a school-specific inspection of the two effects allowed to vary across schools (in this model, the intercept, u_{0k} , and FLUNCH, u_{1k} , effects). More specifically, they indicate how each school's effect departs from what is expected given

the corresponding fixed effects (in this model, the fixed intercept and FLUNCH.SCH effects). These estimates illustrate another way in which hierarchical IRT can be useful, namely, its capacity to provide group-level assessment. Recalling that both residuals have means of zero across schools, we observe that School 1 has a lower intercept (-1.03), implying lower ability levels for non-FLUNCH students, and a more negative FLUNCH effect (-0.26), than would be expected given the school's level on FLUNCH.SCH (.30). School 2, which has a much higher proportion of FLUNCH students than School 1 (.93), has a higher than expected intercept and a more negative than expected FLUNCH effect, while School 3, with FLUNCH.SCH=.17, has a higher than expected intercept, but an FLUNCH effect that is equivalent to what is expected.

Despite the popularity of Kamata's method, it is limited to use with the Rasch model. Other estimation methods have been proposed for more general models. For example, other models within the Rasch family (e.g., Master's partial credit model), can be estimated using a general EM algorithm in the software CONQUEST [24]. Still more general models, such as hierarchical two parameter IRT models, can

Table 2 Examples of empirical Bayes estimates for individual schools, Texas assessment of academic skills data

School	N	Empirical Bayes estimates				FLUNCH.SCH	Fixed effects	
		$\hat{u}_0$	se	$\hat{u}_1$	se		$\hat{\gamma}_{00} + \hat{\gamma}_{01}\text{FLUNCH.SCH}$	$\hat{\gamma}_{10} + \hat{\gamma}_{11}\text{FLUNCH.SCH}$
1	196	-1.03	.01	-.26	.02	.30	1.41	-.51
2	157	.91	.01	-.23	.04	.93	1.16	-.48
3	62	.34	.02	-.00	.04	.17	1.46	-.52

N = number of students; FLUNCH.SCH = proportion of students receiving free or reduced-price lunch.

be estimated using fully Bayesian methods, such as Gibbs' sampling [6, 7]. Such procedures are appealing in that they provide a full characterization of the joint posterior distribution of model parameters, as opposed to point estimates. Because they are easy to implement, they also permit greater flexibility in manipulation of other features of the hierarchical IRT model. For example, Maier [10] explores use of alternative distributions for the residuals (e.g., inverse chi-square, uniform) in a hierarchical Rasch model.

Several advantages of hierarchical IRT modeling can be attributed to its use of a latent trait. First, the use of a **latent variable** as the outcome in the between-person model allows for more realistic treatment of measurement error. When modeling test scores as outcomes, for example, variability in the standard error of measurement across persons is not easily accounted for, as a common residual variance applies for all persons. Second, the invariance properties of IRT allow it to accommodate a broader array of data designs, such as matrix sampled educational assessments (as in the National Assessment of Educational Progress), or others that may involve missing data. Finally, the interval level properties of the IRT metric can be beneficial. For example, Raudenbush, Samson, and Johnson [16] note the value of a Rasch trait metric when modeling self-reported criminal behavior across neighborhoods, where simple counts of crime tend to produce observed scores that are highly skewed and lack interval scale properties.

Hierarchical IRT Models with Both Random Person and Item Effects

Less common, but equally useful, are hierarchical IRT models that model random item effects. Such models typically introduce a between-item model in which the item parameters of the IRT model become outcomes. Modeling the predictive effects of item features on item parameters can be very useful. Advantages include improved estimation of the item parameters (i.e., collateral information), as well as information about item features that can be useful for item construction and item-level validity checks [4]. A common IRT model used for this purpose is Fischer's [5] linear logistic test model [LLTM]. In the LLTM, Rasch item difficulty is expressed as a weighted linear combination of

prespecified item characteristics, such as the cognitive skill requirements of the item. Hierarchical IRT models can extend models such as the LLTM by allowing item parameters to be random, thus allowing for less-than-perfect prediction of the item difficulty parameters [8].

A hierarchical IRT model with random person and item effects can be viewed as possessing two forms of nesting, as item responses are nested within both persons and items. Van den Noortgate, De Boeck, and Meulders [21] show how random item effect models can be portrayed as *cross-classified hierarchical models* [20] in that each item response is associated with both a person and item. With cross-classified models, it is possible to consider not only main effects associated with item and person variables, but also item-by-person variables [19, 21]. This further extends the range of applications that can be portrayed within the hierarchical IRT framework. For example, hierarchical IRT models such as the random weights LLTM [18], where the LLTM weights vary randomly across persons, and dynamic Rasch models [22], where persons learn over the course of a test [22], can both be portrayed in terms of item-by-person covariates [19]. Similarly, IRT applications such as differential item functioning can be portrayed in a hierarchical IRT model where the product of person group by item is a covariate [19].

Additional levels of nesting can also be defined for the items. For example, in the hierarchical IRT model of Janssen, Tuerlinckx, Meulders, and De Boeck [8], items are nested within target content categories. An advantage of this model is that a prototypical item for each category can be defined, thus allowing criterion-related classification decisions based on each person's estimated trait level.

Different estimation strategies have been considered for random person and item effect models. Using the cross-classified hierarchical model representation, Van den Noortgate et al. [21] propose use of quasi-likelihood procedures implemented in the SAS-macro GLIMMIX [23]. Patz and Junker [14] presented a very general Markov chain Monte Carlo strategy for hierarchical IRT models that can incorporate both item and person covariates. A related application is given in Patz, Junker, Johnson, and Moriano [15], where dependence due to rater effects is addressed. General formulations such as this offer the clearest indication of the future potential for hierarchical IRT

models, which should continue to offer the methodologist exciting new ways of investigating sources of hierarchical structure in item response data.

References

- [1] Adams, R.J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model, *Applied Psychological Measurement* **21**, 1–23.
- [2] Adams, R.J., Wilson, M. & Wu, M. (1997). Multilevel item response models: an approach to errors in variables regression, *Journal of Educational and Behavioral Statistics* **22**, 47–76.
- [3] Bock, R.D. & Zimowski, M.F. (1997). Multi-group IRT, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer-Verlag, New York, pp. 433–448.
- [4] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*, Lawrence Erlbaum, Mahwah.
- [5] Fisher, G.H. (1973). Linear logistic test model as an instrument in educational research, *Acta Psychologica* **37**, 359–374.
- [6] Fox, J.P. (in press) *Multilevel IRT Using Dichotomous and Polytomous Response Items*. British Journal of Mathematical and Statistical Psychology.
- [7] Fox, J.-P. & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling, *Psychometrika* **66**, 271–288.
- [8] Janssen, R., Tuerlinckx, F., Meulders, M. & De Boeck, P. (2000). A hierarchical IRT model for criterion referenced measurement, *Journal of Educational and Behavioral Statistics* **25**, 285–306.
- [9] Kamata, A. (2001). Item analysis by the hierarchical generalized linear model, *Journal of Educational Measurement* **38**, 79–93.
- [10] Maier, K. (2001). A Rasch hierarchical measurement model, *Journal of Educational and Behavioral Statistics* **26**, 307–330.
- [11] Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters, *Applied Psychological Measurement* **11**, 81–91.
- [12] Mislevy, R.J., & Sheehan K.M. (1989). The role of collateral information about examinees in item parameter estimation, *Psychometrika* **54**, 661–679.
- [13] Pastor, D. (2003). The use of multilevel item response theory modeling in applied research: an illustration, *Applied Measurement in Education* **16**, 223–243.
- [14] Patz, R.J. & Junker, B.W. (1999). Application and extensions of MCMC in IRT: multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics* **24**, 342–366.
- [15] Patz, R.J., Junker, B.W., Johnson, M.S. & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data, *Journal of Educational and Behavioral Statistics* **27**, 341–384.
- [16] Raudenbush, S.W., Bryk, A.S., Cheong, Y.F. & Congdon, R. (2001). *HLM 5: Hierarchical Linear and Non-linear Modeling*, 2nd Edition, Scientific Software International, Chicago.
- [17] Raudenbush, S.W., Johnson, C. & Sampson, R.J. (2003). A multivariate multilevel Rasch model with application to self-reported criminal behavior, *Sociological Methodology* **33**, 169–211.
- [18] Rimjen, F. & De Boeck, P. (2002). The random weights linear logistic test model, *Applied Psychological Measurement* **26**, 269–283.
- [19] Rimjen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory, *Psychological Methods* **8**, 185–205.
- [20] Snijders, T. & Bosker, R. (1999). *Multilevel Analysis*, Sage Publications, London.
- [21] Van den Noortgate, W., De Boeck, P. & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics, *Journal of Educational and Behavioral Statistics* **28**, 369–386.
- [22] Verhelst, N.D. & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model, *Psychometrika* **58**, 395–415.
- [23] Wolfinger, R. & O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation* **48**, 233–243.
- [24] Wu, M.L., Adams, R.J. & Wilson, M.R. (1997). *Conquest: Generalized Item Response Modeling Software [Software manual]*, Australian Council for Educational Research, Melbourne.
- [25] Zwinderman, A.H. (1991). A generalized Rasch model for manifest predictors, *Psychometrika* **56**, 589–600.

Further Reading

De Boeck, P. & Witson, M. (2004). *Explanatory Item Response Models*, Springer-Verlag, New York.

DANIEL BOLT AND JEE-SEON KIM

Hierarchical Models

ROBERT E. PLOYHART

Volume 2, pp. 810–816

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hierarchical Models

Many types of data in the behavioral sciences are inherently nested or hierarchical [15]. For example, students are nested within classes that are nested within schools. Different cells are nested within different regions of the brain. A consequence of this nesting is that within-unit observations are unlikely to be independent; to some extent, within-unit observations (e.g., students within classes) are correlated or more similar to each other than between-unit observations (e.g., students from different classes). The statistical consequences of nonindependence can be severe [7], and the assumptions of traditional statistical models, such as the **generalized linear model** (GLM), become violated. This in turn may influence the size of the standard errors and, hence, statistical significance tests.

Hierarchical linear models (HLM; sometimes referred to as random coefficient, mixed-effects, or

multilevel models, *see* **Generalized Linear Mixed Models** and **Linear Multilevel Models**) represent a class of models useful for testing substantive questions when the data are inherently nested or when nonindependence of observations is a concern. This chapter is designed to be a simple introduction to the HLM. Readers should consult these excellent sources for technical details [1, 2, 4–6, 8, 10, 14].

A Simple Introduction to the HLM

Figure 1 visually shows the typical GLM model (here, a regression-based model) and contrasts it to the HLM. Below each figure is the statistical model ((1) for the GLM and (2–4) for the HLM). To illustrate this difference further, consider the following example. Suppose we want to understand how job autonomy relates to job satisfaction. We know they are likely to be positively related and so we may model the data using simple regression. However, what if we suspect supervisory style

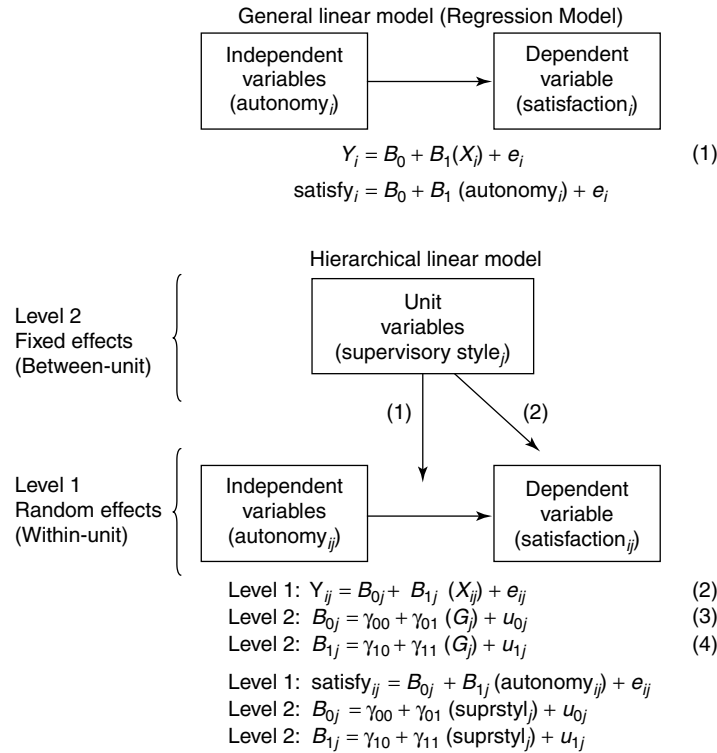


Figure 1 Comparison of GLM to HLM

2 Hierarchical Models

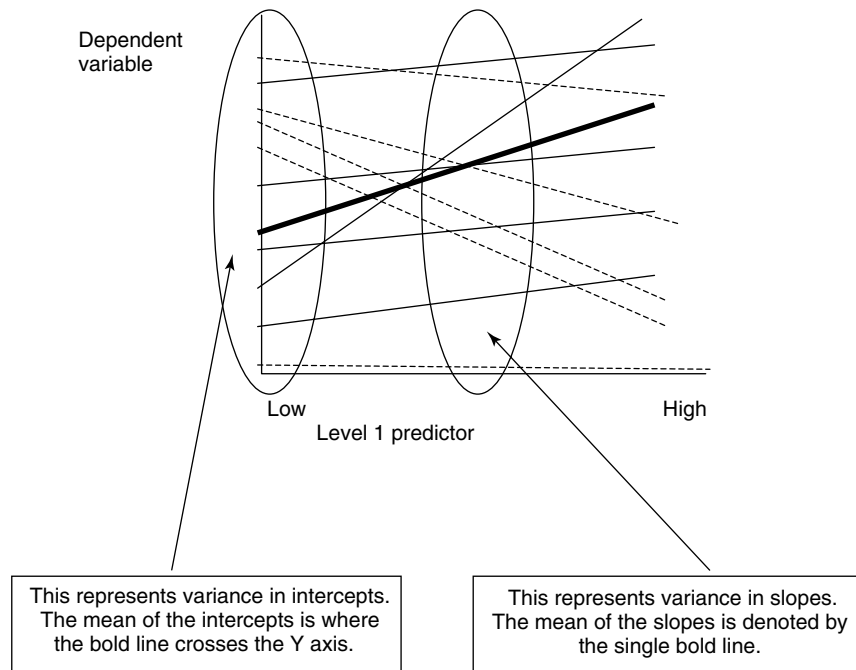


Figure 2 Illustration of what random effects for intercepts and slopes really mean

(e.g., autocratic, democratic) also influences satisfaction? Note employees are nested within supervisors – a given supervisor may be in charge of numerous employees, and, therefore, employees who have a particular supervisor may share some similarities not shared by employees of another supervisor. This makes it necessary to use HLM. Supervisors may have independent and direct effects on satisfaction (arrow 2), or moderate the relationship between autonomy and satisfaction (arrow 1).

Equation (1) is the classic regression model, which assumes errors are independent and normally distributed with a mean of zero and constant variance (see **Multiple Linear Regression**). The model assumes that the regression weights are constant across different supervisors; hence, there are no subscripts for B_0 and B_1 . This exemplifies a *fixed effects model* because the weights do not vary across units (see **Fixed and Random Effects**). In contrast, the HLM exemplifies a *random effects model* because the regression weights B_0 and B_1 vary across supervisors (levels of j ; see (2)) who are assumed to be randomly selected from a population of supervisors (see **Fixed and Random Effects**). Figure 2 illustrates this visually, where hypothetical regression lines are

shown for 10 subjects. The solid lines represent five subjects from one unit, the dashed lines represent five subjects from a different unit. Notice that across both units there is considerable variability in intercepts and slopes, with members of the second unit tending to show negative slopes (denoted by dashed lines). The solid bold line represents the regression line obtained from a traditional regression – clearly an inappropriate representation of this data.

The real benefit of HLM comes in two forms. First, because it explicitly models the nonindependence/heterogeneity in the data, it provides accurate standard errors and, hence, statistical significance tests. Second, it allows one to explain between-unit differences in the regression weights. This can be seen in (3) and (4). Equations (3) and (4) states between-unit differences in the intercept (slope) are explained by supervisory style. Note that in this model, the supervisory effect is a fixed effect.

Level 2 predictors may either be categorical (e.g., experimental condition) or continuous (e.g., individual differences). It is important to center the continuous data to facilitate interpretation of the lower level effects (see **Centering in Linear Multilevel Models**). Often, the most useful centering

method is to center the Level 2 predictors across all units (grand mean centering), and then center the Level 1 predictors within each unit (unit mean centering).

The basic HLM assumptions are (a) errors at both levels have a mean of zero, (b) Level 1 and Level 2 errors are uncorrelated with each other, (c) Level 1 errors (e_{ij}) have a constant variance (sigma-squared), and (d) Level 2 errors take a known form, but this form allows heterogeneity (nonconstant variance), and covariances among the error terms. Note also that HLM models frequently use restricted maximum likelihood (REML) estimation that assumes multivariate normality (*see Maximum Likelihood Estimation; Catalogue of Probability Density Functions*).

Comparing and Interpreting HLM Models

The HLM provides several estimates of model fit. These frequently include the -2 Residual Log Likelihood, **Akaike's Information Criterion** (AIC), and

the Bayesian Information Criterion (BIC). Smaller values for AIC and BIC are better. There are no statistical significance tests associated with these indices, so one must conduct a model comparison approach in which simple models are compared to more complicated models. The difference between the simple and complex models is evaluated via the change in -2 Residual Log Likelihood (distributed as a chi-square), and/or examining which has the smaller AIC and BIC values. Table 1 shows a generic sequence of model comparisons for HLM models. The model comparison approach permits only the minimum amount of model complexity to explain the data.

HLM also provides statistical significance tests of fixed and random effects. Statistical significance tests for the fixed effects are interpreted just like in the usual ordinary least squares (OLS) regression model. However, the statistical significance tests for the random effects should be avoided because they are often erroneous (see [13]). It is better

Table 1 Generic model comparison sequence for HLM

Step	Proc mixed command language
1. Determine the amount of nonindependence (nesting) in the dependent variable via the Intraclass Correlation Coefficient (ICC).	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL dv = /SOLUTION CL DDFM = BW; RANDOM INTERCEPT / TYPE = UN SUB = unit; RUN;
2. Add the Level 1 fixed effects. Interpret the statistical significance values for these effects.	PROC MIXED COVTEST UPDATE; CLASS UNIT; MODEL dv = iv /SOLUTION CL DDFM = BW; RUN;
3. Allow the intercept to be a random effect. Compare the difference between this model to the previous model via the change in loglikelihoods, AIC, and BIC.	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL dv = iv /SOLUTION CL DDFM = BW; RANDOM INTERCEPT / TYPE = UN SUB = unit; RUN;
4. One at a time, allow the relevant Level 1 predictor variables to become random effects. Compare the differences between models via the change in loglikelihoods, AIC, and BIC.	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL dv = iv /SOLUTION CL DDFM = BW; RANDOM INTERCEPT iv / TYPE = UN SUB = unit; RUN;
5. Attempt to explain the random effects via Level 2 predictors. Interpret the statistical significance values for these effects.	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL dv = iv lev2iv /SOLUTION CL DDFM = BW; RANDOM INTERCEPT iv / TYPE = UN SUB = unit; RUN;

Note: Capitalized SAS code refers to SAS commands and options; noncapitalized words are variables. *iv* refers to the Level 1 predictor and *lev2iv* refers to the Level 2 predictor.

to test the random effects using the model comparison approach described above and in Table 1 (see [3, 11]).

Example

Let us now illustrate how to model hierarchical data in SAS. Suppose we have 1567 employees nested within 168 supervisors. We hypothesize a simple two-level model identical to that shown in the lower part of Figure 1.

Following the model testing sequence in Table 1, we start with determining how much variance in satisfaction is explainable by differences between supervisors. This is known as an intraclass correlation coefficient (ICC) and is calculated by taking the variance in the intercept and dividing it by the sum of the intercept variance plus residual variance. Generic SAS notation for running all models is shown in Table 1. The COVTEST option requests significance tests of the random effects (although we know not to put too much faith in these tests), and the UPDATE option asks the program to keep us informed of the REML iteration progress. The CLASS statement identifies the Level 2 variable within which the Level 1 variables are nested (here referred to as ‘unit’). The statement `dv = /SOLUTION CL DDFM = BW` specifies the nature of the fixed effects (‘satisfy’ is the dependent variable; SOLUTION asks for significance tests for the fixed effects, CL requests confidence intervals, and DDFM = BW requests denominator degrees of freedom be calculated using the between-within method, see [9, 13]). The RANDOM statement is where we specify the random effects. Because the INTERCEPT is specified, we allow random variation among the intercept term. TYPE = UN specifies the structure of the random effects, where UN means unstructured (other options include variance components, etc.). SUB = unit again identifies the nesting variable. Because no predictor variable is specified in this model, it is equivalent to a one-way random effects Analysis of Variance (ANOVA) with ‘unit’ as the grouping factor. When we run this analysis, we find a variance component of 0.71 and residual variance of 2.78; therefore the $ICC = 0.71 / (0.71 + 2.78) = 0.20$. This means 20% of the variance in satisfaction can be explained by higher-level effects.

Step 2 includes autonomy and the intercept as fixed effects. This model is identical to the usual

regression model. Table 2 shows the command syntax and results for this model. The regression weights for the intercept (6.28) and autonomy (0.41) are statistically significant.

Step 3 determines whether there are between-unit differences in the intercept, which would represent a random effect. We start with the intercept and compare the fit indices for this model to those from the fixed-effects model. To conserve space, I note only that these comparisons supported the inclusion of the random effect for the intercept. The fourth step is to examine the regression weight for autonomy (also a random effect), and compare the fit of this model to the previous random intercept model. When we compare the two models, we find no improvement in model fit by allowing the slope parameter to be a random effect. This suggests the relationship between autonomy and satisfaction does not differ across supervisors. However, the intercept parameter does show significant variability (0.73) across units.

The last step is to determine whether supervisory style differences explain the variability in job satisfaction. To answer this question, we could include a measure of supervisory style as a Level 2 fixed effects predictor. Table 2 shows supervisory style has a significant effect (0.39).

Thus, one concludes (a) autonomy is positively related to satisfaction, and (b) this relationship is not moderated by supervisory style, but (c) there are between-supervisor differences in job satisfaction, and (d) supervisory style helps explain these differences.

Conclusion

Many substantive questions in the behavioral sciences must deal with nested and hierarchical data. Hierarchical models were developed to address these problems. This entry provided a brief introduction to such models and illustrated their application using SAS. But there are many extensions to this basic model. HLM can also be used to model longitudinal data and growth curves. In such models, the Level 1 model represents intraindividual change and the Level 2 model represents individual differences in intraindividual change (for introductions, see [3, 12]). HLM has many research and real-world applications and provides researchers with a powerful theory testing and building methodology.

Table 2 Sample table of HLM results

Model	df	Fixed parameter	Fixed 95% C.I.	Random parameter	AIC	SBC	-2LLR	SAS code
<u>Level 1 Model</u>								
Intercept (fixed)	1565	6.28*	(6.19; 6.37)	—	6309.7	6315.1	6307.7	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL satisfy = autonomy/SOLUTION CL
Autonomy	1565	0.41*	(0.31; 0.50)	—				DDFM = BW;
Residual	—	—	—	3.27				RUN;
<u>Level 1 and 2 Model</u>								
Intercept (random)	164	6.28*	(6.12; 6.44)	0.73	6161.8	6168.1	6157.8	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL satisfy = autonomy/SOLUTION CL
Autonomy	1401	0.41*	(0.32; 0.49)	—				DDFM = BW;
Residual	—	—	—	2.62				RANDOM INTERCEPT /TYPE = UN SUB = unit; RUN;
<u>Level 1 and 2 Model</u>								
Intercept	163	6.30*	6.16; 6.44)	0.52	6134.1	6140.3	6130.1	PROC MIXED COVTEST UPDATE; CLASS unit; MODEL satisfy = autonomy suprstyl/SOLUTION
Autonomy	1401	0.41*	(0.32; 0.49)	—				CL DDFM = BW;
Supervisory Style	163	0.39*	(0.26; 0.52)					RANDOM INTERCEPT /TYPE = UN SUB = unit;
Residual	—	—	—	2.63				RUN;

References

- [1] Bliese, P.D. (2000). Within-group agreement, non-independence, and reliability: implications for data aggregation and analysis, in *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, K. Klein & S.W.J. Kozlowski, eds, Jossey-Bass, San Francisco, pp. 349–381.
- [2] Bliese, P.D. (2002). Multilevel random coefficient modeling in organizational research: examples using SAS and S-PLUS, in *Modeling in Organizational Research: Measuring and Analyzing Behavior in Organizations*, F. Drasgow & N. Schmitt, eds, Jossey-Bass, Inc, San Francisco, pp. 401–445.
- [3] Bliese, P.D. & Ployhart, R.E. (2002). Growth modeling using random coefficient models: model building, testing, and illustration, *Organizational Research Methods* **5**, 362–387.
- [4] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage, Newbury Park.
- [5] Heck, R.H. & Thomas, S.L. (2000). *An Introduction to Multilevel Modeling Techniques*, Erlbaum, Hillsdale.
- [6] Hofmann, D.A., Griffin, M.A. & Gavin, M.B. (2000). The application of hierarchical linear modeling to organizational research, in *Multilevel Theory, Research, and Methods in Organizations*, K.J. Klein & S.W. Kozlowski, eds, Jossey-Bass, Inc, San Francisco, pp. 467–511.
- [7] Kenny, D.A. & Judd, C.M. (1986). Consequences of violating the independence assumption in analysis of variance, *Psychological Bulletin* **99**, 422–431.
- [8] Kreft, I. & De Leeuw, J. (1998). *Introducing Multilevel Modeling*, Sage Publications, London.
- [9] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*, SAS Institute, Cary.
- [10] Little, T.D., Schnabel, K.U. & Baumert, J. (2000). *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples*, Erlbaum, Hillsdale.
- [11] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects Models in S and S-PLUS*, Springer-Verlag, New York.
- [12] Ployhart, R.E., Holtz, B.C. & Bliese, P.D. (2002). Longitudinal data analysis: applications of random coefficient modeling to leadership research, *Leadership Quarterly* **13**, 455–486.
- [13] Singer, J.D. (1998). Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models, *Journal of Educational and Behavioral Statistics* **24**, 323–355.
- [14] Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage, Thousand Oaks, CA.
- [15] Von Bertalanffy, L. (1968). *General Systems Theory*, Braziller, New York.

ROBERT E. PLOYHART

High-dimensional Regression

JAN DE LEEUW

Volume 2, pp. 816–818

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

High-dimensional Regression

In regression analysis, there are n observations y_i on a dependent variable (also known as outcome or criterion) that are related to n corresponding observations x_i on p independent variables (also known as inputs or predictors). Fitting regression models of some form or another is by far the most common uses of statistics in the sciences (*see Multiple Linear Regression*).

Statistical theory tells us to assume that the observed outcomes y_i are realizations of n random variables $\underline{y}_i$. We model the conditional expectation of $\underline{y}_i$ given x_i , or, to put it differently, we model the expected value of $\underline{y}_i$ as a function of x_i

$$\mathbf{E}(\underline{y}_i | x_i) = F(x_i), \quad (1)$$

where the function F must be estimated from the data. Often the function F is known except for a small number of parameters. This defines parametric regression. Sometimes F is unknown, except for the fact that we know that has a certain degree of continuity or smoothness. This defines **nonparametric regression**.

In this entry, we are specifically concerned with the situation in which the number of predictors is large. Through the years, the meaning of ‘large’ has changed. In the early 1900s, three was a large number, in the 1980s 100 was large, and at the moment we sometimes have to deal with situations in which there are 10 000 predictors. This means, in the regression context, that we have to estimate a function F of 10 000 variables. Modern data collection techniques in, for example, genetics, environmental monitoring, and consumer research lead to these huge datasets, and it is becoming clear that classical statistical techniques are useless for such data. Entirely different methods, sometimes discussed under the labels of ‘data mining’ or ‘machine learning’, are needed [5] (*see Data Mining*).

Until recently **multiple linear regression**, in which F is linear, was the only practical alternative to deal with a large number of predictors. Thus, we specialize our model to

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{s=1}^p \beta_s x_{is}. \quad (2)$$

It became clear rather soon that linear regression with a large number of predictors has many problems. The main ones are multicollinearity, often even singularity, and the resulting numerical instability of the estimated regression coefficients (*see Collinearity*).

An early attempt to improve this situation is using variable selection. We fit the model

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{s=1}^p \beta_s \delta_s x_{is}. \quad (3)$$

where δ_s is either zero or one. In fitting this model, we select a subset of the variables and then do a linear regression. Although variable selection methods appeared relatively early in the standard statistical packages, and became quite popular, they have the major handicap that they must solve the combinatorial problem of finding the optimum selection from among the 2^p possible ones. Since this rapidly becomes unsolvable in any reasonable amount of time, various heuristics have been devised. Because of the instability of high-dimensional linear regression problems, the various heuristics often lead to very different solutions. Two ways out of the dilemma, which both stay quite close to linear regression, have been proposed around 1980. The first is principal component regression (*see Principal Component Analysis*) or PCR, in which we have

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{t=1}^q \beta_t \left[\sum_{s=1}^p \alpha_{ts} x_{is} \right]. \quad (4)$$

Here we replace the p predictors by $q < p$ principal components and then perform the linear regression. This tackles the multicollinearity problem directly, but it inherits some of the problems of principal component analysis. How many components do we keep? And how do we scale our variables for the component analysis?

The second, somewhat more radical, solution is to use the **generalized additive model** or GAM discussed by [6]. This means

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{s=1}^p \beta_s \phi_s(x_{is}), \quad (5)$$

where we optimize the regression fit over both θ and the functions (transformations) ϕ . Usually we require $\phi \in \Phi$ where Φ is some finite dimensional subspace of functions, such as polynomials or splines with a

2 High-dimensional Regression

given knot sequence. Best fits for such models are easily computed these days by using alternating least squares algorithms that iteratively alternate fitting θ for fixed ϕ and fitting ϕ for fixed θ [1]. Although generalized additive models add a great deal of flexibility to the regression situation, they do not directly deal with the instability and multicollinearity that comes from the very large number of predictors. They do not address the data reduction problem, they just add more parameters to obtain a better fit.

A next step is to combine the ideas of PCR and GAM into projection pursuit regression or PPR [4]. The model now is

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{t=1}^q \phi_t \left[\sum_{s=1}^p \alpha_{ts} x_{is} \right]. \quad (6)$$

This is very much like GAM, but the transformations are applied to a presumably small number of linear combinations of the original variables. PPR regression models are closely related to neural networks, in which the linear combinations are the single hidden layer and the nonlinear transformations are sigmoids or other squashers (see **Neural Networks**). PPR models can be fit by general neural network algorithms.

PPR regression is generalized in Li's **slicing inverse regression** or SIR [7, 8], in which the model is

$$\mathbf{E}(\underline{y}_i | x_i) = F \left[\sum_{s=1}^p \alpha_{1s} x_{is}, \dots, \sum_{s=1}^p \alpha_{qs} x_{is} \right]. \quad (7)$$

For details on the SIR and PHD algorithms, we refer to (see **Slicing Inverse Regression**).

Another common, and very general approach, is to use a finite basis of functions h_{st} , with $t = 1, \dots, q_s$, for each of the predictors x_s . The basis functions can be polynomials, piecewise polynomials, or splines, or radical basis functions. We then approximate the multivariate function F by a sum of products of these basis functions. Thus we obtain the model

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{t_1=1}^{q_1} \dots \sum_{t_p=1}^{q_p} \theta_{t_1 \dots t_p} \times h_{1t_1}(x_{i1}) \times \dots \times h_{pt_p}(x_{ip}) \quad (8)$$

This approach is used in multivariate adaptive regression splines, or MARS, by [3]. The basis functions are splines, and they adapt to the data by locating the knots of the splines.

A different strategy is to use the fact that any multivariate function can be approximated by a multivariate step function. This fits into the product model, if we realize that multivariate functions constant on rectangles are products of univariate functions constant on intervals. In general, we fit

$$\mathbf{E}(\underline{y}_i | x_i) = \sum_{t=1}^q \theta_t I(x_i \in R_t). \quad (9)$$

Here, the R_t define a partitioning of the p -dimensional space of predictors, and the $I()$ are indicator functions of the q regions. In each of the regions the regression function is a constant. The problem, of course, is how to define the regions. The most popular solution is to use a recursive partitioning algorithm such as **Classification and Regression Trees**, or by the algorithm CART [2], which defines the regions as rectangles in variable space. Partitionings are refined by splitting along a variable, and by finding the variable and the split which minimize the residual sum of squares. If the variable is categorical, we split into two arbitrary subsets of categories. If the variable is quantitative, we split an interval into two pieces. This recursive partitioning builds up a binary tree, in which leaves are refined in each stage by splitting the rectangles into two parts.

It is difficult, at the moment, to suggest a best technique for high-dimensional regression. Formal statistical **sensitivity analysis**, in the form of standard errors and confidence intervals, is largely missing. Decision procedures, in the form of tests, are also in their infancy. The emphasis is on exploration and on computation. Since the data sets are often enormous, we do not really have to worry too much about significance, we just have to worry about predictive performance and about finding (mining) interesting aspects of the data.

References

- [1] Breiman, L. & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association* **80**, 580–619.
- [2] Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*, Wadsworth.
- [3] Friedman, J. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**, 1–141.

- [4] Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.
- [5] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.
- [6] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman and Hall, London.
- [7] Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion), *Journal of the American Statistical Association* **86**, 316–342.
- [8] Li, K.C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s Lemma, *Journal of the American Statistical Association* **87**, 1025–1039.

JAN DE LEEUW

Hill's Criteria of Causation

KAREN J. GOODMAN AND CARL V. PHILLIPS

Volume 2, pp. 818–820

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hill's Criteria of Causation

The term *criteria of causation* (often called *causal criteria*), applied to Sir Austin Bradford Hill's widely cited list [8] of factors to consider before inferring causation from an observed association, is something of a misnomer. Hill (1897–1991) [4, 13] never called these considerations *criteria* but rather referred to them as *viewpoints*, and he did not believe it useful to elaborate 'hard-and-fast rules of evidence' [8, p. 299] for causation. Nevertheless, the publication of Hill's landmark address to the Royal Society of Medicine [8] is frequently cited as the authoritative source for causal criteria in epidemiologic practice [23, 24].

Hill singled out nine factors to consider 'before we cry causation' [8, p. 299] when observing a statistical association: strength; consistency; specificity; temporality; biological gradient; plausibility; coherence; experiment; analogy (in the order originally presented). Hill was neither the first nor the last to propose such a list [21, 23]. It is curious that causal criteria are so closely associated with Hill's name, particularly given that many authors who apply what they call 'Bradford Hill criteria' select idiosyncratically from his list those items they prefer, producing subsets that often more closely resemble shorter lists proposed by others [23, 24]. The attribution to Hill perhaps reflects his professional stature due to his contributions to medical statistics and epidemiologic study design in the early twentieth century [1–5, 19, 22], although his eloquence has also been proposed as an explanation [7].

Implicit in Hill's articulation of how to decide whether an association is causal is the recognition, dating back to Hume [14], of a fundamental limitation of empirical sciences: cause–effect relations cannot be observed directly or proven true by logic and therefore must be inferred by inductive reasoning [17, 21]. Hill recognized that statistical association does not equate with causation and that all scientific findings are tentative and subject to being upset by advancing knowledge. At the same time, he warned passionately that this limitation of science 'does not confer upon us a freedom to ignore the knowledge that we already have, or to postpone the action it appears to demand at a given

time' [8, p. 300]. These words reveal Hill's conviction that scientific judgments about causation are required so that the knowledge can be used to improve human lives. It is clearly for this purpose that Hill offered his perspective on how to decide whether an association observed in data is one of cause and effect.

Hill's presentation on inferring causation, though replete with insight, did not constitute a complete thesis on causal inference. Hill did not explicitly define what he meant by each of the viewpoints, relying instead on illustrative examples. While qualifying his approach by stating 'none of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis' [8, p. 299], he offered no means of deciding whether these aspects hold when considering a given association, no hierarchy of importance among them, and no method for assessing them to arrive at an inference of causation. Hill included the list of nine viewpoints in four editions of his textbook of medical statistics from 1971 through 1991 [9–12], without further elaboration than appeared in the original paper.

History of Causal Criteria in Epidemiology

Discussions of causal criteria arose from the recognized limitations of the Henle-Koch postulates, formalized in the late nineteenth century to establish causation for infectious agents [6]. These postulates require the causal factor to be necessary and specific, apply only to a subset of infectious agents, and conflict with multifactorial causal explanations. The nonspecificity of causal agents was a particular concern [18]. Two lists of causal criteria published before 1960 did not include specificity (instead including consistency or replication, strength, dose-response, experimentation, temporality, and biological reasonableness) [23]. In 1964, the year before Hill's presentation, the Surgeon General's Committee on Smoking and Health elected to use explicit criteria to determine whether smoking caused the diseases under review; their list included consistency, strength, specificity, temporality, and biological coherence (under which they included biological mechanism, dose-response, and exclusion of alternate explanations such as bias) [23]. According to Hamill, one of the committee members, the committee did not consider the list 'hewn in stone or

intended for all time and all occasions, but as a formal description of how *we* drew our... conclusions... [7, p. 527]. Since the 1970s [20], Susser has advocated the use of causal criteria for discriminating between a true causal factor and 'an imposter' [21, pp. 637–8], proposing a refined list of criteria in 1991, including strength, specificity, consistency (both replicability and survivability on diverse tests of the causal hypothesis), predictive performance, and coherence (including theoretical, factual, biological, and statistical) [21]. Susser's historical analysis argues against ossified causal criteria ('epidemiologists have modified their causal concepts as the nature of their tasks has changed... Indeed, the current set of criteria may well be displaced as the tasks of the discipline change, which they are bound to do.' [21, pp. 646–7]).

Limitations of Criteria for Inferring Causation

With the exception of temporality, no item on any proposed list is necessary for causation, and none is sufficient. More importantly, it is not clear how to quantify the degree to which each criterion is met, let alone how to aggregate such results into a judgment about causation. In their advanced epidemiology textbook, Rothman and Greenland question the utility of each item on Hill's list except temporality [17]. Studies of how epidemiologists apply causal criteria reveal wide variations in how the criteria are selected, defined, and judged [24]. Furthermore, there appear to be no empirical assessments to date of the validity or usefulness of causal criteria (e.g., retrospective studies of whether appealing to criteria improves the conclusions of an analysis). In short, the value of checklists of criteria for causal inference is severely limited and has not been tested.

Beyond Causal Criteria

Although modern thinking reveals limitations of causal criteria, Hill's landmark paper contains crucial insights. Hill anticipated modern statistical approaches to critically analyzing associations, asking 'Is there any way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?' [8, p. 299]. Ironically, although Hill correctly identified this as the fundamental question, consulting a set of criteria does

little to answer this question. Recent developments in methods for uncertainty quantification [15], however, are creating tools for assessing the probability that an observed association is due to alternative explanations, which include random error or study bias rather than a causal relationship. Equally important, Hill, though described as the greatest medical statistician of the twentieth century [4], had his formal academic training in economics rather than medicine or statistics, and anticipated modern expected-net-benefit-based decision analysis [16]. He stated, 'finally, in passing from association to causation... we shall have to consider what flows from that decision' [8, p. 300], and suggested that the degree of evidence required, in so far as alternate explanations appear unlikely, depends on the potential costs and benefits of taking action. Recognizing the inevitable scientific uncertainty in establishing cause and effect, for Hill, the bottom line for causal inference – overlooked in most discussions of causation or statistical inference – was deciding whether the evidence was convincing enough to warrant a particular policy action when considering expected costs and benefits.

References

- [1] Armitage, P. (2003). Fisher, Bradford Hill, and randomization, *International Journal of Epidemiology* **32**, 925–928.
- [2] Chalmers, I. (2003). Fisher and Bradford Hill: theory and pragmatism? *International Journal of Epidemiology* **32**, 922–924.
- [3] Doll, R. (1992). Sir Austin Bradford Hill and the progress of medical science, *British Medical Journal* **305**, 1521–1526.
- [4] Doll, R. (1993). Obituary: Sir Austin Bradford Hill, pp. 795–797, in Sir Austin Bradford Hill, 1899–1991, *Statistics in Medicine* **12**, 795–808.
- [5] Doll, R. (2003). Fisher and Bradford Hill: their personal impact, *International Journal of Epidemiology* **32**, 929–931.
- [6] Evans, A.S. (1978). Causation and disease: a chronological journey, *American Journal of Epidemiology* **108**, 249–258.
- [7] Hamill, P.V.V. (1997). Re: Invited commentary: Response to *Science* article, 'Epidemiology faces its limits', *American Journal of Epidemiology* **146**, 527.
- [8] Hill, A.B. (1965). The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* **58**, 295–300.
- [9] Hill, A.B. (1971). *Principles of Medical Statistics*, 9th Edition, Oxford University Press, New York.

-
- [10] Hill, A.B. (1977). *Short Textbook of Medical Statistics*, Oxford University Press, New York.
 - [11] Hill, A.B. (1984). *Short Textbook of Medical Statistics*, Oxford University Press, New York.
 - [12] Hill, A.B. (1991). *Bradford Hill's Principles of Medical Statistics*, Oxford University Press, New York.
 - [13] Hill, I.D. (1982). Austin Bradford Hill – ancestry and early life, *Statistics in Medicine* **1**, 297–300.
 - [14] Hume, D. (1978). *A Treatise of Human Nature*, (Originally published in 1739), 2nd Edition, Oxford University Press, 1888, Oxford.
 - [15] Phillips, C.V. (2003). Quantifying and reporting uncertainty from systematic errors, *Epidemiology* **14**(4), 459–466.
 - [16] Phillips, C.V. & Goodman K.J. (2004). The missed lessons of Sir Austin Bradford Hill, *Epidemiologic Perspectives and Innovations*, **1**, 3.
 - [17] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, Chap. 2, 2nd Edition, Lippencott-Raven Publishers, Philadelphia, pp. 7–28.
 - [18] Sartwell, P.E. (1960). On the methodology of investigations of etiologic factors in chronic disease – further comments, *Journal of Chronic Diseases* **11**, 61–63.
 - [19] Silverman, W.A. & Chalmers, I. (1992). Sir Austin Bradford Hill: an appreciation, *Controlled Clinical Trials* **13**, 100–105.
 - [20] Susser, M. (1973). *Causal Thinking in the Health Sciences. Concepts and Strategies of Epidemiology*, Oxford University Press, New York.
 - [21] Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology, *American Journal of Epidemiology* **133**, 635–648.
 - [22] *Statistics in Medicine*, Special Issue to Mark the 85th Birthday of Sir Austin Bradford Hill **1**(4), 297–375 1982.
 - [23] Weed, D.L. (1995). Causal and preventive inference, in *Cancer Prevention and Control*, P. Greenwald, B.S. Kramer & D. Weed, eds, Marcel Dekker, pp. 285–302.
 - [24] Weed, D.L. & Gorelic, L.S. (1996). The practice of causal inference in cancer epidemiology, *Cancer Epidemiology, Biomarkers & Prevention* **5**, 303–311.
- (See also **INUS Conditions**)

KAREN J. GOODMAN AND CARL V. PHILLIPS

Histogram

BRIAN S. EVERITT

Volume 2, pp. 820–821

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Histogram

A histogram is perhaps the graphical display that is used most often in the initial exploration of a set of measurements. Essentially, it is a simple graphical representation of a frequency distribution in which each class interval (category) is represented by a vertical bar whose base is the class interval and whose height is proportional to the number of observations in the class interval. When the class intervals are unequally spaced, the histogram is drawn in such a way that the area of each bar is proportional to the frequency for that class interval. Scott [1] considers how to choose the optimal number of classes in a histogram. Figure 1 shows a histogram of the murder rates (per 100 000) for 30 cities in southern USA in 1970.

The histogram is generally used for two purposes, counting and displaying the distribution of a variable, although it is relatively ineffective for both; **stem and leaf plots** are preferred for counting and **box plots** are preferred for assessing distributional properties. A histogram is the continuous data counterpart of the **bar chart**.

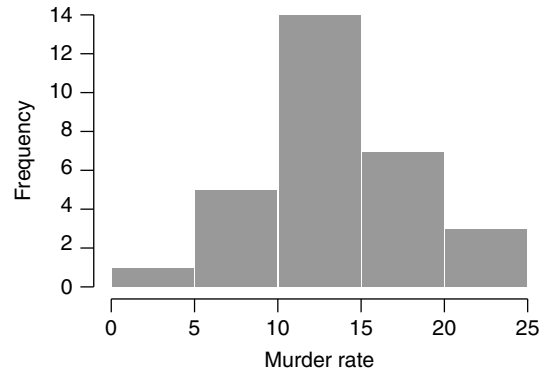


Figure 1 Murder rates for 30 cities in the United States

Reference

- [1] Scott, D.W. (1979). On optimal and data based histograms, *Biometrika* **66**, 605–610.

BRIAN S. EVERITT

Historical Controls

VANCE W. BERGER AND RANDI SHAFRAN

Volume 2, pp. 821–823

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Historical Controls

Randomized trials (*see* **Clinical Trials and Intervention Studies**), in which a single group of units is randomly allocated to two or more treatment conditions and then studied in parallel over time, is the gold standard methodology for reducing biases. Yet not every comparison is amenable to a randomized trial. For example, one cannot randomize subjects to their genders or races. This implies that a comparison of men to women cannot be randomized. Furthermore, even some comparisons that could theoretically be amenable to study with a randomized trial would not be, for ethical reasons. For example, it is possible to randomly assign subjects to different numbers of cigarettes to be smoked per day in an effort to quantify the effects smoking would have on heart disease or lung cancer. Yet for ethical reasons such a study would be highly unlikely to be approved by an oversight committee.

Furthermore, even studies that could be conducted as randomized trials without much ethical objection may be conducted instead as historically controlled trials for financial reasons. For example, Phase II clinical trials often use historical data from the literature to compare the general response rate of experimental treatment to that of a standard treatment. The reality, then, is that some studies are nonrandomized. Nonrandomized studies may still be parallel over time, in that all treatment groups are treated and studied over the same time interval. For example, one could compare two treatments for headaches, both available over-the-counter, based on self-selected samples, that is, one would compare the experiences of those subjects who select one treatment against those subjects who select the other one. Such a design is susceptible to selection bias because those patients selecting one treatment may differ systematically from those selecting another [2, 5]. But at least such a design would not also confound treatments with time. Comparing the experiences of a group of subjects treated one way now to the experiences of a group of subjects treated another way at some earlier point in time has many of the same biases that the aforementioned self-selection design has, but it also confounds treatments and time, and hence introduces additional biases.

For example, suppose that a new surgical technique is compared to an old one by comparing the experiences of a group of subjects treated now with

the new technique to the experiences of a different group of subjects treated 10 years ago with the older technique. If the current subjects live longer or respond better in some other way, then one would want to attribute this to improvements in the surgical technique, and conclude that progress has been made. However, it is also possible that any observed differences are due to improvements not in the surgical technique itself but rather to improvements in ancillary aspects of patient care and management. Note that our interest is in comparing the experimental response rate to its counterfactual, or the response rate of the same subjects to the standard therapy. In a randomized trial, we substitute the experiences of a randomly selected control group for the counterfactual, and in a historically controlled trial, we substitute the experiences of a previously treated and nonrandomly selected control group for the counterfactual.

For a historically controlled trial to provide valid inference, then, we would need the response rate of the previously treated cohort to be the same as the response rate of the present cohort had they been treated with the standard treatment. One condition under which this would be true would be if the control response rate may be treated as a constant, independent of the cohort to which it is applied. Otherwise, use of historical controls may lead to invalid conclusions about experimental treatments [8]. Suppose, for example, that a new treatment is no better than a standard one, as each is associated with a 10% response rate in the population. One may even suppose that the same 10% of the population would respond to each treatment, so these are truly preordained responses, having nothing to do with which treatment is administered. Suppose that historical databases reveal this 10% response rate for the standard treatment, and now the new treatment is to be studied in a new cohort, using only the historical control group (no parallel concurrent control group).

Consider a certain recognizable subset of the population, based on specific values of measured covariates having not a 10% response rate but rather a 50% response rate. Moreover, suppose that this subgroup comprises 20% of the population, and none of the other 80% will respond at all. The overall response rate, then, is 50% of 20%, or the original 10%. A study of a given cohort would be representative of the overall population only to the

2 Historical Controls

extent that it is comprised of roughly 20% potential responders (those having the 50% response rate) and 80% nonresponders, following the population proportions. But if these proportions are distorted in the cohort, then the response rate in the cohort will not reflect the response rate in the population. In the extreme case, if the cohort is comprised entirely of potential responders, then the response rate will be 50%.

The key, though, is that if this distortion from the population is not recognized, then one might be inclined to attribute the increased response rate, 50% versus 10%, not to the composition of the sample but rather to how they were treated. One would go on to associate the 50% response rate with the new treatment, and conclude that it is superior to the standard one. We see that the selection bias discussed above can render historically controlled data misleading, so that observed response rates are not equal to the true response rates. However, historically controlled data can be used to help evaluate new treatments if selection bias can be minimized. The conditions under which selection bias can be demonstrated to be minimal are generally not very plausible, and require a uniform prognosis of untreated patients [3, 7]. An extreme example can illustrate this point. If a vaccine were able to confer immortality, or even the ability to survive an event that currently is uniformly fatal, then it would be clear, even without randomization, that this vaccine is effective. But this is not likely, and so it is probably safe to say that no historically controlled trial can be known to be free of biases.

Of course, if the evidentiary standard required for progress in science were an ironclad guarantee of no biases, then science would probably not make very much progress, and so it may be unfair to single out historically controlled studies as unacceptable based on the biases they may introduce. If historically controlled trials tend to be more biased than concurrent, or especially randomized trials, then this has to be a disadvantage that counts against historically controlled trials. However, it is not the only consideration. Despite the limitations of historical control data, there are, under certain conditions, advantages to employing this technique. For example, if the new treatment turns out to be truly superior to the control treatment, then finding this out with a historically controlled trial would not involve exposing any patients to the less effective control treatment [4].

This is especially important for patients with life-threatening diseases. Besides potential ethical advantages, studies with historical controls may require a smaller number of participants and may require less time than comparable randomized trials [4, 6].

If feasible, then randomized control trials should be used. However, this is not always the case, and historical control trials may be used as an alternative. The limitations of historical controls must be taken into account in order to prevent false conclusions regarding the evaluation of new treatments. It is probably not prudent to use formal inferential analyses with any nonrandomized studies, including historically controlled studies, because without a sample space of other potential outcomes and known probabilities for each, the only outcome that can be considered to have been possible (with a known probability, for inclusion in a sample space) is the one that was observed. This means that the only valid P value is the uninteresting value of 1.00 [1].

References

- [1] Berger, V.W. & Bears, J. (2003). When can a clinical trial be called 'randomized'? *Vaccine* **21**, 468–472.
- [2] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias, *Journal of Modern Applied Statistical Methods* **2**, 80–86.
- [3] Byar, D.P., Schoenfeld, D.A., Green, S.B., Amato, D.A., Davis, R., De Gruttola, V., Finkelstein, D.M., Gatsonis, C., Gelber, R.D., Lagakos, S., Lefkopoulou, M., Tsiatis, A.A., Zelen, M., Peto, J., Freedman, L.S., Gail, M., Simon, R., Ellenberg, S.S., Anderson, J.R., Collins, R., Peto, R., Peto, T. (1990). Design considerations for AIDS trials, *New England Journal of Medicine* **323**, 1343–1348.
- [4] Gehan, E.A. (1984). The evaluation of therapies: historical control studies, *Statistics in Medicine* **3**, 315–324.
- [5] Green, S.B. & Byar, D.B. (1984). Using observational data from registries to compare treatments: the fallacy of omnimetrics, *Statistics in Medicine* **3**, 361–370.
- [6] Hoehler, F.K. (1999). Sample size calculations when outcomes will be compared with historical control, *Computers in Biology and Medicine* **29**, 101–110.
- [7] Lewis, J.A. & Facey, K.M. (1998). Statistical shortcomings in licensing applications, *Statistics in Medicine* **17**, 1663–1673.
- [8] Thall, P. & Simon, R. (1990). Incorporating historical control data in planning phase II clinical trials, *Statistics in Medicine* **9**, 215–228.

VANCE W. BERGER AND RANDI SHAFRAN

History of Analysis of Variance

RYAN D. TWENEY

Volume 2, pp. 823–826

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Analysis of Variance

Following its development by **Sir R.A. Fisher** in the 1920s, the **Analysis of Variance (ANOVA)** first reached a wide behavioral audience via Fisher's 1935 book, *The Design of Experiments* [4], a work that went through many editions and is frequently cited as the origin of behavioral science's knowledge of ANOVA. In fact, ANOVA was first used by an educational researcher in 1934 [10], based on earlier publications by Fisher. The incorporation of ANOVA was initially a gradual one, in which educational researchers and parapsychologists played the earliest roles, followed slightly later by experimental psychologists. Published applications of ANOVA remained few until after World War II, when its use accelerated rapidly. By the 1960s, it had become a routine procedure among experimental psychologists and had been incorporated as a required course in nearly all doctoral programs in psychology.

The history of ANOVA in the behavioral sciences is also the history of Fisherian ideas and of small sample statistics generally. Thus, although Student's t had been developed much earlier than ANOVA, its incorporation by behavioral scientists did not begin until the 1930s and its use grew in parallel to that of ANOVA [11]. This parallelism has been attributed to the incorporation of null hypothesis testing into a broader methodology of experimental design for which ANOVA techniques were especially suited.

The initial uses of statistics in the psychological sciences were primarily correlational and centered on what Danziger [3] characterized as the 'Galtonian' (see **Galton, Francis**) tradition. Statistical analysis was prominent, but the emphasis was upon the measurement of interindividual differences. There were few connections with experimental design methods during this period. The focus was upon variation, a key concept for the post-Darwinian functional approach of the Galtonians. Yet within psychology (as within eugenics) interest in variation was soon displaced by interest in the mean value, or central tendency, a change attributed to increasing bureaucratic pressure to characterize aggregates of people (as in school systems). During the first third of the twentieth century, the use of descriptive and correlational

statistical techniques thus spread rapidly, especially in educational psychology.

The dominant psychological use of statistics during the first decades of the twentieth century reflected a statistics 'without P values'. Such statistical applications were closely related to those prominent among sociological and economic users of statistics. Significance testing, as such, was not used in psychology prior to the adoption of ANOVA, which happened only after the publication of Fisher's 1935 book. Nonetheless, there were precursors to significance testing, the most common being use of the 'critical ratio' for comparing two means, defined as the difference between the means divided by the 'standard deviation.' The latter sometimes represented a *pooled* value across the two groups, although with N instead of $N - 1$ in the denominator. An excellent early account of the critical ratio, one that anticipated later discussions by Fisher of statistical tests, was given in 1909 by Charles S. Myers in his *Text-Book of Experimental Psychology* [8]. Note also that the critical ratio thus defined is nearly identical to the currently widely used measure of **effect size** *Cohen's d* .

The use of inferential statistics such as ANOVA did *not* replace the use of correlational statistics in psychology. Instead, correlational techniques, which were in use by psychologists almost from the establishment of the discipline in America, showed no change in their relative use in American psychological journals. Between 1935 and 1952, the use of correlational techniques remained steady at around 30% of all journal articles; during the same period, the use of ANOVA increased from 0 to nearly 20% and the use of the t Test increased from 0 to 32% [11]. While the relative use of correlational techniques did not change as a result of the incorporation of ANOVA, there was a decline in the use of the critical ratio as a test, which vanished completely by 1960. There was also a decline in the prominence given to 'brass and glass' instruments in psychological publications after World War II, in spite of a continuing ideological strength of experimentation. In fact, statistical analysis strengthened this ideology to a degree that instruments alone had not been able to do; statistics, including ANOVA, became the 'instruments of choice' among experimentalists.

It is also clear that the incorporation of ANOVA into psychology was not driven solely by the logical

need for a method of analyzing complex **experimental designs**. Lovie [7] noted that such designs were used long prior to the appearance of ANOVA techniques and that even **factorial** and nested designs were in occasional use in the 1920s and 1930s. He suggested that the late appearance of ANOVA was instead due in part to the cognitive complexity of its use and the relatively limited mathematical backgrounds of the experimental psychologists who were its most likely clients. Further, Lovie noted the deeply theoretical nature of the concept of **interaction**. Until the simplistic ‘rule of one variable’ that dominated experimental methodological thinking could be transcended, there was no proper understanding of the contribution that ANOVA could make.

In the United States, the demands of war research during World War II exposed many psychologists to new problems, new techniques, and a need to face the limitations of accepted psychological methods. In contrast to the first war, there was much less of the ‘measure everyone’ emphasis that characterized the Yerkes-led mental testing project of World War I. Instead, a variety of projects used the research abilities of psychologists, often in collaboration with other disciplinary scientists. War research also affected the nature of statistical analysis itself, and, in fact, also provided an opportunity for statisticians to establish their autonomy as a distinct profession. Many of the common uses of statistical inference were being extended by statisticians and mathematicians during the war, for example studies of bombing and fire control (**Neyman**), sequential analysis (Wald and others), and quality control statistics (**Deming**). More to the point, significance testing began to find its way into the specific applications that psychologists were working upon. Rucci & Tweney [11] found only 17 articles in psychology journals that used ANOVA between 1934 and 1939, and of these most of the applications were, as Lovie [7] noted, rather unimpressive. Yet the wartime experiences of psychologists drove home the utility of these procedures, led to many more psychologists learning the new procedures, and provided paradigmatic exemplars of their use.

After 1945, there was a rapid expansion of graduate training programs in psychology, driven in large part by a perceived societal need for more clinical and counseling services, and also by the needs of ‘Cold War’ military, corporate, and governmental bureaucracies. Experimental psychologists, newly apprised

of Fisherian statistical testing, and **measurement-oriented** psychologists who had had their psychometric and statistical skills sharpened by war research, were thus able to join hands in recommending that statistical training in *both* domains, ANOVA and correlational, be a requirement for doctoral-level education in psychology. As psychology expanded, experimental psychologists in the United States were therefore entrusted with ensuring the scientific credentials of the training of graduate students (most of whom had little background in mathematics or the physical sciences), even in clinical domains. The adoption of ANOVA training permitted a new generation of psychologists access to a set of tools of perceived scientific status and value, without demanding additional training in mathematics or physical science. As a result, by the 1970s, ANOVA was frequent in all experimental research, and the use of significance testing had penetrated all areas of the behavioral sciences, including those that relied upon correlational and **factor-analytic** techniques. In spite of frequent reminders that there were ‘two psychologies’ [2], one correlational and one ANOVA-based, the trend was toward the statistical merging of the two via the common embrace of significance testing.

In the last decades of the twentieth century, ANOVA techniques displayed a greater sophistication, including repeated measures designs (*see* **Repeated Measures Analysis of Variance**), mixed designs, **multivariate analysis of variance** and other procedures. Many of these were available long before their incorporation in psychology and other behavioral sciences. In addition, recent decades have seen a greater awareness of the formal identity between ANOVA and **multiple linear regression** techniques, both of which are, in effect, applications of a **generalized linear model** [9].

In spite of this growing sophistication, the use of ANOVA techniques has not always been seen as a good thing. In particular, the ease with which complex statistical procedures can be carried out on modern desktop computers has led to what many see as the misuse and overuse of otherwise powerful programs. For example, one prominent recent critic, Geoffrey Loftus [6], has urged the replacement of null hypothesis testing by the pictorial display of experimental effects, together with relevant **confidence intervals** even for very complex designs.

Many of the criticisms of ANOVA use in the behavioral sciences are based upon a claim that

inferential canons are being violated. Some have criticized the ‘mechanized’ inference practiced by many in the behavioral sciences, for whom a significant effect is a ‘true’ finding and a nonsignificant effect is a finding of ‘no difference’ [1]. Gigerenzer [5] argued that psychologists were using an incoherent ‘hybrid model’ of inference, one that inappropriately blended aspects of **Neyman/Pearson** approaches with those of Fisher. In effect, the charge is that a kind of misappropriated Bayesianism (see **Bayesian Statistics**) has been at work, one in which the P value of a significance test, $p(D|H_0)$, was confused with the *posterior* probability, $p(H_0|D)$, and, even more horribly, that $p(H_1|D)$ was equated with $1 - p(D|H_0)$. Empirical evidence that such confusions were rampant – even among professional behavioral scientists – was given by **Tversky & Kahneman** [12].

By the beginning of the twenty-first century, the ease and availability of sophisticated ANOVA techniques continued to grow, along with increasingly powerful graphical routines. These hold out the hope that better uses of ANOVA will appear among behavioral sciences. The concerns over mechanized inference and inappropriate inferential beliefs will not, however, be resolved by any amount of computer software. Instead, these will require better methodological training and more careful evaluation by journal editors of submitted articles.

References

- [1] Bakan, D. (1966). The test of significance in psychological research, *Psychological Bulletin* **66**, 423–437.
- [2] Cronbach, L. (1957). The two disciplines of scientific psychology, *American Psychologist* **12**, 671–684.
- [3] Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*, Cambridge University Press, Cambridge.
- [4] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [5] Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning, in *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, Keren, G. & Lewis, C., eds, Lawrence Erlbaum Associates, Hillsdale, pp. 311–339.
- [6] Loftus, G.R. (1993). A picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age, *Behavior Research Methods, Instruments, & Computers* **25**, 250–256.
- [7] Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934–1945, *British Journal of Mathematical and Statistical Psychology* **32**, 151–178.
- [8] Myers, C.S. (1909). *A Text-book of Experimental Psychology*, Edward Arnold, London.
- [9] Neter, J., Wasserman, W., Kutner, M.H., Nachsteim, C.J. & Kutner, M. (1996). *Applied Linear Statistical Models*, 4th Edition, McGraw-Hill, New York.
- [10] Reitz, W. (1934). Statistical techniques for the study of institutional differences, *Journal of Experimental Education* **3**, 11–24.
- [11] Rucci, A.J. & Tweney, R.D. (1980). Analysis of variance and the “Second Discipline” of scientific psychology: an historical account, *Psychological Bulletin* **87**, 166–184.
- [12] Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers, *Psychological Bulletin* **76**, 105–110.

Further Reading

- Capshaw, J.H. (1999). *Psychologists on the March: Science, Practice, and Professional Identity in America, 1929–1969*, Cambridge University Press, Cambridge.
- Cowles, M. (2001). *Statistics in Psychology: An Historical Perspective*, 2nd Edition, Lawrence Erlbaum Associates, Mahwah.
- Fienberg, S.E. (1985). Statistical developments in World War II: an international perspective, in *A Celebration Of Statistics: The ISI Centenary Volume*, Atkinson, A.C. & Fienberg, S.E., eds, Springer, New York, pp. 25–30.
- Herman, E. (1995). *The Romance of American Psychology: Political Culture in the Age of Experts*, University of California Press, Berkeley.
- Lovie, A.D. (1981). On the early history of ANOVA in the analysis of repeated measure designs in psychology, *British Journal of Mathematical and Statistical Psychology* **34**, 1–15.
- Tweney, R.D. (2003). Whatever happened to the brass and glass? The rise of statistical “instruments” in psychology, 1900–1950, in *Thick Description and Fine Texture: Studies in the History of Psychology*, Baker, D., ed., University of Akron Press, Akron, pp. 123–142 & 200–205, notes.

RYAN D. TWENEY

History of Behavioral Statistics

SANDY LOVIE

Volume 2, pp. 826–829

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Behavioral Statistics

If we follow Ian Hacking [4] and conclude that modern ideas about probability emerged a little less than 300 years ago, and that other historians have pointed to the nineteenth century as the start of statistics proper, then it is clear that statistical analysis is a very recent phenomenon indeed. Further, if one also follows historians like Mackenzie [8] and Porter [9], then modern statistics began in Britain and was a feature of the last half of the nineteenth century only. Of course, the collection of data relating to social, economic, and political concerns was of interest to many people of a wide ideological hue in many countries much earlier on in the nineteenth century, as did the reduction of such material to tabular and graphical forms, but there were no officially blessed attempts to draw conclusions from these exercises or to establish any causal links between the material, although there were plenty of amateur efforts to do so. Indeed, the 1838 motto of the London Statistical Society: '*Aliis exterendum*' ('to be threshed out by others'), together with its pledge 'to exclude all opinion' gives one a feel for the extreme empiricism of the so-called *statists* of the time. However, this rule was not one that sprang from mere conviction, but was designed to placate those who argued that drawing conclusions from such potentially explosive material would create such dissent that the Society would split asunder. What seemed to create a change to the more analytical work of figures such as **Francis Galton** and **Karl Pearson** in the nineteenth century, and **R A Fisher** in the twentieth, was the European work on probability by **Laplace** toward the end of the eighteenth century, and the realization by figures such as **Quetelet** that it could be applied to social phenomena. In other words, such developments were spurred on by the near-simultaneous acceptance that the world was fundamentally uncertain, *and* the recognition of the key role of probability theory in defining and managing such an unpredictable situation.

Thus, once it is accepted that the world is fundamentally uncertain and unpredictable, as was increasingly the case during the nineteenth century, then the means for confronting the reality that *nothing* can be asserted or trusted absolutely or unconditionally had

to be urgently conceived. And this was particularly true of the nascent field of statistics, since its ostensive business was the description and management of an uncertain world. In practice, this did not turn out to be an impossible task, although it took a great deal of time and effort for modern statistical thinking and action to emerge. I want to argue that this was achieved by creating a structure for the world which managed the chaotic devil of uncertainty by both constraining and directing the actions and effects of uncertainty. But this could only be achieved through extensive agreement by the principle players as to what was a realistic version of the world and how it's formal, that is, mathematical, description could be derived. This also went hand in hand with the development of workable models of uncertainty itself, with the one feeding into the other. And it is also the case that many of these structural/structuring principles, for example, the description of the world in terms of variables with defined (or even predefined) properties, were taken over by psychology from much earlier scientific traditions, as was the notion (even the possibility) of systematic and controlled experimentation.

Let me briefly illustrate some of these ideas with Adolphe Quetelet's distinction between a true average and an arithmetic average, an idea that he developed in 1846. This was concerned with choosing not just the correct estimator, but also constraining the situation so that what was being measured was in some senses homogeneous, hence any distribution of readings could be treated as an error distribution (a concept that Quetelet, who started out as an astronomer, would have been familiar with from his earlier scientific work). Thus, collecting an unsorted set of houses in a town, measuring them once and then calculating an average height would yield little information either about the houses themselves or about the height of a particular house. On the other hand, repeatedly measuring the height of one house would, if the measurements were not themselves subject to systematic bias or error, inevitably lead to the true average height for that house. In addition, the set of readings would form a distribution of error around the true mean, which could itself be treated as a probability distribution following its own 'natural law'. Notice the number of hidden but vital assumptions here: first, that a house will remain the same height for a reasonable length of time, hence the measurements are of a homogenous part of the

world; second, that height is a useful property of a house that can be expressed as a measurable variable; third, that the measurements are not biased in any way during their collection, hence any variation represents pure error around the true height and not a variation in the height of the building itself; and finally, that the estimator of height has certain useful properties, which commends it to the analyst. Thus, one has a mixture of consensually achieved background ideas about the context, how the particular aspect of interest can be experimentally investigated, and agreement as to what summary numbers and other representations of the sample yield the most useful information. Statistics is about all of these, since all of them affect how any set of numbers is generated and how they are interpreted.

When we move on quickly to the middle and latter parts of the nineteenth century we find that Quetelet's precepts have been accepted and well learnt (although not perhaps his radical social physics, or his emphasis on *l'homme moyen*, or the average man), and that people are now more than happy to draw conclusions from data, since they are convinced that their approach warrants them to do so. Thus, Galton is happy to use the emerging methods of regression analysis to argue a hereditarian case for all manner of human properties, not just his famous one of heights. Karl Pearson's monumental extension and elaboration of both Galton's relatively simple ideas about relationships and his rather plain vanilla version of Darwinian evolution marks the triumph of a more mathematical look to statistics, which is now increasingly seen as the province of the data modeler rather than the mere gatherer-in of numbers and their tabulation. This is also coincident, for example, with a switch from individual teaching and examining to mass education and testing in the schools, thus hastening the use of statistics in psychology and areas related to it like education (see [2], for more information). One cannot, in addition, ignore the large amount of data generated by psychophysicists like **Fechner** (who was a considerable statistician in his own right), or his model building in the name of *panpsychism*, that is, his philosophy that the whole universe was linked together by a form of psychic energy. The 1880s and 1890s also saw the increasing use of systematic, multifactor experimental designs in the psychological and pedagogical literature, including several on the readability of print. In other words, there was an unrequited thirst in psychology for more

quantitative analyses that went hand in hand with how the psychologist and the educational researcher viewed the complexity of their respective worlds.

The first decade or so of the twentieth century also saw the start of the serious commitment of psychology to statistical analysis in so far as this marks the publication of the first textbooks in the subject by popular and well-respected authors like Thorndike. There was also a lively debate pursued by workers for over 30 years as to the best test for the difference between a pair of means. This had been kick started by Yerkes and Thorndike at the turn of the century and had involved various measures of variation to scale the difference. This was gradually subsumed within Student's *t* Test, but the existence of a range of solutions to the problem meant that, unlike **analysis of variance (or ANOVA)**, the acceptance of the standard analysis was somewhat delayed, but notice that the essentially uncertain nature of the world and any data gathered from it had been explicitly recognized by psychologists in this analysis (see [7] for more details). Unfortunately, the major twentieth-century debates in statistics about inference seemed to pass psychology by. It was only in the 1950s, for example, that the Neyman–Pearson work on Type I and II errors (see **Neyman–Pearson Inference**) from the 1930s had begun to seep into the textbooks (see [7], for a commentary as to why this might have been). An exception could possibly be made for Bayesian inference and its vigorous take up in the 1960s by Ward Edwards [3] and others (see his 1963 paper, for instance), but even this seemed to peter out as psychology came to reluctantly embrace power, sample size, and effects with its obvious line to a Neyman–Pearson analysis of inference. Again, such an analysis brings a strong structuring principle and worldview to the task of choosing between uncertain outcomes.

The other early significant work on psychological statistics, which simultaneously both acknowledged the uncertain nature of the world and sought structuring principles to reduce the effects of this uncertainty was **Charles Spearman's** foray into **factor analysis** from 1904 onwards. Using an essentially **latent variable** approach, Spearman looked for support for the strongly held nineteenth-century view that human nature could be accounted for by a single general factor (sometimes referred to as *mental energy*) plus an array of specific factors whose operation would be determined by the individual demands of the situation

or task. This meant that a hierarchical structure could be imposed, *a priori*, on the intercorrelations between the various scholastic test results that Spearman had obtained during 1903. Factor analysis as a deductive movement lasted until the 1930s when the scepticism of **Godfrey Thomson** and the risky, inductive philosophy of Louis Thurstone combined to turn factor analysis into the exploratory method that it has become today (see [7], and [1], for more detail). But notice that the extraction of an uncertainty-taming *structure* is still the aim of the enterprise, whatever flavor of factor analysis we are looking at. And this is also the case for all the multivariable methods, which were originated by Karl Pearson, from **principal component analysis** to **multiple linear regression**.

My final section is devoted to a brief outline of the rapid acceptance by psychology of the most widely employed method in psychological statistics, that is, ANOVA (see [1, 5, 6, 10] for much more detail). This was a technique for testing the differences between more than two samples, which was developed in 1923 by the leading British statistician of this time, R A Fisher, as part of his work in agricultural research. It was also a method which crucially depended for its validity on the source of the data, specifically from experiments that randomly allocated the experimental material, for instance, varieties of wheat, to the experimental plots. Differences over the need for random allocation was the cause of much bitterness between Fisher and 'Student' (**W S Gosset**), but it is really an extension of Quetelet's rule that any variation in the measurement of a homogenous quality such as a single wheat variety should reflect error and nothing else, a property that only random allocation could guarantee. In psychology, ANOVA and its extension to more than one factor were quickly taken into the discipline after the appearance of Fisher's first book introducing the method (in 1935), and was rapidly applied to the complex, multifactor experiments, which psychology had been running for decades. Indeed, so fast was this process that Garrett and Zubin, in 1943, were able to point to over 30 papers and books that used ANOVA and

variations, while the earliest example that I could find of its application to an area close to psychology was by the statistician Reitz who, in 1934, used the technique to compare student performance across schools. Clearly, psychology had long taken people's actions, beliefs, and thought to be determined by many factors. Here at last was a method that allowed them to quantitatively represent and explore such a structuring worldview.

References

- [1] Cowles, M. (2001). *Statistics in Psychology: an Historical Perspective*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [2] Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology, in *The Probabilistic Revolution: Ideas in Modern Science*, Vol. II, G. Gigerenzer, L. Kruger & M. Morgan, eds, MIT Press, Cambridge.
- [3] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research, *Psychological Review* **70**(3), 193–242.
- [4] Hacking, I. (1976). *The Emergence of Probability*, Cambridge University Press, Cambridge.
- [5] Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934–1945, *British Journal of Mathematical and Statistical Psychology* **32**, 151–178.
- [6] Lovie, A.D. (1981). On the early history of ANOVA in the analysis of repeated measure designs in psychology, *British Journal of Mathematical and Statistical Psychology* **34**, 1–15.
- [7] Lovie, A.D. (1991). A short history of statistics in Twentieth Century psychology, in *New Developments in Statistics for Psychology and the Social Sciences*, Vol. 2, P. Lovie & A.D. Lovie, eds, BPS Books & Routledge, London.
- [8] Mackenzie, D. (1981). *Statistics in Britain, 1865–1930: The Social Construction of Scientific Knowledge*, Edinburgh University Press, Edinburgh.
- [9] Porter, T.M. (1986). *The Rise in Statistical Thinking: 1820–1900*, Princeton University Press, Princeton.
- [10] Rucci, A.J. & Tweney, R.D. (1979). Analysis of variance and the 'Second Discipline' of scientific psychology, *Psychological Bulletin* **87**, 166–184.

SANDY LOVIE

History of the Control Group

TRUDY DEHUE

Volume 2, pp. 829–836

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of the Control Group

A Special Kind of Group

Currently, anyone with a qualification in social science (and also medicine) takes the notion of the 'control group' for granted. Yet, when one comes to think of it, a control group is an exceptional kind of group. Usually, the notion of a 'group' refers to people sharing a particular identity or aim, having a sense of oneness, and most likely also a group leader. Members of control groups, however, do not even need to know one another. One may run into groups of tourists, teenagers, hikers, or hooligans, but never a group of controls. Different from groups in the regular sense, control groups are merely a number of people. They do not exist outside the realm of human science experimentation, and *qua* group, they exist only in the minds of the experimenters who compose and study them. Members of control groups are not supposed to develop group cohesion because that would entail a contaminating factor in the experiment. Researchers are interested in the average response of individual members rather than in their behavior as a group (see **Clinical Trials and Intervention Studies**).

Control groups serve to check the mean effectiveness of an intervention. In order to explain their need, a methods teacher will first suggest that if a particular person shows improvement this does not yet guarantee effectiveness in other people too. Therefore, a number of people must be studied and their average response calculated. This, however (the teacher will proceed), would not be enough. If we know the average result for one group of people, it remains unclear whether the intervention caused the outcome or a simultaneous other factor did. Only if the mean result of the group differs significantly from that of an untreated control group is the conclusion justified that the treatment caused the effect (or, more precisely, that the effect produced was not accidental).

Apart from the idea of comparing group averages, the notion of the true control group entails one pivotal extra criterion. The experimental and control groups must be comparable, that is, the former should not differ from the latter except through the action of

the independent variable that is to be tested. One option is to use the technique of 'matching', that is, to pretest people on factors suspected of causing bias and then create groups with similar test results. Most contemporary methodologists, however, agree that random assignment of the participants to the groups offers a better guarantee of comparability. Assigning people on the basis of chance ensures that both known and unknown invalidating factors are cancelled out, and that this occurs automatically rather than being dependent on individual judgment and trustworthiness. In behavioral and social (as well as clinical) research, the ideal scientific experiment is a so-called *randomized controlled trial*, briefly an RCT.

In view of its self-evident value, and in view of the extensive nineteenth-century interest in human science experimentation, the introduction of the control group may strike us as being remarkably late. Until the early 1900s, the word control was not used in the context of comparative experiments with people, whereas ensuring comparability by matching dates back to the 1910s, and composing experimental and control groups at random was first suggested as late as the 1920s.

The present history connects the seemingly late emergence of the control group to its special nature as a group without a shared extra-individual identity. Moreover, it explains that such groups were inconceivable before considerable changes occurred in society at large. First, however, we need to briefly explain why famous examples of comparison such as that of doctor Ignaz Semmelweis, who fought childbed fever by comparing two maternity clinics in mid-nineteenth-century Vienna [23], are not included in the present account.

The Past and the Present

Comparison is 'a natural thing to do' to anyone curious about the effects of a particular action. Therefore, it should not come as a surprise that instances of comparison can also be found in the long history of interventions into human life. In a scholarly article on the history of experimentation with medical treatments, Ted Kaptchuck discussed various eighteenth-century procedures of comparison (although not to similar groups) such as the deliberately deceptive provision of bread and sugar pills

2 History of the Control Group

to check the claims of homeopathy [24]. And several examples of group comparison in the treatment of illnesses (although without randomization) are also presented in the electronic *James Lind Library* (www.jameslindlibrary.org).

Entertaining, however, as such examples of comparison may be, they are hardly surprising, since checking the effects of one's actions by sometimes withholding them is a matter of everyday logic. Moreover, it would be quite artificial to depict these examples as early, if still incomplete, steps toward the present-day methodological rule of employing control groups. Historians of science use derogatory labels such as 'presentist history', 'finalist history', 'justificationary history', and 'feel good history', for histories applying present-day criteria in selecting 'predecessors' who took 'early steps' toward our own viewpoints, whilst also excusing these 'pioneers' for the understandable shortcomings still present in their ideas. Arranging the examples in chronological order, as such histories do, suggests a progressive trajectory from the past to the present, whereas they actually drew their own line from the present back into the past. Historian and philosopher of science Thomas Kuhn discussed the genre under the name of 'preface history', referring to the typical historical introduction in textbooks. Apart from worshipping the present, Kuhn argued, preface histories convey a deeply misleading view of scientific development as a matter of slow, but accumulative, discovery by a range of mutually unrelated great men [25, pp. 1–10; 136–144].

Rather than lining up unconnected look-alikes through the ages, the present account asks when, why, and how employing control groups became a methodological condition. The many reputed nineteenth-century scholars who explicitly *rejected* experimental comparison are neither scorned nor excused for their 'deficiency'. Rather, their views are analyzed as contributions to debates in their own time. Likewise, the ideas of early twentieth-century scholars who advanced group comparison are discussed as part of debates with their own contemporaries.

Nineteenth-century Qualms

If control groups were not recommended before the early twentieth century, the expression of "social experimentation" did appear in much earlier methodological texts. Eighteenth-century scholars had

already discussed the issue of experimentation as a suitable method for investigating human life [7]. David Hume's *Treatise of Human Nature*, first published in 1739, is subtitled: *Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Hume and his Enlightenment contemporaries, however, borrowed the terminology of experimentation from natural science as a metaphor for major events happening without the intervention of researchers. Observing disturbances of regular life, they argued, is the human science substitute of natural science experimentation.

Nineteenth-century views on social experimentation were largely, but not entirely, the same as those of the eighteenth century. Distinguished scholars such as **Adolphe Quetelet** (1796–1874) in Belgium, Auguste Comte (1798–1857) in France, and George Cornwall Lewis (1806–1863) as well as John Stuart Mill (1806–1873) in Britain used the terminology of experimentation for incidents such as natural disasters, famines, economic crises, and also government interventions. Different, however, from eighteenth-century scholars and in accordance with later twentieth-century views, they preserved the epithet of *scientific* experimentation for experiments with active manipulation by researchers. As scientific experimentation entails intentional manipulation by researchers, they maintained, research with human beings cannot be scientific.

Roughly speaking, there were two reasons why they excluded deliberate manipulation from the useable methods of research with human beings. One reason was of a moral nature. When George Cornwall Lewis in 1852 published his two-volume *Treatise on the Methods of Observation and Reasoning in Politics*, he deliberately omitted the method of experimentation from the title. Experimentation, Lewis maintained, is 'inapplicable to man as a sentient, and also as an intellectual and moral being'. This is not 'because man lies beyond the reach of our powers', but because experiments 'could not be applied to him without destroying his life, or wounding his sensibility, or at least subjecting him to annoyance and restraint' [26, pp. 160–161].

The second reason was of an epistemological nature. In 1843, the prominent British philosopher, economist, and methodologist John Stuart Mill published his *System of Logic* that was to become very influential in the social sciences. This work extensively discussed Mill's 'method of difference', which

entailed comparing cases in which an effect does and does not occur. According to Mill, this ‘most perfect of the methods of experimental inquiry’ was not suitable for research with people. He illustrated this view with the ‘frequent topic of debate in the present century’, that is, whether or not government intervention into free enterprise impedes national wealth. The method of difference is unhelpful in a case like this, he explained, because comparability is not achievable: ‘[I]f the two nations differ in this portion of their institutions, it is from some differences in their position, and thence in their apparent interests, or in some portion or the other of their opinions, habits and tendencies; which opens a view of further differences without any assignable limit, capable of operating on their industrial prosperity, as well as on every other feature of their condition, in more ways than can be enumerated or imagined’ [31, pp. 881–882].

Mill raised the objection of incomparability not only in complex issues such as national economic policies but in relation to all research with people. Even a comparatively simple question such as, whether or not mercury cures a particular disease, was ‘quite chimerical’ as it was impossible in medical research to isolate a single factor from all other factors that might constitute an effect. Although the efficacy of ‘quinine, colchicum, lime juice, and cod liver oil’ was shown in so many cases ‘that their tendency to restore health... may be regarded as an experimental truth’, real experimentation was out of the question, and ‘[S]till less is this method applicable to a class of phenomena more complicated than those of physiology, the phenomena of politics and history’ [31, pp. 451–452].

Organicism and Determinism

How to explain the difference between these nineteenth-century objections and the commonness of experimentation with experimental and control groups in our own time? How could Lewis be compunctious about individual integrity even to the level of not ‘annoying’ people, whereas, in our time, large group experiments hardly raised an eyebrow? And why did a distinguished methodologist like Mill not promote the solution, so self-evident to present-day researchers, of simply *creating* comparable groups if natural ones did not exist?

The answer is that their qualms were inspired by the general holism and determinism of their time. Nineteenth-century scholars regarded communities as well as individuals as organic systems in which every element is closely related to all others, and in which every characteristic is part of an entire pattern of interwoven strands rather than caused by one or more meticulously isolated factors. In addition, they ascribed the facts of life to established laws of God or Nature rather than to human purposes and plans. According to nineteenth-century determinism, the possibilities of engineering human life were very limited. Rather than initiating permanent social change, the role of responsible authorities was to preserve public stability. Even Mill, for whom the disadvantages of a *laissez-faire* economy posed a significant problem, nevertheless, held that government interference should be limited to a small range of issues and should largely aim at the preservation of regular social order.

In this context, the common expression of ‘social experimentation’ could not be more than a metaphor to express the view that careful observation of severe disturbances offers an understanding of the right and balanced state of affairs. The same holistic and determinist philosophy expressed itself in nineteenth-century statistics, where indeterminism or chance had the negative connotation of lack of knowledge and whimsicality rather than the present-day association of something to ‘take’ and as an instrument to make good use of [33, 22]. Nineteenth-century survey researchers, for instance, did not draw representative population samples. This was not because of the inherent complexity of the idea, nor because of sluggishness on the researchers’ part, but because they investigated groups of people as organic entities and prototypical communities [17]. To nineteenth-century researchers, the idea of using chance for deriving population values, or, for that matter, allocating people to groups, was literally unimaginable.

Even the occasional proponent of active experimentation in clinical research rejected chance as an instrument of scientific research. In 1865, the illustrious French physiologist Claude Bernard (1813–1878) published a book with the deliberately provocative title of *Introduction à L’étude de la Médecine Expérimentale* [1] translated into English as *An Introduction to the Study of Experimental Medicine*. Staunchly, Bernard stated that ‘philosophic obstacles’

4 History of the Control Group

to experimental medicine ‘arise from vicious methods, bad mental habits, and certain false ideas’ [2, p. 196]. For the sake of valid knowledge, he maintained, ‘comparative experiments have to be made at the same time and on as comparable patients as possible’ [2, p. 194].

Yet, one searches Bernard’s *Introduction* in vain for comparison of experimental to control groups. As ardently as he defended experimentation, he rejected statistical averages. He sneered about the ‘startling instance’ of a physiologist who collected urine ‘from a railroad station urinal where people of all nations passed’ as if it were possible to analyze ‘the *average* European urine!’ (italics and exclamation mark in original). And he scorned surgeons who published the success rates of their operations because average success does not give any certainty on the next operation to come. Bernard’s expression of ‘comparative experimentation’ did refer to manipulating animals and humans for the sake of research. Instead of comparing group averages, however, he recommended that one should present ‘our most perfect experiment as a type’ [2, pp. 134–135]. To Bernard, the rise of probabilistic statistics meant ‘literally nothing scientifically’ [2, p. 137].

Impending Changes

The British statistician, biometrician, and eugenicist **Sir Francis Galton** (1822–1911) was a crucial figure in the gradual establishment of probabilism as an instrument of social and scientific progress. Galton was inspired by Adolphe Quetelet’s notion of the statistical mean and the normal curve as a substitute for the ideal of absolute laws. In Quetelet’s own writings, however, this novelty was not at odds with determinism. His well-known *L’homme moyen* (average man) represented normalcy and dispersion from the mean signified abnormality. It was Galton who gave Quetelet’s mean a progressive twist.

Combining the evolution theory of his cousin Charles Darwin with eugenic ideals of human improvement, Galton held that ‘an average man is morally and intellectually a very uninteresting being. The class to which he belongs is bulky, and no doubt serves to keep the course of social life in action... But the average man is of no direct help towards evolution, which appears to our dim vision to be the

primary purpose, so to speak, of all living existence’, whereas ‘[E]volution is an unrelenting progression,’ Galton added, ‘the nature of the average individual is essentially unprogressive’ [20, p. 406].

Galton was interested in finding more ways of employing science for the sake of human progress. In an 1872 article ‘Statistical Inquiries into the Efficacy of Prayer’, he questioned the common belief that ‘sick persons who pray, or are prayed for, recover on the average more rapidly than others.’ This article opened with the statement that there were two methods of studying an issue like the profits of piety. The first one was ‘to deal with isolated instances’. Anyone, however, using that method should suspect ‘his own judgments’ or otherwise would ‘certainly run the risk of being suspected by others in choosing one-sided examples’. Galton vigorously broke a lance for substituting the study of representative types with statistical comparison. The most reliable method was ‘to examine large classes of cases, and to be guided by broad averages’ [19, p. 126].

Galton elaborately explained how the latter method could be applied in finding out the revenues of praying: ‘We must gather cases for statistical comparison, in which the same object is keenly pursued by two classes similar in their physical but opposite in their spiritual state; the one class being spiritual, the other materialistic. Prudent pious people must be compared with prudent materialistic people and not with the imprudent nor the vicious. We simply look for the final result - whether those who pray attain their objects more frequently than those who do not pray, but who live in all other respects under similar conditions’ [19, p. 126].

As it seems, Galton was the first to advocate comparison of group averages. Yet, his was not an example of treating one group while withholding the treatment from a comparison group. The emergence of the control group in the present-day sense occurred when his fears of ‘being suspected by others in choosing one-sided examples’ began to outgrow anxieties on doing injustice to organic wholes. This transition took place with the general changeover from determinism to progressivism in a philosophical as well as social sense.

Progressivism and Distrust

By the end of the nineteenth century, extreme destitution among the working classes led to social

movements for mitigation of *laissez faire* capitalism. Enlightened members of the upper middle class pleaded for some State protection of laborers via minimum wage bills, child labor bills, and unemployment insurances. Their appeals for the extension of government responsibility met with strong fears that help would deprive people of their own responsibility and that administrations would squander public funds. It was progressivism combined with distrust that constituted a new definition of social experimentation as statistical comparison of experimental and control groups. Three interrelated maxims of twentieth-century economic liberalism were crucial to the gradual emergence of the present-day ideal experiment.

The first maxim was that of *individual responsibility*. Social success and failure remained an individual affair. This implied that ameliorative attempts were to be directed first and foremost at problematic individuals rather than on further structural social change. Helping people implied trying to turn them into independent citizens by educating, training, punishing, and rewarding them. The second maxim was that of *efficiency*. Ameliorative actions financed with public money had to produce instant results with simple economical means. The fear that public funds would be squandered created a strong urge to attribute misery and backwardness to well-delineated causes rather than complex patterns of individual and social relations. And the third maxim was that of *impersonal procedures*. Fears of abuse of social services evoked distrust of people's own claims of needs, and the consequent search for impersonal techniques to establish the truth 'behind' their stories [38]. In addition, not only was the self-assessment of the interested recipients of help to be distrusted but also that of the politicians and administrators providing help. Measurement also had to control administrators' claims of efficiency [34].

Academic experts on psychological, sociological, political, and economical matters adapted their questions and approaches to the new demands. They began to produce technically useful data collected according to standardized methodological rules. Moreover, they established a partnership with statisticians who now began to focus on population varieties rather than communalities. In this context, the interpretation of chance as something one must make good use of replaced the traditional one of chance as something to defeat [17, 22, 33].

The new social scientists measured people's abilities, motives, and attitudes, as well as social phenomena such as crime, alcoholism, and illiteracy. Soon, they arrived at the idea that these instruments could also be used for establishing the results of ameliorative interventions. In 1917, the well-reputed sociologist F. Stuart Chapin lengthily discussed the issue. Simple, before and after measurement of one group, he stated, would not suffice for excluding personal judgement. Yet, Chapin rejected comparison of treated and untreated groups. Like Mill before him, he maintained that fundamental differences between groups would always invalidate the conclusions of social experiments. Adding a twentieth-century version to Lewis' moral objections, he argued that it would be immoral to withhold help from needy people just for the sake of research [9, 10]. It was psychologists who introduced the key idea to create equal groups rather than search for them in natural life, and they did so in a context with few ethical barriers.

Creating Groups

Psychologists had a tradition of psychophysiological experimentation with small groups of volunteers in laboratory settings for studying the law-like relationships between physical stimuli and mental sensations. During the administrative turn of both government and human science, many of them adapted their psychophysiological methods to the new demands of measuring progress rather than just discovering laws [14, 15]. One of these psychologists was John Edgar Coover, who studied at Stanford University in Palo Alto (California) with the psychophysical experimenter Frank Angell. As a former school principal, Coover gave Angell's academic interests an instrumental twist. He engaged in a debate among school administrators on the utility of teaching subjects such as Latin and formal mathematics. Opponents wanted to abolish such redundant subjects from the school curriculum, but proponents argued that 'formal discipline' strengthens general mental capacities. Coover took part in this debate with laboratory experiments testing whether or not the training of one skill improves performance in another ability. In a 1907 article, published together with Angell, he explained that in the context of this kind of research a one-group design does not do. Instead, he compared the achievements of experimental 'reagents'

who received training with those of control ‘reagents’ who did not [13]. Coover and Angell’s article seems to be the first report of an experiment in which one group of people is treated and tested, while another one is only tested.

From the 1910s, a vigorous movement started in American schools for efficiency and scientific (social) engineering [6]. In the school setting, it was morally warrantable and practically doable to compare groups. Like the earlier volunteers in laboratories, school children and teachers were comparatively easy to handle. Whereas historian Edwin Boring found no control groups in the 1916 volume of the *American Journal of Psychology* [3, page 587], historian Kurt Danziger found 14 to 18% in the 1914–1916 volumes of the *Journal of Educational Psychology* [14, pp. 113–115].

Psychological researchers experimented in real classrooms where they tested the effects of classroom circumstances such as fresh versus ventilated air, the sex of the teacher, memorizing methods, and educational measures such as punishing and praising. They also sought ways of excluding the possibility that their effects are due to some other difference between the groups than the variable that is tested. During the 1920s, it became customary to handle the problem by matching. Matching, however, violated the guiding maxims of efficiency and impersonality. It was quite time- and money-consuming to test each child on every factor suspected of creating bias. And, even worse, determining these factors depended on the imaginative power and reliability of the researchers involved. Matching only covered possibly contaminating factors that the designers of an experiment were aware of, did not wish to neglect, and were able to pretest the participants on.

In 1923, William A. McCall at Columbia University in New York, published the methodological manual *How to Experiment in Education* in which he emphasized the need of comparing similar groups [30]. In the introduction to this volume, McCall predicted that enhancing the efficiency of education could save billions of dollars. Further on, he proposed to equate the groups on the basis of chance as ‘an economical substitute’ for matching. McCall did not take randomization lightly. For example, he rejected the method of writing numbers on pieces of paper because papers with larger numbers contain more ink and are therefore likely to sink further to the bottom of a container. But, he stated, ‘any

device which will make the selection truly random is satisfactory’ [30, pp.41–42].

Fisher’s Support

In the meantime, educational psychologists were testing various factors simultaneously, which made the resulting data hard to handle. The methodological handbook *The Design of Experiments* published in 1935 by the British biometrician and agricultural statistician **Ronald A. Fisher** provided the solution of **analysis of variance (ANOVA)**. As Fisher repeatedly stressed, random allocation to groups was a central condition to the validity of this technique. When working as a visiting professor at the agricultural station of Iowa State College, he met the American statistician **George W. Snedecor**. Snedecor published a book based on Fisher’s statistical methods [37] that was easier to comprehend than Fisher’s own, rather intricate, writings and that was widely received by methodologists in biology as well as psychology [28, 35]. Subsequently, Snedecor’s Iowa colleague, the educational psychologist Everett Lindquist, followed with the book *Statistical Analysis in Educational Research* which became a much-cited source in the international educational community [27].

Fisher’s help was welcomed with open arms by methodologists, not only because it provided a means to handle multi factor research but also because it regulated experimentation from the stage of the experimental design. As Snedecor expressed it in 1936, the designs researchers employed often ‘bafled’ the statisticians. ‘No more than a decade past, the statistician was distinctly on the defence’, he revealed, but ‘[U]nder the leadership of R. A. Fisher, the statistician has become the aggressor. He has found that the key to the problem is the intimate relation between the statistical method and the experimental plan’ [36, p. 690]. This quote confirms the thesis of historians that the first and foremost motive to prescribe randomization was not the logic of probabilistic statistics, but the wish to regulate the conduct of practicing researchers [8, 16, 29, 34]. Canceling out personal judgment, together with economical reasons, was the predominant drive to substitute matching by randomization. Like Galton in 1872, who warned against eliciting accusations of having chosen one-sided examples, early twentieth-century

statisticians and methodologists cautioned against the danger of selection bias caused by high hopes on particular outcomes.

Epilogue

It took a while before **randomization** became more than a methodological ideal. Practicing physicians argued that the hopes of a particular outcome are often a substantial part of the treatment itself. They also maintained that it is immoral to let chance determine which patients gets the treatment his doctor believes in and which patient does not, as well as keeping it a secret as to which group a patient has been assigned. Moreover, they put forward the argument that subjecting patients to standardized tests rather than examining them in a truly individual way would harm, rather than enhance, the effectiveness of diagnoses and treatments.

In social research, there were protests too. After he learned about the solution of random allocation, sociologist F. Stuart Chapin unambiguously rejected it. Allocating people randomly to interventions, he maintained, clashes with the humanitarian mores of reform [11, 12]. And the Russian-American anthropologist Alexander Goldenweiser objected that human reality ‘resents highhanded manipulation’ for which reason it demands true dictators to ‘reduce variety by fostering uniformity’ [21, p. 631]. An extensive search for the actual use of random allocation in social experiments led to the earliest instance in a 1932 article on educational counseling of university students, whereas the next seven appeared in research reports dating from the 1940s (all but one in the field of educational psychology) [18].

Nevertheless, the more twentieth-century welfare capitalism replaced nineteenth-century *laissez-faire* capitalism, the more administrators and researchers felt that it is both necessary and morally acceptable to experiment with randomized groups of children as well as adults. From about the 1960s onward, therefore, protesting doctors could easily be accused of an unwillingness to give up an outdated elitist position for the truly scientific attitude. Particularly in the United States, the majority of behavioral and social researchers too began to regard experiments with randomly composed groups as the ideal experiment. Since President Johnson’s War on Poverty, many such social experiments have been conducted, sometimes with thousands of people. Apart from school

children and university students, also soldiers, slum dwellers, spouse beaters, drug abusers, disabled food-stamp recipients, bad parents, and wild teenagers have all participated in experiments testing the effects of special training, social housing programs, marriage courses, safe-sex campaigns, health programs, income maintenance, employment programs, and the like, in an impersonal, efficient, and standardized way [4, 5, 32].

References

- [1] Bernard, C. (1865). *Introduction à L’étude de la Médecine Expérimentale*, Ballière, Paris.
- [2] Bernard, C. (1957). *An Introduction to the Study of Experimental Medicine*, Dover Publications, New York.
- [3] Boring, E.G. (1954). The nature and history of experimental control, *American Journal of Psychology* **67**, 573–589.
- [4] Boruch, R. (1997). *Randomised Experiments for Planning and Evaluation*, Sage Publications, London.
- [5] Bulmer, M. (1986). Evaluation research and social experimentation, in *Social Science and Social Policy*, M. Bulmer, K.G. Banting, M. Carley & C.H. Weiss, eds, Allen and Unwin, London, pp. 155–179.
- [6] Callahan, R.E. (1962). *Education and the Cult of Efficiency*, The University of Chicago Press, Chicago.
- [7] Carrithers, D. (1995). The enlightenment science of society, in *Inventing Human Science. Eighteenth-Century Domains*, C. Fox, R. Porter & R. Wokler, eds, University of California Press, Berkeley, pp. 232–270.
- [8] Chalmers, I. (2001). Comparing like with like. Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments, *International Journal of Epidemiology* **30**, 1156–1164.
- [9] Chapin, F.S. (1917a). The experimental method and sociology. I. The theory and practice of the experimental method, *Scientific Monthly* **4**, 133–144.
- [10] Chapin, F.S. (1917b). The experimental method and sociology. II. Social legislation is social experimentation, *Scientific Monthly* **4**, 238–247.
- [11] Chapin, F.S. (1938). Design for social experiments, *American Sociological Review* **3**, 786–800.
- [12] Chapin, F.S. (1947). *Experimental Designs in Social Research*, Harper & Row, New York.
- [13] Coover, J.E. & Angell, F. (1907). General practice effect of special exercise, *American Journal of Psychology* **18**, 328–340.
- [14] Danziger, K. (1990). *Constructing the Subject*, Cambridge University Press, Cambridge.
- [15] Dehue, T. (2000). From deception trials to control reagents. The introduction of the control group about a century ago, *American Psychologist* **55**, 264–269.
- [16] Dehue, T. (2004). Historiography taking issue. Analyzing an experiment with heroin maintenance, *Journal of the History of the Behavioral Sciences* **40**(3), 247–265.

8 History of the Control Group

- [17] Desrosières, A. (1998). *The Politics of Large Numbers. A History of Statistical Reasoning*, Harvard University Press, Cambridge.
- [18] Forsetlund, L., Bjørndal, A. & Chalmers, I. (2004, submitted for publication). Random allocation to assess the effects of social interventions does not appear to have been used until the 1930s.
- [19] Galton, F. (1872). Statistical inquiries into the efficacy of prayer, *Fortnightly Review* **XII**, 124–135.
- [20] Galton, F. (1889). Human variety, *Journal of the Anthropological Institute* **18**, 401–419.
- [21] Goldenweiser, A. (1938). The concept of causality in physical and social science, *American Sociological Review* **3**(5), 624–636.
- [22] Hacking, I. (1990). *The Taming of Chance*, Cambridge University Press, New York.
- [23] Hempel, C.G. (1966). *Philosophy of Natural Science*, Prentice-Hall, Englewood Cliffs.
- [24] Kaptchuck, T.J. (1998). Intentional ignorance: a history of blind assessment and placebo controls in medicine, *Bulletin of the History of Medicine* **72**, 389–433.
- [25] Kuhn, T.S. (1962, reprinted 1970). *The Structure of Scientific Revolutions*, Chicago University Press, Chicago.
- [26] Lewis, C.G. (1852, reprinted 1974). *A Treatise on the Methods of Observation and Reasoning in Politics*, Vol 1, Arno Press, New York.
- [27] Lindquist, E.F. (1940). *Statistical Analysis in Educational Research*, Houghton-Mifflin, Boston.
- [28] Lovie, A.D. (1979). The analysis of variance in experimental psychology: 1934–1945, *British Journal of Mathematical and Statistical Psychology* **32**, 151–178.
- [29] Marks, H.M. (1997). *The Progress of Experiment. Science and Therapeutic Reform in the United States, 1900–1990*, Cambridge University Press, New York.
- [30] McCall, W.A. (1923). *How to Experiment in Education*, McMillan McCall, New York.
- [31] Mill, J.S. (1843, reprinted 1973). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*, University of Toronto, Toronto.
- [32] Orr, L. (1999). *Social Experiments. Evaluating Public Programs with Experimental Methods*, Sage Publications, London.
- [33] Porter, T.M. (1986). *The Rise of Statistical Thinking, 1820–1900*, Princeton University Press, Princeton.
- [34] Porter, T.M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton.
- [35] Rucci, A.J. and Ryan D.T. (1980). Analysis of variance and the ‘second discipline’ of scientific psychology: a historical account, *Psychological Bulletin* **87**, 166–184.
- [36] Snedecor, G.W. (1936). The improvement of statistical techniques in biology, *Journal of the American Statistical Association* **31**, 690–701.
- [37] Snedecor, G.W. (1937). *Statistical Methods*, Collegiate Press, Ames, Iowa.
- [38] Stone, D.A. (1993). Clinical authority in the construction of citizenship, in *Public Policy for Democracy*, H. Ingram & S. Rathgeb Smith, eds, Brookings Institution, Washington, pp. 45–68.

TRUDY DEHUE

History of Correlational Measurement

MICHAEL COWLES

Volume 2, pp. 836–840

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Correlational Measurement

Shakespeare in *Julius Caesar* avers, ‘Yond Cassius has a lean and hungry look; He thinks too much. Such men are dangerous.’ This and more mundane generalities, such as, ‘Like father, like son,’ have been quoted for centuries. They illustrate perceived relationships between both different and similar variables, as in those between offspring and parental characteristics. However, it was not until the nineteenth century that attempts began to quantify such relationships, work that is generally attributed to **Sir Francis Galton** (1822–1911) and his disciple and friend, **Karl Pearson** (1857–1936).

However, a number of mathematicians had worked on the measurement of the relationships among variables. Helen Walker [5] has observed that Plana, Adrain, **Gauss**, and **Laplace** examined probabilities in the context of the simultaneous occurrence of pairs of errors, and formulae that included a product term were suggested. De Forest and Czuber also discussed the problem. This work took place in the context of the *Law of Frequency of Error*, now termed the *normal distribution*. It is not too surprising that the mathematics of error estimation were adopted by social scientists especially when **Quetelet** equated error with the deviations from the mean of many measured human and characteristics.

But perhaps the best-known writer in the field was a French mathematician who was examining the question of error estimation in the context of astronomy, although it was treated as a mathematical exercise rather than the development of a practical statistical tool. Auguste Bravais formulated a mathematical theory, and published a lengthy paper on the matter in 1846, but apparently failed to appreciate that the topic had wider application than in the fields of astronomy [1]. Also, land surveying, physics, and gambling problems were recognized as having relevance to the estimation of errors and pairs of errors. However, Bravais’ work has been suggested as the natural precursor to Karl Pearson’s work, at least until as recently as the 1960s and beyond. What we now commonly refer to as **Pearson’s product-moment coefficient** was called the *Bravais–Pearson coefficient* by some researchers.

What is more, in 1896, Pearson himself acknowledged Bravais’ work in the paper that gave us the product-moment coefficient, but later denied that he had been helped by his contribution. But, it would be foolish to reject Bravais’ work for it is certainly the earliest most complete account of the mathematical foundations of the correlation coefficient.

In 1892, Edgeworth examined correlation and methods of estimating correlation coefficients in a series of papers. The first of these papers was *Correlated Averages*. Edgeworth, an economist who appears to be a self-taught mathematician, was someone who Pearson said that the biometricians might claim as one of their own. As early as 1885, Edgeworth was working on calculations related to **analysis of variance**. But, as so often was the case, Pearson fell out with his ingenious and knowledgeable statistical colleague and denied that Edgeworth’s work had influenced him.

In modern notation, the product-moment formula is $r = \sum xy / \sqrt{\sum x^2 \sum y^2}$.

This formula is based on x and y , the deviations of the measurements from the means of the two sets of measurements. In precomputer days, it was the formula of choice, and Pearson [3] noted that the formula ‘presents no practical difficulty in calculation, and therefore we shall adopt it.’ A further requirement of this correlational procedure is that the variables that are put into the exercise are *linearly* related. That is to say that when graphs of the pairs of variables are examined they are best shown as a straight line. A model is produced that is termed the general **linear model**, which may be depicted thus, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + e$. The independent variables are chosen and/or manipulated by the investigator, Y is the dependent variable and e is random error. The model is often termed a *probabilistic* model because it is based on a sample drawn randomly from the population. The coefficient of multiple correlation may be calculated from the separate simple relationships.

For the situation where we have two independent variables,

$$R^2_{Y.XZ} = \frac{r^2_{YX} + r^2_{YZ} - 2r_{YX}r_{YZ}r_{XZ}}{1 - r^2_{XZ}} \quad (1)$$

where the r s are the simple correlations of the labelled variables. Essentially, what we have here

2 History of Correlational Measurement

is a correlation coefficient that takes into account the correlation, the overlap between the independent variables.

This procedure is just one of the techniques used in *univariate statistical analysis*.

It must be appreciated that the model is applicable to cases where the fundamental question is, ‘what goes with what?’ – the correlational study, and ‘How is the dependent variable changed by the independent variables that have been chosen or manipulated?’ – the formal experiment. Some workers have been known to reject the correlational study largely because the differences in the dependent variable are, in general, individual differences or ‘errors.’ The true experiment attempts to reduce error so that the effect of the independent variables is brought out. Moreover, the independent variables in the correlational study are most often, but not always, continuous variables, whereas these variables in, for example, the analysis of variance are more likely to be categorical. The unnecessary disputes arise from the historical investigation of variate and categorical data and do not reflect the mathematical bases of the applications.

Among the earliest of studies that made use of the idea of correlational measurement in the fields of the biological and social sciences was the one carried out in 1877 by an American researcher, Henry Bowditch, who drew up correlation charts based on data from a large sample of Massachusetts school children. Although he did not have a method to compute measures of correlation, there is no doubt that he thought that one was necessary, as was an assessment of partial correlation. It was at this time that Sir **Francis Galton**, a founding father of statistical techniques, was in correspondence with Bowditch and who was himself working on the measurement of what he termed *correlation* and on the beginnings of regression analysis.

The **partial correlation** is the correlation of two variables when a third is held constant. For example, there is a correlation between height and weight in children, but the relationship is affected by the fact that age will influence the variables.

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (2)$$

and, for four variables a, b, c , and d , we find that $\frac{r_{ac}}{r_{ad}} = \frac{r_{bc}}{r_{bd}}$ and therefore $r_{ac} \cdot r_{bd} - r_{ad} \cdot r_{bc} = 0$. If there are two variables a and b and g is a constant, then

$r_{ab} = r_{ag} \cdot r_{bg}$, and, when a is set equal to b , the variance in a would be accounted for by g and this leads us to what are called the *communalities* in a correlation matrix. This approach leads us to Spearman’s *tetrad differences* and the beginnings of what he thought was a mathematical approach to his two-factor theory of intelligence and the development of the methods of **factor analysis**.

The general aim of using correlation to identify specific and general intelligences began to occupy a number of researchers, notably **Thurstone** and Kelley in the United States, and **Thomson** and **Burt** in Britain. The ongoing problem for many researchers was the subjective element in the methods. The fact that they did not produce determinate results reflected an argument that has not yet totally expired, ‘what goes into the analysis’ say the critics, ‘reflects what comes out.’ But increasing attention was given to developing methods that avoid subjective decisions. Apart from its beginnings in the study of intelligence and ability, factor analysis is used by a number of workers in the field of personality research in attempts to produce nonsubjective assessments of the existence of personality traits.

The growing use of factor analytic techniques produced a burgeoning of interest in the assessment of the *reliability* of tests. They were becoming increasingly sophisticated as researchers worried not only about their *validity* – and validity had been largely a matter of subjective face validity – but also of the respectability of their reliability. A leading scholar in this field was Cronbach who listed those aspects of test reliability that are of concern. They are test-retest reliability – is a test consistent over repeated administrations?; internal consistency – do the test items relate to the whole set of items?; alternate or parallel forms reliability – do equivalent forms of the test show high correlations? Cronbach himself offered a useful test, Cronbach’s α . A popular early test of reliability, the Kuder–Richardson estimate of reliability was developed to offset the difficulties of split-half methods, and the Spearman–Brown formula that compares the reliabilities of tests with their length. $r_{nn} = nr_{11}/1 + (n - 1)r_{11}$, where n is test length and r_{11} is the reliability of the test of unit length.

Galton’s view that ability, talent, and intellectual power are characteristics that are primarily innately determined – the *nature side* of the *nature-nurture* issue – sparked a series of investigations that examined the weights and the sizes of sweet pea seeds

over two generations. Later, he looked at human characteristics in a similar context, these latter data being collected by offering prizes for the submission of family records of physical endowments and from visitors to an anthropometric laboratory at the International Health Exhibition, held in 1884. He pondered on the data and noted (the occasion was when he was waiting for a train, which shows that his work was never far from his thoughts) that the frequency of adult children's measurements of height charted against those of the parents (he had devised a measure that incorporated the heights of both parents) produced a set of ellipses centred on the mean of all the measurements. This discovery provided Galton with a method of describing the relationship between parents and offspring using the regression slope.

An event of greatest importance led to the investigations that were to lie at the heart of the new science of biometrics. In his memoirs, [2] Galton noted that,

As these lines are being written, the circumstances under which I first clearly grasped the important generalisation that the laws of Heredity were solely concerned with deviations expressed in statistical units, are vividly recalled in my memory. It was in the grounds of Naworth Castle, where an invitation had been given to ramble freely. A temporary shower drove me to seek refuge in a reddish recess in the rock by the side of the pathway. There the idea flashed across me, and I forgot everything for a moment in my great delight. (p. 300).

An insight of the utmost utility shows us that if the characteristics of interest are measured on a scale that is based on its variability, then the regression coefficient could be applied to these data.

The formula is, of course, the mean of the products of what we now call z scores – the standard scores $r = \sum z_x z_y / n$.

It may be shown that the best estimate of the slope of the regression line is $b = r_{XY} s_Y / s_X$, where s is the sample standard deviation of Y and X , the two variables of interest.

The **multiple linear regression model** is given by $Y' = b_{YX.Z} X + b_{YZ.X} Z + a$, where Y' is termed the *dependent* or *criterion variable* and X and Z are the independent or predictor variables and a is a constant. The b 's are the constants that represent the weight given to the independent (predictor) variables in the estimation (prediction) of Y , the dependent variable. In other words, the regression model may be used to predict the value of a dependent variable from

a set of independent variables. When we have just one dependent and one independent variable, then the slope of the regression line is equivalent to r . For the values of b , we have constants that represent the weights given to the independent variables, and these are calculated on the basis of the partial regression coefficients.

The first use of the word correlation in a statistical context is by Galton in his 1888 paper, *Correlations and their measurement, chiefly from anthropometric data*. Pearson maintains that Galton had first approached the idea of correlation via the use of ranked data before he turned to the measurement of variates (see **Spearman's Rho**). The use of ranks in these kinds of data is usually attributed to **Charles Spearman**, who became the first Professor of Psychology at University College, London. He was, then, for a time a colleague of Pearson's in the same institution, but the two men disliked each other and were critical of each other's work so that a collaboration, that may have been valuable, was never entertained. A primary reason was that Pearson did not relish his approach to correlation that was central to his espousal of eugenics being sullied by methods that did not openly acknowledge the use of variates, essential to the law of ancestral heredity. It can, in fact, be rather easily shown that the modern formula for correlation using ranked data may be derived directly from the product-moment formula.

Spearman first offered the formula for the rank differences. $R = 1 - 3Sd/n^2 - 1$. Here, he uses S for the sum, rather than the modern version Σ and d is the difference in ranks. Later, the formula becomes $r_s = 1 - 6 \sum d^2 / n(n^2 - 1)$. An alternative measure of correlation using ranks was suggested by **Kendall**. This is his *tau* statistic (see **Kendall's Tau** – τ).

If two people are asked to rank the quality of service in four restaurants, the data may be presented thus:

Restaurant	a	b	c	d
Judge 1	3	4	2	1
Judge 2	3	1	4	2

Reordered

Restaurant	d	c	a	b
Judge 1	1	2	3	4
Judge 2	2	4	3	1

What is the degree of correspondence between the judges? We examine the data from Judge 2. Considering the rank of 2 and comparing it with the other ranks, 2 precedes 4, 2 precedes 3, but 2 does not precede 1. These outcomes produce the ‘scores’ +1, +1, and –1. When they are summed, we obtain +1. We proceed to examine each of the possible pairs of ranks and their totals. The maximum possible total obtained if there was perfect agreement between Judge 1 and Judge 2 would be 6. $\tau = (\text{actual total})/(\text{maximum possible total}) = -2/6 = -0.33$. This is a measure of agreement. This index is not identical with that of Spearman, but they both reflect association in the population.

George Udny Yule initially trained as an engineer, but turned to statistics when Pearson offered him a post at University College, London. Although, at first, the two maintained a friendly relationship, this soured when Yule’s own work did not meet with Pearson’s favor. In particular, Yule’s development of a coefficient of association in 2×2 contingency tables created disagreement and bitter controversy between the two men. In general, $X^2 = \sum (f_o - f_e)^2 / f_e$, where f_o is the observed, and f_e the expected, frequency of the observations. In a 2×2 table, this becomes

$$X^2 = (f_o - f_e)^2 \sum \frac{1}{f_e}. \quad (3)$$

When two variables, X and Y , have been reduced to two categories, it is possible to compute the **tetrachoric correlation coefficient**. This measure demands normality of distribution of the continuous variables and a linear relationship. The basic calculation is difficult and approximations to the formula are available. The procedure was just one of a number of methods provided by Pearson, but it lacks reliability and is rarely used nowadays.

The contingency coefficient is also an association method for two sets of attributes (see **Measures of Association**). However, it makes no assumptions

about an underlying continuity in the data and is most suitable for nominal variables. The technique is usually associated with Yule, and this, together with Pearson’s insistence that the variables should be continuous and normally distributed, almost certainly contributed toward the Pearson–Yule disputes.

The correlation technique of Spearman [4] is well known, but his legacy must be his early work on what is now called *factor analysis*. Factor analyses applied to matrices of intercorrelations among observed score variables are techniques that psychology can call its own for they were developed in that discipline, particularly in the context of the measurement of ability.

All the developments discussed here have led us to modern approaches of increasing sophistication. But these approaches have not supplanted the early methods, and correlational techniques produced in the nineteenth century and the later approaches to regression analysis will be popular in the behavioral sciences for a good while yet.

References

- [1] Bravais, A. (1846). Sur les probabilités des erreurs de situation d’un point [on the probability of errors in the position of a point], *Memoirs de l’Académie Royale des Sciences de l’Institut de France* **9**, 255–332.
- [2] Galton, F. (1908). *Memories of My Life*, Methuen, London.
- [3] Pearson K. (1896). Mathematical contributions to the theory of evolution. III. regression, heredity and panmixia in 1896, *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318.
- [4] Spearman, C. (1927). *The Abilities of Man, their Nature and Measurement*, Macmillan Publishers, New York.
- [5] Walker, H.M. (1975). *Studies in the History of Statistical Method*, Arno Press, New York, (facsimile edition from the edition of 1929).

MICHAEL COWLES

History of Discrimination and Clustering

SCOTT L. HERSHBERGER

Volume 2, pp. 840–842

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Discrimination and Clustering

There are two types of problems in classification. In problems of the first type, which are addressed using **discriminant analysis**, we are given the existence of two or more groups and using a sample of individuals from each, the object is to develop a rule under which an individual whose group membership is unknown can be assigned to the correct group. On the other hand, in problems of the second type, which are addressed using cluster analysis (*see* **Cluster Analysis: Overview**), the groups themselves are unknown *a priori* and must be determined by the data so that members within the same group are more similar than those that belong to different groups.

Although discriminant analysis and cluster analysis are distinct procedures, they address complementary classification problems, and thus, are often used together. Such is the case in numerical taxonomy. In numerical taxonomy, the entities to be classified are different animals, and observations on how different animals differ in their characteristics establish a notion of similarity (or dissimilarity) between them. The characteristics chosen by taxonomists vary from morphological attributes (e.g., weight), genetic (e.g., the number of chromosome pairs) to ecological and geographical data describing the habitat of animals. Animals are ‘close’ if their respective mean values on the selected characteristics are similar. Cluster analysis is useful in dealing with the multivariate data required to identify categories of similar animals. In taxonomy, the categories identified by cluster analysis are thought to correspond to natural taxa in the environment, the taxa that comprise the familiar seven-level hierarchy of kingdoms, phyla, classes, orders, families, genera, and species. Once the taxonomic groups have been identified, discriminant analysis can then be used to place an animal into the correct group within each level of the hierarchy.

Nonetheless, historically, discriminant analysis and cluster analysis have been developed independently. Initially, the problems addressed by discriminant analysis were different from what they are today. In the early work on what was then referred to as ‘discriminatory analysis’, classification problems did not involve assigning cases to known groups with

the least amount of error; rather, they involved confirming the distinctiveness of two or more known groups by testing the equality of their distributions [3]. For this reason, statistics for testing the equality of the distributions played an important role, as exemplified by **Pearson’s** ‘coefficient of racial likeness’ [7]. Problems in numerical taxonomy initiated the development of discriminant analysis in its contemporary form, a development undertaken principally by **Fisher** [4]. Fisher was concerned with the univariate classification of observations into one of two groups. For this problem, he suggested a rule that classifies the observation x into the i th population if $|x - \bar{x}_i|$ is the smaller of $|x - \bar{x}_1|$ and $|x - \bar{x}_2|$. For a p -variable observation vector ($p > 1$), Fisher reduced the problem to the univariate one by considering an ‘optimum’ linear combination, called the *discriminant function of the p -variables*. The criterion for defining a discriminant function was to maximize the ratio between the difference in the sample means and the pooled within-groups variance.

Following Fisher, the probabilistic approaches of Welch [16], von Mises [14], and Rao [8] predominated. Summarizing briefly, Welch derived the forms of Bayes rules and the minimax Bayes rule when the groups’ distributions were multivariate normal (*see* **Catalogue of Probability Density Functions**) and their covariance matrices were equal; von Mises specified a rule which maximized the probability of correct classification; and Rao suggested a distance measure between observations and groups, whose minimum value maximized the probability of correctly assigning observations to groups. Rao’s generalized distance measure built upon Pearson’s coefficient of racial likeness and **Mahalanobis**’ [6]. Wald [15] took a decision theoretic approach to discriminant analysis. Lately, nonparametric approaches to discriminant analysis have been a popular area of development [5]. In addition, of historical interest is a recent review [9] of the largely unknown but substantial and important work on discriminant analysis in the former Soviet Union that was initiated by Kolmogorov and his colleagues at Moscow University.

Early forms of cluster analysis included Zubin’s [17] method for sorting a correlation matrix that would yield clusters, and Stephenson’s [11] use of inverted (‘Q’) **factor analysis** to find clusters of personality types (*see* **R & Q Analysis**). However, the first systematic work was performed by Tryon [12], who viewed cluster analysis (‘a poor man’s

factor analysis' according to Tryon) as an alternative to using **factor analysis** for classifying people into types. Most of the methods developed by Tryon were in fact variants of multiple factor analysis [13]. **Cattell** [1], who also emphasized the use of cluster analysis for classifying types of persons, discussed four clustering methods: (a) 'ramifying linkage', which was a variation on what is now termed *single linkage*, (b) a 'matrix diagonal method' which was a graphical procedure, (c) 'Tryon's method' which is related to what currently would be described as average linkage (see **Hierarchical Clustering**), and (d) the 'approximate delimitation method' which was Cattell's extension of the ramifying linkage method. Cattell et al. [2] presented an iterative extension of the ramifying linkage method in order to identify two general classes of types: 'homostats' and 'segregates.' A homostat is a group in which every member has a high degree of resemblance with every other member in the group. On the other hand, a segregate is a group in which each member resembles more members of that group than other groups.

Since the 1960s, interest in cluster analysis has increased considerably, and a large number of different methods for clustering have been proposed. The new interest in cluster analysis was primarily due to two sources: (a) the availability of high-speed computers, and (b) the advocacy of cluster analysis as a method of numerical taxonomy [10]. The introduction of high-speed computers permitted the development of sophisticated cluster analysis methods, methods nearly impossible to carry out by hand. Most of the methods available at the time when high-speed computers first became available required the computation and analysis of an $N \times N$ similarity matrix, where N refers to the number of observations to be clustered. For example, if a sample consisted of 100 observations, this would require the analysis of a 100×100 matrix, which would contain 4950 unique values, hardly an analysis to be undertaken without mechanical assistance.

Cluster analysis appears now to be in a stage of consolidation, in which synthesizing and popularizing currently available methods, rather than introducing new ones, are emphasized. Consolidation is important, if for no other reason than to remove existing discrepancies and ambiguities. For example, the same methods of cluster analysis are often confusingly called by different names. 'Single linkage' is the standard name for a method of hierarchical

agglomerative clustering, but it is also referred to pseudonymously as 'nearest neighbor method', the 'minimum method', the 'space contracting method', 'hierarchical analysis', 'elementary linkage analysis', and the 'connectedness method'.

References

- [1] Cattell, R.B. (1944). A note on correlation clusters and cluster search methods, *Psychometrika* **9**, 169–184.
- [2] Cattell, R.B., Coulter, M.A. & Tsuijoka, B. (1966). The taxonomic recognition of types and functional emergents, in *Handbook of Multivariate Experimental Psychology*, R.B. Cattell, ed., Rand-McNally, Chicago, pp. 288–329.
- [3] Das Gupta, S. (1974). Theories and methods of discriminant analysis: a review, in *Discriminant Analysis and Applications*, T. Cacoullos, ed., Academic Press, New York, pp. 77–138.
- [4] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.
- [5] Huberty, C.J. (1994). *Applied Discriminant Analysis*, John Wiley & Sons, New York.
- [6] Mahalanobis, P.C. (1930). On tests and measurements of group divergence, *The Journal and Proceedings of the Asiatic Society of Bengal* **26**, 541–588.
- [7] Pearson, K. (1926). On the coefficient of racial likeness, *Biometrika* **13**, 247–251.
- [8] Rao, C.R. (1947). A statistical criterion to determine the group to which an individual belongs, *Nature* **160**, 835–836.
- [9] Raudys, S. & Young, D.M. (2004). Results in statistical discriminant analysis: a review of the former Soviet Union literature, *Journal of Multivariate Analysis* **89**, 1–35.
- [10] Sokal, R.R. & Sneath, P. (1963). *Principles of Numerical Taxonomy*, Freeman, San Francisco.
- [11] Stephenson, W. (1936). Introduction of inverted factor analysis with some applications to studies in orexia, *Journal of Educational Psychology* **5**, 553–567.
- [12] Tryon, R. (1939). *Cluster Analysis*, McGraw-Hill, New York.
- [13] Tryon, R. & Bailey, D.E. (1970). *Cluster Analysis*, McGraw-Hill, New York.
- [14] von Mises, R. (1944). On the classification of observation data into distinct groups, *Annals of Mathematical Statistics* **16**, 68–73.
- [15] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups, *Annals of Mathematical Statistics* **15**, 145–162.
- [16] Welch, B.L. (1939). Note on discriminant functions, *Biometrika* **31**, 218–220.
- [17] Zubin, J.A. (1938). A technique for measuring likemindedness, *Journal of Abnormal Psychology* **33**, 508–516.

SCOTT L. HERSHBERGER

History of Factor Analysis: A Psychological Perspective

ROBERT M. THORNDIKE

Volume 2, pp. 842–851

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Factor Analysis: A Psychological Perspective

In 1904, **Charles Spearman** published two related papers that have had an immense influence on psychology in general and psychometrics in particular. In the first, [23] he outlined the classical or true-score model of reliability, dividing test performance, and therefore the variance in test scores, into a portion that was due to the individual's 'true' level on the trait in question and a portion that was random error of measurement. This can be represented symbolically as

$$X_{ij} = T_i + e_{ij} \quad (1)$$

The observed score of individual i on variable X at occasion j (X_{ij}) is composed of the true score of individual i on X (T_i) plus the error made in the measurement of person i at time j (e_{ij}). If T_i is constant, variation in a person's performance on successive measurements is due to randomly fluctuating errors. This distinction has formed the cornerstone of classical measurement theory (see **Measurement: Overview**) and is still highly influential.

By applying (1) to the scores for a sample on N individuals and finding the variance, we can break the variance in the observed scores down into two components, the variance in true scores, and the variance of the errors of measurement.

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \quad (2)$$

The ratio of true-score variance to total variance yields the reliability coefficient, and the square root of the error variance is the standard error of measurement, which can be used to determine an interval of uncertainty for a predicted score.

In his second paper, Spearman [24] enunciated one of the most influential theories of human cognitive abilities of the twentieth century, his theory of general intelligence, and laid the foundations for the method of data analysis that has come to be known as **factor analysis**. In this paper, Spearman divided the score of a person on an observed variable into a portion that represented what that variable had in common with the other variables in the analysis, which he called g or general intelligence, and a

portion that was unique to the variable in question, which he called s or specific performance.

$$X_{ij} = g_i + s_{ij} \quad (3)$$

The score of individual i on variable j (X_{ij}) is composed of person i 's score on the general ability variable g (g_i) plus the individual's score on the specific part of X_j (s_{ij}). Applying the logic of (2) to a set of scores yields the conclusion that the variance of a variable can be decomposed into a portion that is due to the common factor and another portion that is due to the specific factor.

$$\sigma_X^2 = \sigma_g^2 + \sigma_s^2 \quad (4)$$

Because there were two sources, or factors, contributing to the variance of each variable, this theory came to be known as the *two-factor theory*.

Three years earlier **Karl Pearson** [20] had derived what he called the **principal component** of a set of variables to account for the largest amount of variance explainable by a single dimension of the set (later generalized by **Hotelling** [10] to provide the full set of principal components). Because this procedure was not associated with a psychological theory and was computationally demanding, it did not get much attention from psychologists at the time.

There is an important ontological difference between component analysis as conceived by Pearson and Hotelling and factor analysis as conceived by Spearman. Component analysis is properly viewed as a data-reduction procedure. It results in an orthogonal (uncorrelated) representation of the variable space, but implies nothing about constructs underlying the variables. Factor analysis, on the other hand, has been viewed from its inception as a method for uncovering meaningful causal constructs to account for the correlations between variables. Some writers, for example, Velicer and Jackson [38], have argued that the distinction is unnecessary, and in one sense they are right. One should get a similar description of the data from either approach. However, as we shall see, the common factor approach generally yields better results in terms of one important index of the quality of the solution, ability to reproduce the original data.

Spearman's initial proposal of a single general factor of cognitive ability sparked an immediate trans-Atlantic debate between Spearman and E. L. Thorndike [27], who argued that there were many

2 History of Factor Analysis: A Psychological Perspective

factors of ‘intellect’ (his preferred term; see [29] for a description of the debate). In the face of this criticism, Spearman was forced to develop an analytic method to support his claim that a single factor was sufficient to account for the correlations among a set of tests. He was able to show [8] that a sufficient condition for the existence of a single factor was that an equation of the form

$$r_{ab}r_{cd} - r_{ac}r_{bd} = 0 \quad (5)$$

be satisfied for all possible sets of four tests. This criterion, known as the tetrad difference equation, would not be exactly satisfied for all possible sets of four tests with real data, but it might be approximated.

Debate over the nature of intelligence continued as one side produced a set of data satisfying the tetrad criterion and the other side countered with one that did not. Then, in 1917, **Cyril Burt** [2] offered a method for extracting a factor from a matrix of correlations that approximated Pearson’s principal component, but at great computational savings. Because his method placed the first factor through the average or geometric center of the set of variables, it became known as the centroid method for extracting factors (determining the initial location of a factor is called *factor extraction*). The centroid method was computationally straightforward and yielded useful factors. In the hands of **L. L. Thurstone**, it would become the standard method of factor extraction until computers became widely available in the late 1950s.

Although Spearman continued to offer his tetrad criterion as providing evidence of a single general factor of intelligence [25], the two-factor theory was dealt a serious blow in 1928 by Truman Kelley [17]. Using the method of partial correlation to remove g from the matrix of correlations among a set of ability variables, Kelley showed that additional meaningful factors could be found in the matrix of residual correlations. He argued that the distribution of residual correlations after extracting g could be used to test (and reject) the hypothesis of a single general factor and that an important goal for psychological measurement should be to construct tests that were pure measures of the multiple factors that he had found. Somewhat earlier, Thompson [26] had proposed a sampling approach to the conceptualization of factors that resulted logically in a hierarchy of factors depending on the breadth of the sampling. The concept of a hierarchy was later explicitly developed by

Vernon [39] and Humphreys [11] into general theories about the organization of human abilities.

Enter **L. L. Thurstone**, the most important single contributor to the development of factor analysis after Spearman himself. In 1931, Thurstone [30] published an important insight. He recognized that satisfying the tetrad criterion for any set of four variables was equivalent to saying that the rank of the 4×4 correlation matrix was 1. (We can roughly define the rank of a matrix as the number of independent dimensions it represents. More formal definitions require a knowledge of matrix algebra.) In this important paper, Thurstone argued that the rank of a matrix is the equivalent of the number of factors required to account for the correlations. Unless the rank of a matrix was 1, it would require more than one factor to reproduce the correlations (see below). He also showed how the centroid method could be used to extract successive factors much more simply and satisfactorily than Kelley’s partial correlation procedure.

Through the remainder of the 1930s, Thurstone continued to expand his conception of common factor analysis. He undertook a massive study of mental abilities, known as the Primary Mental Abilities study, in which 240 college-student volunteers took a 15-hour battery of 56 tests [31–34]. From analysis of this battery, he identified as many as 12 factors, seven of which were sufficiently well defined to be named as *scientific constructs of ability*. In addition, he developed the geometric interpretation of factors as the axes of a multidimensional space defined by the variables. This insight allowed him to recognize that the location of any factor is arbitrary. Once the multidimensional space (whose dimensionality is defined by the rank of the correlation matrix) is defined by the variables, centroid factors or principal components are used to define the nonzero axes of the space by satisfying certain conditions (see below), but these initial factors seldom seemed meaningful. Thurstone argued that one could (and should) move the axes to new positions that had the greatest psychological meaning. This process was called *factor rotation* (see **Factor Analysis: Exploratory**).

In his original work, Thurstone rotated the factors rigidly, maintaining their orthogonal or uncorrelated character. By 1938, he was advocating allowing the factors to become correlated or oblique. Geometrically, this means allowing the factors to assume

positions at other than 90 degrees to each other. Others, such as Vernon [39] and Humphreys [11] would later apply factor analysis to the matrices of correlations among the ‘first-order factors’ to obtain their hierarchical models.

Thurstone’s insights created three significant problems. First, the actual rank of any proper correlation matrix would always be equal to the number of variables because the diagonal entries in the matrix included not only common variance (the *g*-related variance of Spearman’s two-factor model) but also the specific variance. Thurstone suggested that the correlation matrix to be explained by the factors should not be the original matrix but one in which an estimate of the common variance of each variable had been placed in the appropriate location in the diagonal. This left investigators with the problem of how to estimate the common variance (or communality, as it came to be known).

The problem of estimating the communality was intimately related to the problem of how many factors were needed to account for the correlations. More factors would always result in higher communalities. In an era of hand computation, one did not want to extract factors more than once, so good communality estimates and a correct decision on the number of factors were crucial. Thurstone himself tended to favor using the largest correlation that a variable had with any other variable in the matrix as an estimate of the communality. Roff [22] argued that the squared multiple correlation of each variable with the other variables in the matrix provided the best estimate of the communality, and this is a starting point commonly used today. Others suggested an estimate of the reliability of each variable provided the best communality estimate.

The third problem resulted from the practice of rotation. The criteria for factor extraction provided a defined solution for the factors, but once rotation was introduced, there were an infinite number of equally acceptable answers. Thurstone attempted to solve this problem with the introduction of the concept of simple structure. In its most rudimentary form, the principle of simple structure says that each observed variable should be composed of the smallest possible number of factors, ideally one. In his most comprehensive statement on factor analysis, Thurstone [35, p. 335] offered five criteria that a pattern of factor loadings should meet to qualify as satisfying the simple structure principle, but most

attention has been directed to finding a rotation that produces a small number of nonzero loadings for any variable.

There are two primary arguments in favor of factor patterns that satisfy simple structure. First, they are likely to be the most interpretable and meaningful. A strong argument can be made that meaningfulness is really the most important property for the results of a factor analysis to have. Second, Thurstone argued that a real simple structure would be robust across samples and with respect to the exact selection of variables. He argued convincingly that one could hardly claim to have discovered a useful scientific construct unless it would reliably appear in data sets designed to reveal it.

Thurstone always did his rotations graphically by inspection of a plot of the variables and a pair of factors. However, this approach was criticized as lacking objectivity. With the advent of computers in the 1950s, several researchers offered objective rotation programs that optimized a numerical function of the factor loadings [e.g., 3, 19]. The most successful of these in terms of widespread usage has been the varimax criterion for rotation to an orthogonal simple structure proposed by Kaiser [14], although the direct oblimin procedure of Jennrich and Sampson [12] is also very popular as a way to obtain an oblique simple structure.

In addition to making analytic rotation possible, the rise of computers also sounded the death knell for centroid extraction. By the late 1960s the Pearson–Hotelling method of principal axis factor extraction had replaced all others. Several alternatives had been offered for how to estimate the communalities, including maximum likelihood [13, 18], alpha [16] and minimum residuals [7], but all employed the same basic extraction strategy that is described below.

There was also progress on the number-of-factors question that can be traced to the availability of computers. Although Hotelling [10] and Bartlett [1] had provided tests of the statistical significance of principal components (Bartlett’s sphericity test is still an option in SPSS), neither was used until computers were available because they did not apply to centroid factors. Rippe [21] offered a general test for the number of factors in large samples, and Lawley [18] had provided the foundation of a significance test for use with maximum likelihood factors. Others,

notably Kaiser [15] and **Cattell** [4] offered nonstatistical rules of thumb for the number of principal components to retain for rotation. Kaiser's criterion held that only factors that have **eigenvalues** (see below) greater than 1.0 should be considered, and Cattell suggested that investigators examine the plot of the eigenvalues to determine where a 'scree' (random noise factors) began. Kaiser's criterion became so popular that it is the default in SPSS and some other computer programs, and many programs will output the plot of the eigenvalues as an option.

Statistical criteria for the number of factors were criticized as being highly sensitive to sample size. On the other hand, one person's scree is another person's substantive factor, and Kaiser's criterion, although objective, could result in keeping a factor with an eigenvalue of 1.0001 and dropping one at 0.999. To solve these problems, Horn [9] proposed that in a study with m variables, $m \times m$ matrices of correlations from random data be analyzed and only factors from the real data with eigenvalues larger than the paired eigenvalue from random data be kept. This approach has worked well in simulation studies, but has not seen widespread application. A method with similar logic by Velicer [36] based on average squared partial correlations has also shown promise but seen little application.

By the early 1970s, the development of common factor analysis was all but complete. That this is so can be inferred from the fact that there has not been a major book devoted to the subject since 1983 [5], while before that date several important treatments appeared every decade. This does not mean that the method has been abandoned. Far from it; unrestricted (exploratory) factor analysis remains one of the most popular data analytic methods. Rather, work has focused on technical issues such as rules for the number of factors to extract, how large samples need to be, and how many variables need to be included to represent each factor. Although many investigators have contributed to developments on these topics, Wayne Velicer and his associates have been among the most frequent and influential contributors [37].

Overview of Factor Analysis

There are two basic ways to conceptualize factor analysis, an algebraic approach and a graphic or geometric approach. In this section, we will review

each of these briefly. For a further description, see the entry on common factor analysis. A thorough description of both approaches can also be found in Harman [6] or Gorsuch [5].

Algebraic Approach

Spearman [24] and Thurstone [35] both considered factors to represent real latent causal variables that were responsible for individual differences in test scores. Individuals are viewed as having levels of ability or personality on whatever traits the factors represent. The task for factor analysis is to determine from the correlations among the variables how much each factor contributes to scores on each variable. We can therefore think of a series of regression equations with the factors as predictors and the observed variables as the criteria. If there are K factors and p observed variables, we will have p regression equations, each with the same K predictors, but the predictors will have different weights in each equation reflecting their individual contributions to that variable.

Suppose we have a set of six variables, three measures of verbal ability and three measures of quantitative ability. We might expect there to be two factors in such a set of data. Using a generalization of Spearman's two-factor equation, we could then think of the score of a person (call him i for Ishmael) on the first test (X_{1i}) as being composed of some part of Ishmael's score on factor 1 plus some part of his score on factor 2, plus a portion specific to this test. For convenience, we will put everything in standard score form.

$$Z_{X_{1i}} = \beta_{X_1 F_1} Z_{F_{1i}} + \beta_{X_1 F_2} Z_{F_{2i}} + U_{X_{1i}} \quad (6)$$

Ishmael's score on variable X_1 ($Z_{X_{1i}}$) is composed of the contribution factor 1 makes to variable X_1 ($\beta_{X_1 F_1}$) multiplied by Ishmael's score on factor 1 ($Z_{F_{1i}}$) plus the contribution factor 2 makes to X_1 ($\beta_{X_1 F_2}$) times Ishmael's score on factor 2 ($Z_{F_{2i}}$) plus the residual or unique part of the score, $U_{X_{1i}}$. U is whatever is not contributed by the common factors and is called *uniqueness*. We shall see shortly that uniqueness is composed of two additional parts. Likewise, Ishmael's scores on each of the other variables are composed of a contribution from each of the factors plus a unique part. For example,

$$Z_{X_{2i}} = \beta_{F_1 X_2} Z_{F_{1i}} + \beta_{F_2 X_2} Z_{F_{2i}} + U_{X_{2i}} \quad (7)$$

If we have scores on variable 1 for a set of people, we can use (6) to see that factor analysis decomposes the variance of these scores into contributions by each of the factors. That is, for each variable X_j we can develop an expression of the following form

$$\sigma_{X_j}^2 = \beta_{X_j F_1} \sigma_{F_1}^2 + \beta_{X_j F_2} \sigma_{F_2}^2 + \sigma_{U_j}^2 \quad (8)$$

The variance of each observed variable is a weighted combination of the factor variances plus a unique contribution due to that variable.

The betas are known as *factor pattern coefficients* or *factor loadings*. As is the case in multiple correlations generally, if the predictors (factor) are uncorrelated, the regression weights are equal to the predictor-criterion correlations. That is, for orthogonal factors, the pattern coefficients are also the correlations between the factors and the observed variables. In factor analysis, the correlations between the variables and the factors are called *factor structure coefficients*. One of the major arguments that has been made in favor of orthogonal rotations of the factors is that as long as the factors are orthogonal the equivalence between the pattern and structure coefficients is maintained, so interpretation of the results is simplified.

Let us consider an example like the one above. Table 1 contains hypothetical correlations among three verbal and three quantitative tests from the Stanford–Binet Fourth Edition [28]. The matrix was constructed to be similar to the results obtained with the actual instrument.

Applying principal axis factor analysis (see the entry on common factor analysis) to this matrix yields the factor matrix in Table 2. This matrix is fairly typical of results from factoring sets of ability variables. There is a large first factor with all positive loadings and a smaller second factor with about half positive and half negative loadings (called a *bipolar*

Table 1 Hypothetical correlations between six subtests of the Stanford–Binet, Fourth Edition

Variable	1	2	3	4	5	6
Vocabulary	1.000					
Comprehension	0.710	1.000				
Absurdities	0.586	0.586	1.000			
Equation building	0.504	0.460	0.330	1.000		
Number series	0.562	0.563	0.522	0.630	1.000	
Quantitative	0.570	0.567	0.491	0.594	0.634	1.000

Table 2 Initial factor matrix for six Stanford–Binet Tests

	Factor		h^2
	1	2	
Vocabulary	0.80	−0.20	0.68
Comprehension	0.79	−0.26	0.69
Absurdities	0.67	−0.26	0.52
Equation building	0.71	0.43	0.69
Number series	0.78	0.18	0.64
Quantitative	0.76	0.14	0.60
Factor variances	3.40	0.41	3.81

factor). The large first factor has often been equated with Spearman's g factor.

One interpretation of a correlation is that its square corresponds to the proportion of variance in one variable that is accounted for by the other variable. For orthogonal factors, this means that a squared factor loading is the proportion of the variable's variance contributed by that factor. Summed across all common factors, the result is the proportion of the variable's variance that is accounted for by the set of common factors (note that uniqueness is not included). This quantity is called the *common variance* of the variable or its *communality*. For the case of K factors,

$$\text{Communality of } X_1 = \sum_{j=1}^K \beta_{X_1 F_j}^2 = h_{X_1}^2 \quad (9)$$

The communality of each variable is given in the last column of Table 1. The symbol h^2 is often used for communality and represents a variance term.

The remainder of each variable's variance, $(1 - h^2)$, is the variance unique to that variable, its uniqueness (symbolized u^2 , also a variance term). The unique variance is composed of two parts, variance that is due to reliable individual differences that are not accounted for by the common factors and random errors of measurement. The first is called *specificity* (symbolized s^2) and the second is simply error (e^2). Thus, Spearman's two 1904 papers lead to the following way to view a person's score on a variable

$$Z_{X_j i} = \beta_{F_1 X_j} Z_{F_{1i}} + \beta_{F_2 X_j} Z_{F_{2i}} + s_{X_j i} + e_{X_j i} \quad (10)$$

and link common factor theory with measurement theory. If we once again think of the scores for N people on the set of variables, the variance of each

6 History of Factor Analysis: A Psychological Perspective

variable (we are still considering standard scores, so each variable's total variance is 1.0) can be viewed in three interlocking ways (each letter corresponds to a kind of variance derived from (10)).

$$\begin{array}{lll} \text{Factor theory} & 1.0 = h^2 + u^2 & u^2 = s^2 + e^2 \\ & 1.0 = h^2 + s^2 + e^2 & r^2 = h^2 + s^2 \\ \text{Measurement} & 1.0 = r^2 + e^2 & \\ \text{theory} & & \end{array}$$

The symbol r^2 is used to indicate the reliable variance in test scores.

We can also consider the factor loading as revealing the amount of a factor's variance that is contributed by each variable. Again taking the squared factor loadings, but this time summing down the column of each factor, we get the values at the bottom of Table 2. These values are often referred to, somewhat inappropriately, as eigenvalues. This term is really only appropriate in the case of principal components. In this example, we used squared multiple correlations as initial communality estimates, so *factor variance* is the correct term to use. Note that the first factor accounts for over half of the variance of the set of six variables and the two factors combined account for about 2/3 of the variance.

Now, let us see what happens if we apply varimax rotation to these factors. What we would expect for a simple structure is for some of the loadings on the first factor to become small, while some of the loadings on the second factor become larger. The results are shown in Table 3. The first factor now has large loadings for the three verbal tests and modest loadings for the three quantitative tests and the reverse pattern is shown on the second factor. We would be inclined to call factor 1 a verbal ability factor and factor 2 a quantitative ability factor. Note

that the factors still account for the same amount of each variable's variance, but that variance has been redistributed between the factors. That is, the communalities are unchanged by rotation, but the factor variances are now more nearly equal.

There are two things about the varimax factor matrix that might cause us concern. First, the small loadings are not that small. The structure is not that simple. Second, there is no particular reason why we would or should expect the factors to be orthogonal in nature. We will allow the data to speak to us more clearly if we permit the factors to become correlated. If they remain orthogonal with the restriction of orthogonality relaxed, so be it, but we might not want to force this property on them. Table 4 contains the factor pattern matrix after rotation by direct oblimin.

There are two things to notice about these pattern coefficients. First, the large or primary coefficients display the same basic pattern and size as the coefficients in Table 3. Second, the secondary loadings are quite a lot smaller. This is the usual result of an oblique rotation. The other important statistic to note is that the factors in this solution are correlated 0.70. That is, according to these data, verbal ability and quantitative ability are quite highly correlated. This makes sense when we observe that the smallest correlation between a verbal test and a quantitative test in Table 1 is 0.33 and most are above 0.50. It is also what Spearman's theory would have predicted.

We can note one final feature of this analysis, which addresses the question of whether there is a difference between principal components analysis and common factor analysis. The factors provide a model for the original data and we can ask how well the model fits the data. We can reproduce the

Table 3 Varimax-rotated factor matrix for six Stanford-Binet Tests

	Factor		h^2
	1	2	
Vocabulary	0.73	0.38	0.68
Comprehension	0.76	0.34	0.69
Absurdities	0.68	0.26	0.52
Equation building	0.24	0.79	0.69
Number series	0.46	0.66	0.64
Quantitative	0.48	0.61	0.60
Factor variances	2.07	1.76	3.81

Table 4 Pattern matrix from a direct oblimin rotation for six Stanford-Binet Tests

	Factor ^a	
	1	2
Vocabulary	0.76	0.10
Comprehension	0.82	0.02
Absurdities	0.75	-0.04
Equation building	-0.09	0.89
Number series	0.27	0.59
Quantitative	0.30	0.53

^aFactors correlate +0.70.

original correlations, as accounted for by the factors, by multiplying the factor matrix by its transpose (most factor programs will give you this output if you ask for it). If the model fits well, the difference between the original correlations and the correlations as reproduced from the factor model should be similar (the difference is called the *residual*). Applying this test to the factors in Table 2, we find that all of the residuals with the correlations in Table 1 are less than .05, indicating quite good fit. If we had applied a principal components analysis to the same data, over half of the residual correlations would exceed .05, a much poorer fit. The lack of fit for principal components is a result of including unique variance in the correlation matrix; the reproduced correlations will be inflated.

Geometric Approach

The foundation of the geometric approach to factor analysis rests on the fact that variables can be represented by lines in space. The correlation between any pair of variables is directly related to the cosine of the angle between the lines representing the two variables; a right angle means $r = 0$, and a small angle means the correlation is large. Thus, highly correlated variables lie close to each other in space. Thurstone rotated factors by placing them close to clusters of variables in such a graphic display.

A proper correlation matrix will require as many dimensions as there are variables to completely represent the data. Our small six-variable example would require a six-dimensional space. However, we can make a plot showing the model of the data represented by any pair of factors by simply laying out the factors as axes of the space and plotting the variables as lines given by the factor loadings. The tip of each line is defined by the variable's factor loadings. Figure 1 contains the plot of the factor matrix from Table 2.

The variables form a fan-shaped array around the positive end of the first factor with the quantitative variables on the positive side of factor 2 and the verbal variables on the negative side of this factor. The factor matrix in Table 2 and the graphic display in Figure 1 give identical representations of the relations among the variables. The square of the length of the line representing each variable is equal to its communality. (From the Pythagorean theorem, $c^2 = a^2 + b^2$. The line representing a variable is c ,

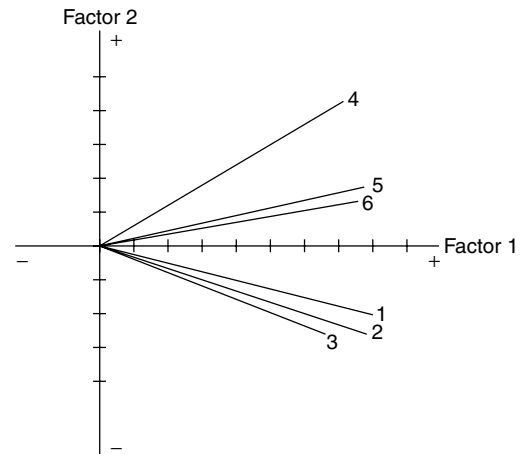


Figure 1 Plot of the six variables and first two factors before rotation. Note that all loadings on the first factor are positive and the second factor is bipolar

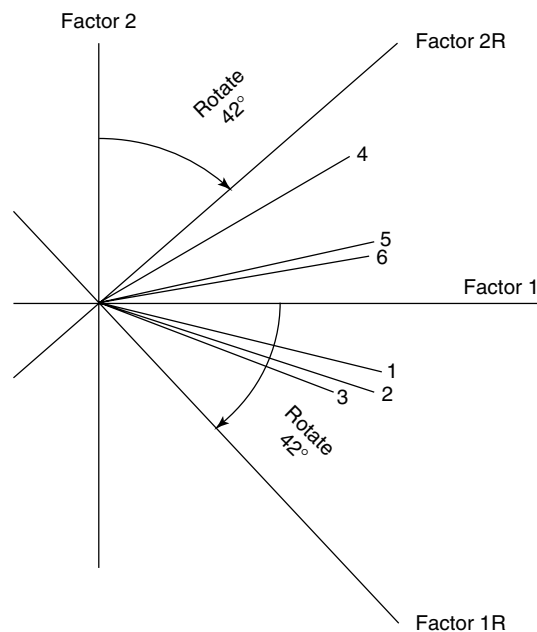


Figure 2 Plot of the six variables and first two factors in which the factors have been rotated orthogonally 42 degrees clockwise

and a and b are the factor loadings, so this is the geometric equivalent of (9).)

Now consider the rotation concept. Factor 1 is a complex combination of all the variables while factor

2 contrasts verbal and quantitative tests. If we rotate the factors clockwise, factor 1 will come to represent the verbal tests more clearly and factor 2 will align with the quantitative ones. It looks like a rotation of about 45 degrees will do the trick.

When we apply the varimax rotation criterion, a rotation of 42 degrees produces an optimum solution. The plot of the rotated solution is shown in Figure 2. Notice that the variables stay in the same place and the factors rotate to new locations. Now, all of the variables project toward the positive ends of both factors, and this fact is reflected by the uniformly positive loadings in Table 3.

Figure 3 is a plot of the direct oblimin rotation from Table 4. Here we can see that the two factors have been placed near the centers of the two clusters of variables. The verbal cluster is a relatively pure representation of the verbal factor (1). None of the variables are far from the factor and all of their pattern coefficients on factor 2 are essentially zero. Equation building is a relatively pure measure of the quantitative factor, but two of the quantitative variables seem to also involve some elements of verbal behavior. We can account for this fact in the case of the Quantitative test because it is composed

in part of word problems that might involve a verbal component. The reason for the nonzero coefficient for Number Series is not clear from the test content.

The algebraic and graphical representations of the factors complement each other for factor interpretation because they provide two different ways to view exactly the same outcome. Either one allows us to formulate hypotheses about causal constructs that underlie and explain a set of observed variables. As Thurstone [35] pointed out many years ago, however, this is only a starting point. The scientific value of the constructs so discovered must be tested in additional studies to demonstrate both their stability with respect to the specific selection of variables and their generality across subject populations. Often they may be included in studies involving experimental manipulations to test whether they behave as predicted by theory.

References

- [1] Bartlett, M.S. (1950). Tests of significance in factor analysis, *British Journal of Psychology, Statistical Section* **3**, 77–85.
- [2] Burt, C. (1917). *The Distributions and Relations of Educational Abilities*, London County Council, London.
- [3] Carroll, J.B. (1953). Approximating simple structure in factor analysis, *Psychometrika* **18**, 23–38.
- [4] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.
- [5] Gorsuch, R.L. (1983). *Factor Analysis*, Lawrence Erlbaum, Mahwah.
- [6] Harman, H.H. (1976). *Modern Factor Analysis*, University of Chicago Press, Chicago.
- [7] Harman, H.H. & Jones, W.H. (1966). Factor analysis by minimizing residuals (Minres), *Psychometrika* **31**, 351–368.
- [8] Hart, B. & Spearman, C. (1912). General ability, its existence and nature, *British Journal of Psychology* **5**, 51–84.
- [9] Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis, *Psychometrika* **30**, 179–185.
- [10] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441.
- [11] Humphreys, L.G. (1962). The organization of human abilities, *American Psychologist* **17**, 475–483.
- [12] Jennrich, R.I. & Sampson, P.F. (1966). Rotation for simple loadings, *Psychometrika* **31**, 313–323.
- [13] Joreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis, *Psychometrika* **32**, 443–482.
- [14] Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

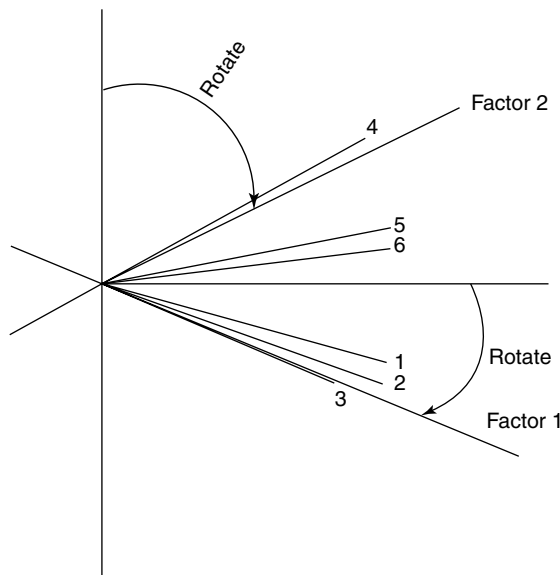


Figure 3 Plot of the six variables and first two factors after a direct Oblimin rotation. Note that the positions of the factors are the same as in Figures 1 and 2, but the factors have been rotated by different amounts

- [15] Kaiser, H.F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement* **20**, 141–151.
- [16] Kaiser, H.F. & Caffrey, J. (1965). Alpha factor analysis, *Psychometrika* **30**, 1–14.
- [17] Kelley, T.L. (1928). *Crossroads in the Mind of Man*, Stanford University Press, Stanford.
- [18] Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood, *Proceedings of the Royal Society of Edinburgh* **60**, 64–82.
- [19] Neuhaus, J.O. & Wrigley, C. (1954). The quartimax method: an analytic approach to orthogonal simple structure, *British Journal of Statistical Psychology* **7**, 81–91.
- [20] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine, Series B* **2**, 559–572.
- [21] Rippe, D.D. (1953). Application of a large sampling criterion to some sampling problems in factor analysis, *Psychometrika* **18**, 191–205.
- [22] Roff, M. (1936). Some properties of the communality in multiple factor theory, *Psychometrika* **1**, 1–6.
- [23] Spearman, C. (1904a). The proof and measurement of the association between two things, *American Journal of Psychology* **15**, 72–101.
- [24] Spearman, C. (1904b). “General intelligence,” objectively determined and measured, *American Journal of Psychology* **15**, 201–293.
- [25] Spearman, C. (1927). *The Abilities of Man*, Macmillan Publishing, New York.
- [26] Thomson, G.H. (1920). General versus group factors in mental activities, *Psychological Review* **27**, 173–190.
- [27] Thorndike, E.L., Lay, W. & Dean, P.R. (1909). The relation of accuracy in sensory discrimination to general intelligence, *American Journal of Psychology* **20**, 364–369.
- [28] Thorndike, R.L., Hagen, E.P. & Sattler, J.M. (1986). The Stanford-Binet intelligence scale, 4th Edition, *Technical Manual*, Riverside, Itasca.
- [29] Thorndike, R.M. (1990). *A Century of Ability Testing*, Riverside, Itasca.
- [30] Thurstone, L.L. (1931). Multiple factor analysis, *Psychological Review* **38**, 406–427.
- [31] Thurstone, L.L. (1935). *The Vectors of Mind*, University of Chicago Press, Chicago.
- [32] Thurstone, L.L. (1936a). A new conception of intelligence and a new method of measuring primary abilities, *Educational Record* **17**(Suppl. 10), 124–138.
- [33] Thurstone, L.L. (1936b). A new conception of intelligence, *Educational Record* **17**, 441–450.
- [34] Thurstone, L.L. (1938). *Primary Mental Abilities*, Psychometric Monographs No. 1.
- [35] Thurstone, L.L. (1947). *Multiple Factor Analysis*, University of Chicago Press, Chicago.
- [36] Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations, *Psychometrika* **41**, 321–327.
- [37] Velicer, W.F. & Fava, J.L. (1998). The effects of variable and subject sampling on factor pattern recovery, *Psychological Methods* **3**, 231–251.
- [38] Velicer, W.F. & Jackson, D.N. (1990). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure, *Multivariate Behavioral Research* **25**, 1–28.
- [39] Vernon, P.E. (1950). *The Structure of Human Abilities*, Wiley, New York.

(See also **Factor Analysis: Confirmatory; Factor Analysis: Multitrait–Multimethod; Factor Analysis of Personality Measures**)

ROBERT M. THORNDIKE

History of Factor Analysis: A Statistical Perspective

DAVID J. BARTHOLOMEW

Volume 2, pp. 851–858

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Factor Analysis: A Statistical Perspective

Origins

Factor analysis is usually dated from **Charles Spearman's** paper '*General Intelligence' Objectively Determined and Measured* published in the *American Journal of Psychology* in 1904 [18]. However, like most innovations, traces of the idea can be found in earlier work by **Karl Pearson** [17] and others. All the same, it was a remarkable idea. Spearman, of course, did not invent factor analysis in the full glory of its later development. He actually proposed what would now be called a *one-factor* model though then it was, perversely, called a *two-factor model*. It arose in the context of the theory of correlation and partial correlation, which was one of the few topics in multivariate statistics that was reasonably well developed at that time. Technically speaking, it was not such a great step forward but it proved enough to unlock the door to a huge field of applications.

Spearman and most of his immediate followers were interested in measuring human abilities and, in particular, general intelligence. There was no interest in developing the general method of **multivariate analysis** which factor analysis later became. Factor analysis is unusual among multivariate statistical techniques in that it was developed almost entirely within the discipline of psychology. Its line of development was therefore subservient to the needs of psychological measurement of abilities in particular. This has had advantages and disadvantages. On the positive side, it has 'earthed' or grounded the subject, ensuring that it did not wander off into theoretical irrelevancies. Negatively, it had a distorting effect that emphasized some aspects and ignored others.

Returning to Spearman and the origins of factor analysis; the theory quickly grew. Sir **Cyril Burt**, see for example [5], was one of the first on the scene and, with his access to large amounts of data from the London County Council, was able to press ahead with practical applications.

The Key Idea

The key idea was that it might be possible to explain the correlations in sets of observable variables by the hypothesis that they all had some dependence on a common factor (or, later, factors). The fact that, in practice, the correlations were not wholly accounted for in this way was explained by the influence of other variables 'specific' to each observable variable. If this hypothesis were correct, then conditioning on the common variables (factors) should render the variables independent. In that sense, their correlations were 'explained' by the common factors. It was then but a short step to show that the variances of each variable could be partitioned into two parts, one arising from the common factor(s) and the other from the rest. The importance of each variable (its 'saturation' with the common factor) could be measured by its correlation with that factor – and this could be estimated from the observed correlations. In essence, this was achieved by Spearman in 1904.

In 1904, there was little statistical theory available to help Spearman but what there was proved to be enough. Correlation had been a major field of study. The invention of the product-moment correlation (see **Pearson Product Moment Correlation**) had been followed by expressions for **partial correlations**. A first-order partial correlation gives the correlation between a pair of variables when a third is fixed. Second-order coefficients deal with the case when two other variables are fixed, and so on. The expressions for the partial correlations presupposed that the relationships between the variables were linear. That was because product-moment correlation is a measure of linear correlation. Inspection of early editions of Yule's *Introduction to the Theory of Statistics* (starting with [21]) will show how prominent a place partial correlation occupied in the early days. Later, the emphasis shifted to multiple regression (see **Multiple Linear Regression**), which offered an alternative way of investigating the same phenomenon.

The result of Spearman's idea is that if the correlation between two variables is due to their common dependence on a third variable, then one can deduce that the form of the correlations has a particularly simple form. It is not entirely clear from the 1904 paper how Spearman went about this or what form of the relationship among the correlations

2 History of Factor Analysis: A Statistical Perspective

he actually used, but a simple way of arriving at his result is as follows.

Suppose we have a set of variables correlated among themselves. We suspect that these correlations are induced by their common dependence on a factor called G (Spearman used G in this context because he was using it to denote general intelligence). If that is the case, then conditioning on G should remove the correlation. Consider two variables i and j with correlation r_{ij} . If our hypothesis is correct, that correlation should vanish if we condition on G . That is, the partial correlation $r_{ij.G}$ should be zero (the ‘dot’ is used to denote ‘given’). Now,

$$r_{ij.G} = \frac{r_{ij} - r_{iG}r_{jG}}{\sqrt{1 - r_{iG}^2}\sqrt{1 - r_{jG}^2}}, \quad (1)$$

and so the necessary and sufficient condition for r to vanish is that

$$r_{ij} = r_{iG}r_{jG} \quad (i, j = 1, 2, \dots, p) \quad (i \neq j). \quad (2)$$

If we can find values r_{iG} ($i = 1, 2, \dots, p$) to satisfy these relations (approximately), then we shall have established that the mutual correlation among the variables can, indeed, be explained by their common dependence on the common factor G . This derivation shows that what came to be called *factor loadings* are, in fact, correlations of the manifest variables with the factor. As we shall see, this idea can easily be extended to cover additional factors but that was not part of Spearman’s original discovery.

The Statistical Strand

The first passing contact of statistics with the developing factor analysis was the publication of **Harold Hotelling’s** seminal paper [6] on **principal component analysis**. PCA is quite distinct from factor analysis but the distinction was, perhaps, less clear in the 1930s. Hotelling himself was critical of factor analysis, especially because of its lack of the statistical paraphernalia of inferential statistics.

Hotelling was followed, quite independently it seems, by Bartlett, [2–4], whose name is particularly remembered in this field for what are known as ‘Bartlett’ scores. These are ‘factor scores’ and we shall return to them below (see **Factor Score Estimation**). He also wrote more widely on the subject

and, through his influence, Whittle [20] made a brief excursion into the field.

There the matter appears to have rested until the immediate postwar period. By then, statistics, in a modern guise, was making great progress. **M. G. Kendall**, who was a great systematizer, turned his attention to factor analysis in [9] and also included it in taught courses at about the time and in one of his early monographs on multivariate analysis. This period also marks D. N. Lawley’s contribution concerned especially with fitting the factor model, see, for example, [10]. His one-time colleague, **A. E. Maxwell**, who collaborated in the writing of the book *Factor Analysis as a Statistical Method* [11], did practical factor analysis in connection with his work at the London Institute of Psychiatry. His expository paper [14], first read at a conference of the Royal Statistical Society in Durham and subsequently published in the *Journal Series A*, is an admirable summary of the state of play around 1961. In particular, it highlights the problems of implementing the methods of fitting the model that had already been developed – uncertain convergence being prominent among them.

However, factor analysis did not ‘catch on’ in a big way within the statistical community and there were a number of critical voices. These tended to focus on the alleged arbitrariness of the method that so often seemed to lead to an unduly subjective treatment. The range of rotations available, oblique as well as orthogonal, left the user with a bewildering array of ‘solutions’ one of which, surely, must show what the analyst desired. Much of this ‘unfriendly fire’ was occasioned by the fact that, in practice, factor analysts showed little interest in sampling error. It was easily possible to demonstrate the pitfalls by simulation studies on the basis of small sample sizes, where sampling error was often mistaken for arbitrariness. To many statisticians, the solidity of principal components analysis provided a surer foundation even if it was, basically, only a descriptive technique. However, to psychologists, ‘meaningfulness’ was at least as important a criterion in judging solutions as ‘statistical significance’.

The immediate postwar period, 1950–1960 say, marks an important watershed in the history of factor analysis, and of statistics in general. We shall come to this shortly, but it owed its origin to two important happenings of this period. One was the introduction of the electronic computer, which was, ultimately,

to revolutionize multivariate statistical analysis. The other was the central place given to probability models in the specification and analysis of statistical problems. In a real sense, statistics became a branch of applied probability in a way that it had not been earlier.

Prior to this watershed, the theory of factor analysis was largely about the numerical analysis of correlation (and related) matrices. In a sense, this might be called a *deterministic* or *mathematical* theory. This became such a deeply held orthodoxy that it still has a firm grip in some quarters. The so-called problem of factor scores, for example, is sometimes still spoken of as a ‘problem’ even though its problematic character evaporates once the problem is formulated in modern terms.

Next Steps

The first main extension was to introduce more than one common factor. It soon became apparent in applied work that the original one-factor hypothesis did not fit much of the data available. It was straightforward, in principle, to extend the theory, and Burt was among the pioneers, though it is doubtful whether his claim to have invented multifactor analysis can be substantiated (see [13]).

At about the same time, the methods were taken up across the Atlantic, most conspicuously by **L. L. Thurstone** [19]. He, too, claimed to have invented multifactor analysis and, for a time at least, his approach was seen as a rival to Spearman’s. Spearman’s work had led him to see a single underlying factor (G) as being common to, and the major determinant of, measures of human ability. Eventually, Spearman realized that this dominant factor could not wholly explain the correlations and that other ‘group’ factors had to be admitted. Nevertheless, he continued to believe that the one-factor model captured the essence of the situation.

Thurstone, on the other hand, emphasized that the evidence could be best explained by supposing that there were several (7 or 9) primary abilities and, moreover, that these were correlated among themselves. To demonstrate the latter fact, it was necessary to recognize that once one passed beyond one factor the solution was not unique. One could move from one solution to another by simple transformations, known as *rotations*, because that is how

they can be regarded when viewed geometrically. Once that fact was recognized, the question naturally arose as to whether some rotations were better or ‘more meaningful’ than others. Strong claims may be advanced for those having what Thurstone called ‘simple’ structure. In such a rotation, each factor depends only (or largely) on a subset of the observable variables. Such variables are sometimes called *group* variables, for obvious reasons.

Two Factors

The question of whether the correlation matrix can be explained by a single underlying factor therefore resolved itself into the question of whether it has the structure (2). If one factor failed to suffice, one could go on to ask whether two factors or more would do the job better. The essentials can be made clear if we first limit ourselves to the case of two factors.

Suppose, then, we introduce two factors G_1 and G_2 . We then require $r_{ij.G_1G_2}$ to be zero for all $i \neq j$. If G_1 and G_2 are uncorrelated, it turns out that

$$r_{ij} = r_{iG_1}r_{jG_1} + r_{iG_2}r_{jG_2} \quad (i \neq j) \quad (3)$$

$$= \lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2}, \text{ say.} \quad (4)$$

Pursuing this line of argument to incorporate further factors, we find, in the q -factor case, that

$$r_{ij} = \sum_{k=1}^q \lambda_{ik}\lambda_{jk} \quad (i \neq j). \quad (5)$$

In matrix notation,

$$\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}, \quad (6)$$

where $\mathbf{\Lambda} = \{\lambda_{ik}\}$ and $\mathbf{\Psi}$ is a diagonal matrix whose elements are chosen to ensure that the diagonal elements of the matrix on the right add up to 1 and so match those of $\mathbf{R}$. The complements of the elements of $\mathbf{\Psi}$ are known as *the communalities* because they provide a measure of the variance attributable to the common factor.

The foregoing, of course, is not a complete account of the basis of factor analysis, even in its original form but it shows why the structure of the correlation matrix was the focal point. No question of a probability model arose and there was no discussion, for example, of standard errors of estimates. Essentially and originally, factor analysis

4 History of Factor Analysis: A Statistical Perspective

was the numerical analysis of a correlation matrix. This approach dominated the development of factor analysis before the Second World War and is still sometimes found today. For this reason, (6) was (and sometimes still is) spoken of as the factor analysis model.

The Model-based Approach

From about the 1950s onward, a fundamental change took place in statistics. This was the period when the ‘model’ became the point at which most statistical analysis began. A statistical model is a specification of the joint distribution of a set of random variables. Thus, if we have a set of variables $x_1, x_2, \dots, x_n$, a model will say something about their joint distribution.

Thus, Lawley and Maxwell’s *Factor Analysis as a Statistical Method* [11], which appeared in 1963, places what we would now call the normal linear factor model at the center. In factor analysis, there are three kinds of random variable. First, there are the manifest variables that we observe. We shall denote them by $x_1, x_2, \dots, x_p$ and make no distinction between random variables and the values they take. Then there are the factors, or **latent variables** denoted by $y_1, y_2, \dots, y_q$. The normal factor model supposes that

$$x_i \sim N\left(\mu_i + \sum_{j=1}^q \lambda_{ij} y_j, \Psi_i\right) \quad (i = 1, 2, \dots, p) \quad (7)$$

$$y_j \sim N(0, 1) \quad (j = 1, 2, \dots, q) \quad (8)$$

Or, equivalently,

$$x_i = \mu_i + \sum_{j=1}^q \lambda_{ij} y_j + e_i \quad (i = 1, 2, \dots, p), \quad (9)$$

where $e_i \sim N(0, \Psi_i)$ and where the e_i ’s are independent of the y_j ’s. The e ’s are the third kind of random variable referred to above. Their mutual independence expresses the conditional independence of the x ’s given the y ’s. The covariance matrix of the x_i ’s for this model is

$$\Sigma = \Lambda\Lambda' + \Psi, \quad (10)$$

which is of exactly the same form as (6) and so justifies us in regarding it as a stochastic version of the old (Spearman) ‘model’. The difference is that Σ is the covariance matrix rather than the correlation matrix. This is often glossed over by supposing that the x_i ’s have unit variance. This, of course, imposes a further constraint on Λ and Ψ by requiring that

$$\Psi_i + \sum_{j=1}^q \lambda_{ij}^2 = 1 \quad (i = 1, 2, \dots, p). \quad (11)$$

Viewed in a statistical perspective, we would now go on to fit the model, which amounts to finding estimates of Λ and Ψ to optimize some fitting function. The usual function chosen is the likelihood and the method is that of maximum likelihood (*see Maximum Likelihood Estimation*). In essence, the likelihood is a measure of the distance between Σ , as given by (10) and the sample covariance matrix. Other measures of distance, such as weighted or unweighted least squares, have also been used.

Prior to the 1950s, the problem of fitting the model was essentially that of finding Λ and Ψ in (10) to make it as close as possible to the sample covariance matrix without regard to the probabilistic interpretation.

When viewed in the statistical perspective, one can go on to construct tests of goodness-of-fit or calculate standard errors of estimates. That perspective also, as we shall see, provides a natural way of approaching other, related problems, which under the old approach were intractable, such as the so-called problem of *factor scores*.

Recent History of Factor Analysis

In the last few decades, factor analysis has developed in two different directions. One is in what is assumed about the factors and the other in what is assumed about the observable variables. In both cases, the scope of the basic factor model is enlarged.

What we have described so far is often known as *exploratory factor analysis*. Here, nothing is assumed *a priori* about the factor structure. The purpose of the analysis is simply to uncover whatever is there. Sometimes, on the other hand, there is prior information based either on previous empirical work or prior knowledge. For example, it may be known,

or suspected, that only the members of a given subset are indicators of a particular factor. This amounts to believing that certain factor loadings are zero. In cases like these, there is a prior hypothesis about the factor structure and we may then wish to test whether this is confirmed by a new data set.

This is called *confirmatory factor analysis* (see **Factor Analysis: Confirmatory**). Confirmatory factor analysis is a rather special case of a more general extension known as *linear structural relations modeling* or **structural equation modeling**. This originated with [7] and has developed enormously in the last 30 years. In general, it supposes that there are linear relationships among the latent variables. The object is then to not only determine how many factors are needed but to estimate the relationships between them. This is done, as in factor analysis, by comparing the observed covariance matrix with that predicted by the model and choosing the parameters of the latter to minimize the distance between them. For obvious reasons, this is often called *covariance structure analysis*.

Another ‘long-standing’ part of factor analysis can also be cast into the mold of linear structural relations modeling. This is what is known as *hierarchical factor analysis*, and it has been mainly used in intelligence testing. When factor analysis is carried out on several sets of test scores in intelligence testing, it is common to find that several factors are needed to account for the covariances – perhaps as many as 8 or 9. Often, the most meaningful solution will be obtained using an oblique rotation in which the resulting factors will themselves be correlated. It is then natural to enquire whether their covariances might be explained by factors at a deeper level, to which they are related. A second stage analysis would then be carried out to reveal this deeper factor structure. It might even be possible to carry the analysis further to successively deeper levels. In the past, hierarchical analysis has been carried out in an *ad hoc* way much as we have just described it. A more elegant way is to write the dependence between the first-level factors and the second level as linear relations to be determined. In this way, the whole factor structure can be estimated simultaneously.

The second kind of recent development has been to extend the range of observed variables that can be considered. Factor analysis was born in the context

of continuous variables for which correlations are the appropriate measure of association. It was possible, as we have seen, because the theory of partial correlation already existed. At the time, there was no such theory for categorical variables, whether ordered or not. This lopsided development reflected much that was going on elsewhere in statistics. Yet, in practice, categorical variables are very common, especially in the behavioral sciences, and are often mixed up with continuous variables. There is no good reason why this separation should persist. The logic of the problem does not depend, essentially, on the type of variable.

Extension to Variables of Other Types

Attempts to cope with this problem have been made in a piecemeal fashion, centering, to a large extent, on the work of Lazarsfeld, much of it conveniently set out in [12]. He introduced latent structure analysis to do for categorical – and especially binary variables – what factor analysis had done for continuous variables. Although he noted some similarities, he seemed more interested in the differences that concerned the computational rather than the conceptual aspects. What was needed was a broader framework within which a generalized form of factor analysis could be carried out regardless of the type of variable. Lazarsfeld’s work also pointed to a second generalization that was needed. This concerns the factors, or latent variables. In traditional factor analysis, the factors have been treated as continuous variables – usually normally distributed. There may be circumstances in which it would be more appropriate to treat the factors as categorical variables. This was done by Lazarsfeld with his latent class and latent profile models. It may have been partly because the formulae for models involving categorical variables look so different from those for continuous variables, that their essential unity was overlooked.

The key to providing a generalized factor analysis was found in the recognition that the exponential family of distributions provided a sufficient variety of forms to accommodate most kinds of observed variable. It includes the normal distribution, of course, but also the Bernoulli and multinomial distributions, to cover categorical data and many other forms as well that have not been much considered in latent variables analysis. A full development on these lines will be found in [1].

In this more general approach, the normal linear factor model is replaced by one in which the canonical parameter (rather than the mean) of the distribution is expressed as a linear function of the factors. Many features of the standard linear model carry over to this more general framework. Thus, one can fit the model by maximum likelihood, rotate factors, and so on. However, in one important respect it differs. It moves the focus away from correlations as the basic data about dependencies and toward the more fundamental conditional dependencies that the model is designed to express. It also resolves the disputes that have raged for many years about factor scores. A factor score is an ‘estimate’ or ‘prediction’ of the value of the factor corresponding to a set of values of the observed variables (*see Factor Score Estimation*). Such a value is not uniquely determined but, within the general framework, is a random variable. The factor score may then be taken as the expected value of the factor, given the data. It is curious that this has been the undisputed practice in **latent class analysis** from the beginning where allocation to classes has been based on posterior probabilities of class membership. Only recently is it becoming accepted that this is the obvious way to proceed in all cases.

Posterior probability analysis also shows that, in a broad class of cases, all the information about the latent variables is contained in a single statistic, which, in the usual statistical sense, is ‘sufficient’ for the factor. It is now possible to have a single program for fitting virtually any model in this wider class when the variables are of mixed type. One such is GENLAT due to Moustaki [15]. A general account of such models is given in [16].

Computation

Factor analysis is a computer-intensive technique. This fact made it difficult to implement before the coming of electronic computers. Various methods were devised for estimating the factor loadings and communalities for use with the limited facilities then available. The commonest of these, known as *the centroid method*, was based on geometrical ideas and it survived long enough to be noted in the first edition of Lawley and Maxwell [11]. Since then, almost all methods have involved minimizing the distance between the observed covariance (or correlation)

matrix $\mathbf{S}$ and its theoretical equivalent given by

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}. \quad (12)$$

These methods include least squares, weighted (or generalized) least squares, and maximum likelihood. The latter has been generally favored because it allows the calculation of standard errors and measures of goodness-of-fit. It is not immediately obvious that this involves a minimization of distance but this becomes apparent when we note that the log (likelihood) turns out to be

$$\begin{aligned} \log(\text{likelihood}) = \text{constant} + \frac{n}{2} \ln \det [\mathbf{\Sigma}^{-1}\mathbf{S}] \\ - \frac{n}{2} \text{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}], \end{aligned} \quad (13)$$

where $\mathbf{\Sigma}$ is the covariance matrix according to the model and $\mathbf{S}$ is the sample covariance matrix. We note that $\mathbf{\Sigma}^{-1}\mathbf{S} = \mathbf{I}$ if $\mathbf{\Sigma} = \mathbf{S}$ and, otherwise, is positive. This means that, even if the distributional assumptions required by the model are not met, the maximum likelihood method will still be a reasonable fitting method. There are two principal approaches to minimizing (13). One, adopted by Jöreskog and Sörbom [8] uses the Fletcher–Powell (*see Optimization Methods*) algorithm. The second is based on the E-M algorithm. The latter has the conceptual advantage that it can be developed for the much wider class of models described in the section titled Extension to Variables of Other types.

The major software packages that are now available, also allow for various kinds of rotation, the calculation of factor scores, and many other details of the analysis.

In spite of the fact that the main computational problems of fitting have been solved, there are still complications inherent in the model itself. Most noteworthy are what are known as *Heywood cases*. These arise from the fact that the elements of the diagonal matrix $\mathbf{\Psi}$ are variances and must, therefore, be nonnegative. Viewed geometrically, we are looking for a point in the parameter space (of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$) that maximizes the likelihood. It may then happen that the maximum is a boundary point at which one or more elements of $\mathbf{\Psi}$ is zero. The problem arises because such a boundary solution can, and often does, arise when the ‘true’ values of all the elements of $\mathbf{\Psi}$ are strictly positive. There is nothing inherently impossible about a zero value of a residual variance but they do seem practically implausible.

Heywood cases are an inconvenience but their occurrence emphasizes the inherent uncertainty in the estimation of the parameters. They are much more common with small sample sizes and the only ultimate ‘cure’ is to use very large samples.

What Then is Factor Analysis?

Factor analysis has appeared under so many guises in its 100-year history that one may legitimately ask whether it has retained that unitary character that would justify describing it as a single entity. Retrospectively, we can discern three, overlapping, phases that have coexisted. The prominence we give to each may depend, to some extent, on what vantage point we adopt – that of psychologist, statistician, or general social scientist.

At the beginning, and certainly in Spearman’s view, it was concerned with explaining the pattern in a correlation matrix. Why, in short, are the variables correlated in the way they are – it thus became a technique for explaining the pattern of correlation coefficients in terms of their dependence on underlying variables. It is true that the interpretation of those correlations depended on the linearity of relations between the variables but, in essence, it was the correlation coefficients that contained the relevant information. Obviously, the technique could only be used on variables for which correlation coefficients could be calculated or estimated.

The second approach is to write down a probability model for the observed (manifest) variables. Traditionally, these variables have been treated as continuous and it is then natural to express them as linear in the latent variables, or factors. In the standard normal linear factor model, the joint distribution of the manifest variables is multivariate normal and thus depends, essentially, on the covariance matrix of the data. We are thus led to the *covariance* rather than the *correlation* matrix as the basis for fitting. Formally, we have reached almost the same point as in the first approach though this is only because of the particular assumptions we have made. However, we can now go much further because of the distributional assumptions we have made. In particular, we can derive standard errors for the parameter estimates, devise goodness-of-fit tests, and so forth.

The third and final approach is to drop the specific assumptions about the kinds of variable and their

distributions. The focus then shifts to the essential question that has underlain factor analysis from the beginning. That is, is the interdependence among the manifest variables indicative of their dependence on a (small) number of factors (latent variables)? It is then seen as one tool among many for studying the dependence structure of a set of random variables. From that perspective, it is seen to have a much wider relevance than Spearman could ever have conceived.

References

- [1] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd Edition, Arnold, London.
- [2] Bartlett, M.S. (1937). The statistical concept of mental factors, *British Journal of Psychology* **28**, 97–104.
- [3] Bartlett, M.S. (1938). Methods of estimating mental factors, *Nature* **141**, 609–610.
- [4] Bartlett, M.S. (1948). Internal and external factor analysis, *British Journal of Psychology (Statistical Section)* **1**, 73–81.
- [5] Burt, C. (1941). *The Factors of the Mind: An Introduction to Factor Analysis in Psychology*, Macmillan, New York.
- [6] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.
- [7] Jöreskog, K.G. (1970). A general method for analysis of covariance structures, *Biometrika* **57**, 239–251.
- [8] Jöreskog, K.G. & Sörbom, D. (1977). Statistical models and methods for analysis of longitudinal data, in *Latent Variables in Socioeconomic Models*, D.J. Aigner & A.S. Goldberger, eds, North-Holland, Amsterdam.
- [9] Kendall, M.G. (1950). Factor analysis as a statistical technique, *Journal of the Royal Statistical Society. Series B* **12**, 60–63.
- [10] Lawley, D.N. (1943). The application of the maximum likelihood method to factor analysis, *British Journal of Psychology* **33**, 172–175.
- [11] Lawley, D.N. & Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*, Butterworths, London.
- [12] Lazarsfeld, P.E. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton Mifflin, New York.
- [13] Lovie, A.D. & Lovie, P. (1993). Charles Spearman, Cyril Burt and the origins of factor analysis, *Journal of the History of the Behavioral Sciences* **29**, 308–321.
- [14] Maxwell, A.E. (1961). Recent trends in factor analysis, *Journal of the Royal Statistical Society. Series A* **124**, 49–59.
- [15] Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* **56**, 337–357.

8 History of Factor Analysis: A Statistical Perspective

- [16] Moustaki, I. & Knott, M. (2000). Generalized latent trait models, *Psychometrika* **65**, 391–411.
- [17] Pearson, K. (1901). On lines and planes of closest fit to a system of points in space, *Philosophical Magazine* (6th Series), **2**, 557–572.
- [18] Spearman, L. (1904). “General intelligence” objectively determined and measured, *American Journal of Psychology* **15**, 201–293.
- [19] Thurstone, L.L. (1947). *Multiple Factor Analysis*, University of Chicago Press, Chicago.
- [20] Whittle, P. (1953). On principal components and least square methods of factor analysis, *Skandinavisk Aktuarietidskrift* **55**, 223–239.
- [21] Yule, G.U. (1911). *Introduction to the Theory of Statistics*, Griffin, London.

(See also **Structural Equation Modeling: Categorical Variables; Structural Equation Modeling: Latent Growth Curve Analysis; Structural Equation Modeling: Multilevel**)

DAVID J. BARTHOLOMEW

History of Intelligence Measurement

NADINE WEIDMAN

Volume 2, pp. 858–861

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Intelligence Measurement

The history of intelligence measurement can be roughly divided into three chronological periods: the first, a period of initial interest in defining intelligence and the establishment and use of intelligence tests; the second, a period of repudiation of the hereditarian assumptions underlying the tests; and the third, a period of resurgence of belief in the connection between intelligence and heredity. The first period lasted from about 1870 to the 1920s and is associated with **Francis Galton**, Alfred Binet, **Charles Spearman**, Robert Yerkes, and Lewis Terman. The second period, from the 1930s to the 1960s, is associated with the critics of intelligence testing such as Otto Klineberg and Horace Mann Bond. Finally, the period of resurgence began in 1969 with Arthur Jensen's controversial claims that intelligence is largely hereditary, claims repeated and enlarged upon by Richard Herrnstein and Charles Murray in their 1994 work, *The Bell Curve*.

Like his first cousin Charles Darwin, Francis Galton (1822–1911) substituted a belief in evolution and the power of heredity for religious orthodoxy. Like Darwin, he assumed that variation of traits in a population was key to understanding evolution and that most variation was hereditary. But rather than analyzing hereditary variation by seeking to understand its physiological cause, Galton chose to treat it statistically: by examining its distribution in a population. He noted that any given trait – height, say, or weight – was ‘normally’ distributed in a population, along a ‘bell curve’, with most individuals in the population displaying moderate height or weight, and fewer outliers at either extreme. But Galton did not stop at the measurement of physical traits: indeed, he believed it was even more important to measure mental traits, and these too he thought were normally distributed. In his 1869 work *Hereditary Genius: An Inquiry into Its Laws and Consequences*, Galton argued that genius, or talent, was inborn, that it tended to run in families, and that one's reputation was an accurate measure of one's inborn ability [5]. During the 1870s, Galton built up an arsenal of statistical concepts to treat heredity as a measurable relationship between generations. These concepts included **regression to**

the mean (in which certain characteristics revert to more typical values with each passing generation) and the coefficient of correlation (the degree to which one variable depends upon another). Galton's passion for measurement and belief in the power of heredity, combined with his concern for future social progress and fears about the decline of civilization, led him to advocate a program of ‘eugenics’ (a term he coined in 1883): a system of controlled mating in which those with desirable hereditary traits were encouraged to marry and produce offspring, while those deemed unfit were prevented from mating [9].

Charles Spearman (1863–1945), professor of psychology at University College, London, took up Galton's interest in using statistical tools to measure hereditary mental traits. Using the concept of the coefficient of correlation, Spearman determined that an individual's level of ability tends to hold steady in many different tasks, whether those be understanding a literary text or doing a mathematical calculation. The sameness of these abilities, the fact of their correlation, could be accounted for by their dependence on the individual's general intelligence – which Spearman identified as the ‘*g* factor’. General intelligence pervaded all of the individual's mental abilities and mental processes, was hereditarily determined, and held constant over an individual's lifetime. For Spearman, any given ability depended on two factors, the *g* factor and a special factor (*s*) that determined facility in a specific task. Spearman also believed that general intelligence was a real thing, an actual entity that exists in the brain, and for which a physiological correlate must ultimately be found [6]. While debates raged over the reification of intelligence, psychologists used Spearman's method of factor analysis, by which the variability of a trait can be reduced to one or more underlying factors or variables, to claim scientific status in the 1920s and 1930s [12]. Spearman's student **Cyril Burt** (1883–1971) broadened Spearman's use of factor analysis from intelligence and in his 1940 work *Factors of the Mind* applied it to analyzing emotion and personality [1].

In France, the psychologist Alfred Binet (1857–1911) developed an approach to understanding intelligence that was very different from Spearman's. The French government commissioned Binet in 1904 to produce a test of ability to identify sub-normal children in school classrooms, so that they

2 History of Intelligence Measurement

could be removed and given special education, allowing the other children to progress normally. Binet had previously been interested in the experimental study of the highest and most complex mental processes, and of individuals of high ability; with his colleague Theodore Simon (1873–1961), Binet determined what tasks a normal child, of a given age, could be expected to do, and then based his test on a series of 30 tasks of graded difficulty. Binet and Simon published their results in *L'annee Psychologique* in 1905, revising their test in 1908 and again in 1911. Though some identified the general capacity that such a test seemed to assess with Spearman's general factor of intelligence, Binet and Simon referred to the ability being tested as 'judgment'. They were, unlike Spearman, more interested in the description of individuals than in developing a theory of general intelligence, and their work did not have the strong hereditarian overtones that Spearman's did [6].

Binet and Simon's test for mental ability was refined and put to use by many other psychologists. In Germany, the psychologist William Stern argued in 1912 that the mental age of the child, as determined by the test, should be divided by the child's chronological age, and gave the number that resulted the name 'IQ' for intelligence quotient [9]. But it was in the United States that the Binet–Simon IQ test found its most receptive audience, and where it was put to the hereditarian ends that both Binet and Stern had renounced. The psychologist Henry Herbert Goddard (1866–1957), for example, used the test to classify patients at the New Jersey Training School for Feeble-minded Boys and Girls, a medical institution housing both children and adults diagnosed with mental, behavioral, and physical problems. Goddard subsequently developed, in part on the basis of his experience at the Training School, a theory that intelligence was unitary and was determined by a single genetic factor. He also used IQ tests on immigrants who came to America through the Ellis Island immigration port [13].

At Stanford University, the educational psychologist Lewis Terman (1877–1956) and his colleagues used IQ tests to determine the mental level of normal children, rather than to identify abnormal ones, an application that represented a significant departure from Binet's original intention. Terman called his elaboration of Binet's test the 'Stanford-Binet', and it became the predecessor and prototype

of the standardized, multiple-choice tests routinely taken by American students from the elementary grades through the college and postgraduate years [3]. But, psychologists moved the intelligence test beyond its application to performance in school. With the entrance of the United States into World War I in 1917, the comparative psychologist Robert M. Yerkes, supported by the National Research Council, proposed to the US Army a system of mass testing of recruits, which would determine whether they were fit for army service and, if so, what tasks best suited them [2]. Mass testing of thousands of soldiers differed greatly from Binet's individualist emphasis, but it raised psychology's public profile considerably: after the war, psychologists could justifiably call themselves experts in human management [4, 10]. Again, the results of the army testing were interpreted in hereditarian ways: psychologists argued that they showed that blacks and immigrants, especially from southern and eastern Europe, were less intelligent than native-born whites. Such arguments lent support to the call for immigration restriction, which passed into law in 1924.

Even as the IQ testers achieved these successes, they began to receive harsh criticism. Otto Klineberg (1899–1992), a psychologist trained under the anthropologist Franz Boas, made the best known and most influential attack. Klineberg argued that the supposedly neutral intelligence tests were actually compromised by cultural factors and that the level of education, experience, and upbringing so affected a child's score that it could not be interpreted as a marker of innate intelligence. Klineberg's work drew on that of lesser-known black psychologists, most notably Horace Mann Bond (1904–1972), an educator, sociologist, and university administrator. Bond showed that the scores of blacks from the northern states of New York, Ohio, and Pennsylvania were higher than those of southern whites, and explained the difference in terms of better access to education on the part of northern blacks. Such an argument flew in the face of innatist explanations. Nonetheless, despite his criticisms of hereditarian interpretations of the tests, Bond never condemned the tests outright and in fact used them in his work as a college administrator. Intelligence tests could, he argued, be used to remedy the subjectivity of individual teachers' judgments. If used properly – that is, for the diagnosis of learning problems –

and if interpreted in an environmentalist way, Bond believed that the tests could actually subvert bias. By the mid-1930s, Bond's evidence and arguments had severely damaged the hereditarian interpretation of IQ test results [11].

By 1930, too, several prominent psychologists had made public critiques or undergone well-publicized reversals on testing. E. G. Boring (1886–1968) expressed skepticism that intelligence tests actually measured intelligence. And Carl Brigham (1890–1943), who had in 1923 published a racist text on intelligence, recanted his views by the end of that decade [6].

The trend toward environmentalist and cultural critiques of intelligence testing met a strong opponent in Arthur Jensen, a psychologist at the University of California, Berkeley. In 1969, his controversial article 'How Much Can We Boost I.Q. and Scholastic Achievement' claimed that 'compensatory education has been tried and it apparently has failed' [8]. Jensen argued that it was in fact low IQ, not discrimination, cultural or social disadvantages, or racism that accounted for minority students' poor performance in intelligence tests and in school. His claim relied to an extent on Cyril Burt's twin studies, which purported to show that identical twins separated at birth and raised in different environments were highly similar in mental traits and that such similarity meant that intelligence was largely genetically determined. (In 1974, Leon J. Kamin investigated Burt's twin studies and concluded that Burt had fabricated his data.) Jensen's argument was in turn echoed by the Harvard psychologist Richard Herrnstein (1930–1994), who argued that because IQ was so highly heritable, one should expect a growing stratification of society based on intelligence and that this was in fact happening in late twentieth-century America. Expanded and developed, this same argument appeared in *The Bell Curve: Intelligence and Class Structure in American Life*, which Herrnstein published with the political scientist Charles Murray in 1994 [7]. Both in 1970 and 1994, Herrnstein's argument met a firestorm of criticism.

Attempts to define and measure intelligence are always tied to social and political issues, so the controversy that attends such attempts should come as

no surprise. Just as the post-World War I enthusiasm for IQ testing must be understood in the context of immigration restriction, Jensen's and Herrnstein's interest in intelligence and heredity arose against a background of debates over civil rights, affirmative action, and multiculturalism. From Galton's day to the present, IQ testers and their critics have been key players in the ongoing conversation about the current state and future direction of society.

References

- [1] Burt, C. (1940). *The Factors of the Mind*, University of London Press, London.
- [2] Carson, J. (1993). Army alpha, army brass, and the search for army intelligence, *Isis* **84**, 278–309.
- [3] Chapman, P.D. (1988). *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890–1930*, New York University Press, New York.
- [4] Fancher, R. (1985). *The Intelligence Men: Makers of the I.Q. Controversy*, Norton, New York.
- [5] Galton, F. (1869). *Hereditary Genius: An Inquiry into Its Laws and Consequences*, Horizon Press, New York.
- [6] Gould, S.J. (1981). *The Mismeasure of Man*, Norton, New York.
- [7] Herrnstein, R. & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*, Free Press, New York.
- [8] Jensen, A. (1969). How much can we boost I.Q. and scholastic achievement? *Harvard Educational Review* **39**, 1–123.
- [9] Smith, R. (1997). *The Norton History of the Human Sciences*, Norton, New York.
- [10] Sokal, M.M., ed. (1987). *Psychological Testing and American Society*, Rutgers University Press, New Brunswick.
- [11] Urban, W.J. (1989). The black scholar and intelligence testing: the case of Horace Mann bond, *Journal of the History of the Behavioral Sciences* **25**, 323–334.
- [12] Wooldridge, A. (1994). *Measuring the Mind: Education and Psychology in England, c. 1860–c. 1990*, Cambridge University Press, Cambridge.
- [13] Zenderland, L. (1998). *Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing*, Cambridge University Press, Cambridge.

NADINE WEIDMAN

History of Mathematical Learning Theory

SANDY LOVIE

Volume 2, pp. 861–864

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Mathematical Learning Theory

For those psychologists with a sense of their discipline's past, the best known mathematical learning theorist has to be **Clark L. Hull** whose *Principles of Behavior* [10] and related works appeared to epitomize modern formal thinking in the behavioral sciences. Unfortunately, a closer look at the models and the modeling process shows Hull to be more of a nineteenth-century figure than one from the twentieth in that the models were fundamentally deterministic, and had been created by less than transparent or accepted mathematical means such as least squares. Consequently, this account of mathematical learning theory will not start or concern itself with Hull, but with a small but skillful group of psychologists and statisticians whose work was considerably more professional and more up-to-date than Hull's, and hence more worthy of the label 'modern'. These include **Robert R. Bush**, Frederick Mosteller, William K. Estes, and R. Duncan Luce. Between them, they created what quickly became known as Mathematical Learning Theory (MLT), although their ideas were rapidly taken up by workers in many other areas, thus subsuming MLT into the larger activity of an emerging *mathematical psychology* (see the three volumes of the *Handbook of Mathematical Psychology*, [14, 15], edited by Luce, Bush, and Galanter, for a detailed picture of the field's scope in the early 1960s).

What distinguished the approach of all four workers from the earlier efforts by Hull are their views about behavior: this was taken to be *intrinsically* uncertain and probabilistic. Early on, Estes made this explicit by referring to his brand of MLT as *statistical* learning theory, while Bush and Mosteller titled their seminal 1955 text *Stochastic Models for Learning* [5]. Hull, in comparison, had generated models which assumed that all behavior could be represented by a nonprobabilistic process, with any variation being bolted onto this essentially deterministic framework as error or 'behavioral oscillation' (to use Hull's phrase), in much the same way as a linear regression model (see **Multiple Linear Regression**) consists of fixed, unvarying components plus a random error variable. This nineteenth-century Newtonian worldview had long vanished in physics and related sciences under the onslaught of quantum theory, leaving

Hull (who was often claimed to be psychology's Isaac Newton) as a conceptually conservative figure in spite of the apparent novelty of, say, his use of modern symbolic logic in formalizing an axiomatic system for rote learning studies [11]. What the modern advocates of MLT did was to base all their modeling on the *probability* of action, with the clear assumption that this was the only way in which one could approach behavior. Many also drew on quantum theory formulations, in particular, **Markov chains**, which had originally been developed to model the probabilistic emission of particles from radioactive sources (see the writings of the MLT guru William Feller: for example, his 1957 textbook [8]). Thus, the conceptual basis of all flavors of these early versions of MLT was the probability of the event of interest, usually a response, or the internal processes that generated it, including the sampling and linking of stimulus and response elements by some probabilistic conditioning mechanism. (Notice that in the 1950s, conditioning, whether of the Pavlovian, Skinnerian, or Guthrieian variety, or mixtures of the three, tended to dominate empirical and theoretical work in learning.)

Although the honor of publishing the first paper embodying the new approach has to go to Estes in 1950, the most ambitious early programme into MLT was undoubtedly the one undertaken by Bush and Mosteller. This culminated in their 1955 textbook, which not only laid down a general system of considerable maturity and sophistication for modeling behavior, but also applied this to the detailed analysis of results culled from five areas of human and animal learning: imitation, avoidance learning, maze running, free recall verbal learning, and symmetric choice [5]. However, the claimed generality of this work, together with its lack of commitment to any particular theory of learning in that it attempted to embrace (and model) them all, meant that from the start there existed a certain distance or conceptual tension between Estes, on the one hand, and Bush and Mosteller, on the other. Thus, Estes had early on explicitly positioned his work within a Guthrieian setting by referring to it as 'Stimulus Sampling Theory' [7], while Bush and Mosteller commented that 'Throughout this book we have attempted to divorce our model from particular psychological theories' [4, p. 332].

This tension was somewhat increased by Bush and Mosteller's formal claim that Estes's stimulus sampling theory could be subsumed under their

2 History of Mathematical Learning Theory

system of *linear operators* (see their [5, Chapter 2]; also their initial comments that ‘a stimulus model is not necessary to the operator approach’, [5, p. 46]). What they were attempting was the development of a flexible mathematical system which could be tweaked to model many theoretical approaches in psychology by varying the range (and meaning) of allowable parameter values (but not model type) according to both the theory and the experimental domain. So ambitious a project was, however, almost impossible to carry out in practice, particularly as it also assumed a narrowly defined class of models, and was eminently mis-understandable by learning theorist and experimentalist alike. And so it proved. What now happened to MLT from the late 1950s was an increasing concentration on technical details and the fragmentation of the field as a result of strong creative disagreements, with infighting over model fit replacing Bush and Mosteller’s 1955 plea for a cumulative process of model development; tendencies, which, paradoxically, they did little to discourage. Indeed, their eight model comparison in [6] using the 1953 Solomon and Wynne shock avoidance data could be said to have kick-started the competitive phase of MLT, a direction hastened by the work of Bush, Galanter, and Luce in the same 1959 volume, which pitted Luce’s *beta* model for individual choice against the linear operator one, in part using the same Solomon and Wynne summary numbers [4].

Of course, comparing one model with another is a legitimate way of developing a field, but the real lack at the time of any deep or well-worked out theories of learning meant that success or failure in model-fitting was never unambiguous, with the contenders usually having to fall back on informal claims of how much closer their (*unprioritized* and *unprioritizable*) collective predictions were to the data than those of their opponents. This also made the issue of *formal* tests of goodness of fit, such as chi-square, problematic for many workers (see Bower’s careful but ultimately unsatisfactory trip around this issue in [9, pp. 375–376]). Furthermore, the epistemological deficit meant that MLT would sooner or later have to face up to the problem of *identifiability*, that is, how well do models need to be substantively and formally specified in order to *uniquely* and *unambiguously* represent a particular data set. Not to do so opened up the possibility of finding that MLT’s theories are typically

underdetermined by the data, to quote the standard postpositivist mantra. (Consult [13, especially Chapter 12], for a carefully drawn instance of how to handle some problems of identifiability in reaction time studies originally raised by Townsend in, for example, [16]; see also [12] for a case study in the history of factor analysis).

Meanwhile, and seemingly almost oblivious to this debate, Estes single-mindedly pursued his vision of MLT as Stimulus Sampling Theory (SST), which claimed to be closer than most versions of MLT to psychological theorizing. Increasingly, however, SST was viewed as a kind of *meta-theory* in that its major claim to fame was as a creative resource rather than its instantiation in a series of detailed and specific models. Thus, according to Atkinson et al. [1, p. 372], ‘Much as with any general heuristic device, stimulus sampling theory should not be thought of as provable or disprovable, right or wrong. Instead, we judge the theory by how useful it is in suggesting specific models that may explain and bring some degree of orderliness into the data’. Consequently, from the 1966 edition of their authoritative survey of learning theory onwards, Hilgard and Bower treated MLT as if it was SST, pointing to the approach’s ability to generate testable models in just about every field of learning, from all varieties of conditioning to concept identification and two person interactive games, via signal detection and recognition, and spontaneous recovery and forgetting. In fact, Bower, on pages 376 to 377 of his 1966 survey of MLT [9], lists over 25 distinctive areas of learning and related fields into which MLT, in the guise of SST, had infiltrated (see also [1], in which Atkinson et al. take the same line over SST’s status and success). By the time of the 1981 survey [3], Bower was happy to explicitly equate MLT with SST, and to impute genius to Estes himself (p. 252).

Finally, all these developments, together with the increasing power of the computer metaphor for the human cognitive system, also speeded up the recasting and repositioning of the use of mathematics in psychology. For instance, Atkinson moved away from a completely analytical and formal approach by mixing semiformal devices such as flow charts and box models (used to sketch in the large scale anatomy of such systems as human memory) with mathematical models of the process side, for example, the operation of the rehearsal buffer linking the short and long term memory stores (see [2]). Such

hybrid models or approaches allowed MLT to remain within, and contribute to, the mainstream of experimental psychology for which a more thoroughgoing mathematical modeling was a minority taste. Interestingly, a related point was also advanced by Bower in his survey of MLT [9], where a distinction was made between rigorous mathematical systems with only minimal contact with psychology (like the class of linear models proposed by Bush and Mosteller) and overall ones like SST, which claimed to represent well-understood psychological processes and results, but which made few, if any, specific predictions. Thus, on page 338 of [9], Bower separates *specific-quantitative* from *quasi-quantitative* approaches to MLT, but then opts for SST on the pragmatic grounds that it serves as a persuasive example of *both*.

References

- [1] Atkinson, R.C., Bower, G.H. & Crothers, E.J. (1965). *An Introduction to Mathematical Learning Theory*, Wiley, New York.
- [2] Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes, in *The Psychology of Learning and Motivation*, Vol. 2, K.W. Spence & J.T. Spence, eds, Academic Press, New York.
- [3] Bower, G.H. & Hilgard, E.R. (1981). *Theories of Learning*, 5th Edition, Prentice Hall, Englewood Cliffs.
- [4] Bush, R.R., Galanter, E. & Luce, R.D. (1959). Tests of the beta model, in *Studies in Mathematical Learning Theory*, R.R. Bush & W.K. Estes, eds, Stanford University Press, Stanford.
- [5] Bush, R.R. & Mosteller, F. (1955). *Stochastic Models for Learning*, Wiley, New York.
- [6] Bush, R.R. & Mosteller, F. (1959). A comparison of eight models, in *Studies in Mathematical Learning Theory*, R.R. Bush & W.K. Estes, eds, Stanford University Press, Stanford.
- [7] Estes, W.T. (1950). Towards a statistical theory of learning, *Psychological Review* **57**, 94–107.
- [8] Feller, W. (1957). *An Introduction to Probability Theory and its Applications*, Vol. 1, 2nd Edition, Wiley, New York.
- [9] Hilgard, E.R. & Bower, G.H. (1966). *Theories of Learning*, 3rd Edition, Appleton-Century-Crofts, New York.
- [10] Hull, C.L. (1943). *Principles of Behavior*, Appleton-Century-Crofts, New York.
- [11] Hull, C.L., Hovland, C.J., Ross, R.T., Hall, M., Perkins, D.T. & Fitch, F.G. (1940). *Mathematico-deductive Theory of Rote Learning*, Yale University Press, New York.
- [12] Lovie, P. & Lovie, A.D. (1995). The cold equations: Spearman and Wilson on factor indeterminacy, *British Journal of Mathematical and Statistical Psychology* **48**(2), 237–253.
- [13] Luce, R.D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organisation*, Oxford University Press, New York.
- [14] Luce, R.D., Bush, R.R. & Galanter, E., eds (1963). *Handbook of Mathematical Psychology*, Vols. 1 & 2, Wiley, New York.
- [15] Luce, R.D., Bush, R.R. & Galanter, E., eds (1965). *Handbook of Mathematical Psychology*, Vol. 3, Wiley, New York.
- [16] Townsend, J.T. (1974). Issues and models concerning the processing of a finite number of inputs, in *Human Information Processing: Tutorials in Performance and Cognition*, Kantowitz, B.H., ed., Erlbaum, Hillsdale.

SANDY LOVIE

History of Multivariate Analysis of Variance

SCOTT L. HERSHBERGER

Volume 2, pp. 864–869

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Multivariate Analysis of Variance

Origins of Multivariate Analysis of Variance (MANOVA)

Building on the work of **Karl Pearson**, who derived the chi-square distribution in 1900, and of ‘Student’ **W.S. Gossett** who derived the t distribution in 1908, **Fisher** in 1923 introduced the **analysis of variance (ANOVA)** as a useful approach to studying population differences on a single ($p = 1$) dependent variable. The multivariate generalization of ANOVA – **multivariate analysis of variance (MANOVA)** – for studying population differences on $p > 1$ dependent variables soon followed. (NB Although Bartlett [2] and Roy [16–18] used variations of the term, the term ‘multivariate analysis of variance’, exactly written as such, is attributable to Roy [19].)

The MANOVA procedure was originally developed by Wilks [22] in 1932 on the basis of the generalized likelihood-ratio (LR), an application of Fisher’s **maximum likelihood** principle (*see Maximum Likelihood Estimation*). Fisher had introduced maximum likelihood in germinal form in 1912 [5] but did not provide a full development until 1922 [6]. The principle of maximum likelihood provides a statistical criterion for evaluating the consistency of a set of data with hypotheses concerning the data. Suppose we have N independent and identically distributed random variables denoted

$$\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]' \quad (1)$$

in column vector notation, a corresponding column vector of observed data

$$\mathbf{y} = [y_1, y_2, \dots, y_N]' \quad (2)$$

drawn from $\mathbf{Y}$, and a joint probability density function (pdf) given by $f(\mathbf{y}; \Theta)$ with q unknown pdf parameters denoted as a column vector

$$\Theta = [\theta_1, \theta_2, \dots, \theta_q]' \quad (3)$$

The principle of maximum likelihood recommends that an estimate for Θ be found such that it maximizes the likelihood of observing those data that were

actually observed. In other words, given a sample of observations $\mathbf{y}$ for the random vector $\mathbf{Y}$, find the solution for Θ that maximizes the joint probability density function $f(\mathbf{y}; \Theta)$.

Importantly, the likelihood computed for a set of data is based on a hypothesis concerning that data: The likelihood will vary under different hypotheses. That hypothesis which produces the ‘maximum’ likelihood is the most consistent with the distribution of the data. By examining the ratio of likelihoods computed under two different hypotheses, we can determine which likelihood is more consistent with data. Suppose that the likelihood of the sample on H_0 is L_0 and that of H_1 is L_1 . The ratio L_0/L_1 gives us some measure of the ‘closeness’ of the two hypotheses. If they are identical the ratio is unity. As they diverge, the ratio diminishes to zero.

The LR provides a criterion by which we can compare the two hypotheses typically specified in MANOVA: (a) a null hypothesis H_0 that several k population centroids μ are equal ($\mu_1 = \mu_2 = \dots = \mu_k$.) and (b) a non-null hypothesis H_1 that at least two population centroids are not equal ($\mu_1 \neq \mu_2 \neq \dots \neq \mu_k$.) Lower ratios suggest less probable null hypotheses, conversely, higher ratios suggest more probable null hypotheses. The LR underlies the development of the test statistic Λ , Wilks’s Lambda, for comparing the means of several dependent variables between more than two groups. Arguably, it was not Fisher’s work that was most directly responsible for Wilks’s development of the generalized LR . While Fisher emphasized the use of the LR principle for parameter estimation, **Jerzy Neyman** and **Egon Pearson** focused on the hypothesis testing possibilities of the LR . These authors in a 1928 paper [12] used the LR for hypothesis testing that was restricted to comparing any number of k groups on a *single* dependent variable.

We briefly describe the derivation of Wilks’s Λ from the LR of two hypotheses; considerably greater detail can be found in [1]. In MANOVA, under the null hypothesis, we assume that a common multivariate normal probability density function (*see Catalogue of Probability Density Functions*) describes each group’s data.

Therefore, L_0 is

$$L_0 = \frac{e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}Np}} \cdot \frac{1}{|S|^{\frac{1}{2}}}, \quad (4)$$

2 History of Multivariate Analysis of Variance

where S is the pooled within-groups covariance matrix. For L_1 , we need the likelihood of k separate multivariate normal distributions; in other words, each of the k multivariate normal distribution is described by a different mean and covariance structure. The likelihood for the non-null hypothesis, of group inequality, is

$$L_1 = \frac{e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}Np}} \prod_{t=1}^k \frac{1}{|S_t|^{\frac{1}{2}n_t}}, \quad (5)$$

where n_t is the sample size of an individual group. Thus to test the hypothesis that the k samples are drawn from the same population as against the alternative that they come from different populations, we test the ratio L_0/L_1 :

$$\begin{aligned} LR = \frac{L_0}{L_1} &= \frac{\frac{e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}Np}} \cdot \frac{1}{|S|^{\frac{1}{2}N}}}{\frac{e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}Np}} \prod_{t=1}^k \frac{1}{|S_t|^{\frac{1}{2}n_t}}} \\ &= \frac{\prod_{t=1}^k |S_t|^{\frac{1}{2}n_t}}{|S|^{\frac{1}{2}N}} \\ &= \prod_{t=1}^k \left\{ \frac{|S_t|}{|S|} \right\}^{\frac{1}{2}n_t}. \end{aligned} \quad (6)$$

Further simplification of the LR is possible when we recognize that the numerator represents the between-groups variance and the denominator represents the total variance. Therefore, we have

$$LR = \prod_{t=1}^k \{ |S_t|/|S| \}^{\frac{1}{2}n_t} = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|}, \quad (7)$$

where $|\mathbf{W}|$ is the determinant of the within-groups sum of squares (SS_{within}) and cross-products (CP_{within}) matrix, $|\mathbf{B} + \mathbf{W}|$ is the determinant of the sum of the between-groups sum of squares (SS_{between}) and cross-products (CP_{between}) matrix and the within-groups SS_{within} , CP_{within} matrix, and $|\mathbf{T}|$ is the determinant of the total sample sum of squares (SS_{total}) and cross-products (CP_{total}) matrix. The ratio $|\mathbf{W}|/|\mathbf{T}|$ is Wilks's Λ . Note that as $|\mathbf{T}|$ increases relative to $|\mathbf{W}|$ the ratio decreases in size with an

accompanying increase in the probability of rejecting H_0 .

Lambda is a family of three-parameter curves, with parameters based on the number of groups, the number of subjects, and the number of dependent variables, and is thus complex. Although Λ has been tabled for specific values of its parameters [7, 8, 10, 20], the utility of Λ depends on its transformation to either an exact or approximate χ^2 or F statistic. Bartlett [2] proposed an approximation to Λ in 1939 based on the chi-square distribution:

$$\chi^2 = -[(N-1) - 0.5(p+k)] \ln \Lambda, \quad (8)$$

which is evaluated at $p(k-1)$ degrees of freedom. Closer asymptotic approximations have been given by Box [4] and Anderson [1]. Transformations to exact chi-squared distributions have been given by Schatzoff [21], Lee [11], and Pillai and Gupta [14].

Rao [15] derived an F statistic in 1952 which provides better approximations to Λ cumulative probability densities compared to approximate chi-square statistics, especially when sample size is relatively small:

$$F = \left[\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \right] \left[\frac{ms - i(j-1)/2 + 1}{i(k-1)} \right], \quad (9)$$

where $m = N - 1 - (p+k)/2$, $s = \sqrt{[(p^2(k-1)^2 - 4/p^2 + (k-1)^2 - 5)]}$, and with $p(k-1)$, $ms - p(k-1)/2 + 1$ degrees of freedom.

In general, the LR principle provides several optimal properties for reasonably sized samples, and is convenient for hypotheses formulated in terms of multivariate normal parameters. In particular, the attractiveness of the LR presented by Wilks is that it yields test statistics that reduce to the familiar univariate F and t statistics when $p = 1$. If only one dependent variable is considered, $|\mathbf{W}| = SS_{\text{within}}$ and $|\mathbf{B} + \mathbf{W}| = SS_{\text{between}} + SS_{\text{within}}$. Hence, the value of Wilks's Λ is

$$\Lambda = \frac{SS_{\text{within}}}{SS_{\text{between}} + SS_{\text{within}}}. \quad (10)$$

Because the F ratio in a traditionally formulated as

$$F = \frac{SS_{\text{between}}}{SS_{\text{between}} + SS_{\text{within}}}, \quad (11)$$

Wilks's Λ can also be written as

$$\Lambda = \frac{1}{1 + \left[\frac{(k-1)}{(N-k)} \right] F}, \quad (12)$$

where N = the total sample size. This indicates that the relationship between Λ and F is somewhat inverse. The larger the F ratio is, the smaller the Wilks's Λ .

Most computational algorithms for Wilks's Λ take advantage of the fact that Λ can be expressed as a function of the **eigenvalues** of a matrix. Consider Wilks's Λ rewritten as

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} = \frac{1}{|\mathbf{B}\mathbf{W}^{-1} + \mathbf{I}|}. \quad (13)$$

Also consider for any matrix $\mathbf{X}$ there are λ_i eigenvalues, and for a matrix $(\mathbf{X} + \mathbf{I})$ there are $(\lambda_i + 1)$ eigenvalues. In addition, the product of the eigenvalues of a matrix is always equal to the determinant of the matrix (i.e., $\prod \lambda_i = |\mathbf{X}|$). Hence, $\prod (\lambda_i + 1) = |\mathbf{X} + \mathbf{I}|$. Based on this information, the value of Wilks's Λ can be written as the product of the eigenvalues of the matrix $\mathbf{B}\mathbf{W}^{-1}$:

$$\Lambda = \frac{1}{\prod (\lambda_i + 1)}. \quad (14)$$

MANOVA Example

We illustrate MANOVA using as an example one of its earliest applications [13]. In this study, there were five samples, each with 12 members, of aluminum diecastings ($k = 5$, $n_k = 12$, $N = 60$). On each specimen $p = 2$ measurements are taken: tensile strength

(TS , 1000 lb per square inch) and hardness (H , Rockwell's E). The data may be summarized as shown in Table 1.

We wish to test the multivariate null hypothesis of sample equality with the χ^2 approximation to Wilks's Λ . Recall that $\Lambda = |\mathbf{W}|/|\mathbf{B} + \mathbf{W}|$, so $\mathbf{W}$ and $\mathbf{B}$ are needed. First we calculate $\mathbf{W}$. Recognizing that each sample provides an estimate of $\mathbf{W}$, we use a pooled estimate of the within-sample variability for the two variables:

$$\begin{aligned} \mathbf{W} &= \mathbf{W}_1 + \mathbf{W}_2 + \mathbf{W}_3 + \mathbf{W}_4 + \mathbf{W}_5 \\ &= \begin{bmatrix} 78.95 & 214.18 \\ 214.18 & 1247.18 \end{bmatrix} + \begin{bmatrix} 223.70 & 657.62 \\ 657.62 & 2519.31 \end{bmatrix} \\ &\quad + \begin{bmatrix} 57.45 & 190.63 \\ 190.63 & 1241.78 \end{bmatrix} + \begin{bmatrix} 187.62 & 375.91 \\ 375.91 & 1473.44 \end{bmatrix} \\ &\quad + \begin{bmatrix} 88.46 & 259.18 \\ 259.18 & 1171.73 \end{bmatrix} \\ &= \begin{bmatrix} 636.17 & 1697.52 \\ 1697.52 & 7653.44 \end{bmatrix} \end{aligned} \quad (15)$$

The diagonal elements of $\mathbf{B}$ are defined as follows:

$$b_{ii} = \sum_{j=1}^k n_j (\bar{y}_{ij} - \bar{\bar{y}}_i)^2, \quad (16)$$

where n_j is the number of specimens in group j , $\bar{y}_{ij}$ is the mean for variable i in group j , and $\bar{\bar{y}}_i$ is the grand mean for variable i .

The off diagonal elements of $\mathbf{B}$ are defined as follows:

$$b_{mi} = b_{im} = \sum_{j=1}^k n_j (\bar{y}_{ij} - \bar{\bar{y}}_i)(\bar{y}_{mj} - \bar{\bar{y}}_m). \quad (17)$$

Table 1 Data from five samples of aluminum diecastings

Sample	TS		H		TS, H CP _{within}
	Mean	SS _{within}	Mean	SS _{within}	
1	33.40	78.95	68.49	1247.18	214.18
2	28.22	223.70	68.02	2519.31	657.62
3	30.31	57.45	66.57	1241.78	190.63
4	33.15	187.62	76.12	1473.44	375.91
5	34.27	88.46	69.92	1171.73	259.18
	$\bar{\bar{TS}} = 31.87$	$\sum SS_{w_{TS}} = 636.17$	$\bar{\bar{H}} = 69.82$	$\sum SS_{w_H} = 7653.44$	$\sum SS_{w_{TS,H}} = 1697.52$

4 History of Multivariate Analysis of Variance

Now we can find the elements of **B**:

$$\begin{aligned} b_{11} &= 12(33.40 - 31.87)^2 + 12(28.22 - 31.87)^2 \\ &\quad + 12(30.13 - 31.87)^2 + 12(33.15 - 31.87)^2 \\ &\quad + 12(34.27 - 31.87)^2 \\ &= 313.08 \end{aligned}$$

$$\begin{aligned} b_{22} &= 12(68.49 - 69.82)^2 + 12(68.02 - 69.82)^2 \\ &\quad + 12(66.57 - 69.82)^2 + 12(76.12 - 69.82)^2 \\ &\quad + 12(69.92 - 69.82)^2 \\ &= 663.24 \end{aligned}$$

$$\begin{aligned} b_{12} &= 12(33.40 - 31.87)(68.49 - 69.82) \\ &\quad + 12(28.22 - 31.87)(68.02 - 69.82) \\ &\quad + 12(30.13 - 31.87)(66.57 - 69.82) \\ &\quad + 12(33.15 - 31.87)(76.12 - 69.82) \\ &\quad + 12(34.27 - 31.87)(69.92 - 69.82) \\ &= 221.88 \end{aligned} \quad (18)$$

Therefore,

$$\mathbf{B} = \begin{bmatrix} 313.08 & 221.24 \\ 221.24 & 663.24 \end{bmatrix}. \quad (19)$$

Now we can obtain Wilks's Λ :

$$\begin{aligned} \Lambda &= \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|} \\ &= \frac{\begin{vmatrix} 636.17 & 1697.52 \\ 1697.52 & 7653.44 \end{vmatrix}}{\begin{vmatrix} 313.08 & 221.24 \\ 221.24 & 663.24 \end{vmatrix} + \begin{vmatrix} 636.17 & 1697.52 \\ 1697.52 & 7653.44 \end{vmatrix}} \\ &= \frac{1987314.77}{158700.04 + 1987314.77} = 0.93. \end{aligned} \quad (20)$$

Finally, we compute the chi-square approximation to Λ :

$$\chi^2 = -[(60 - 1) - 0.5(2 + 5)] \ln(0.93)$$

$$= -55.5(-0.07)$$

$$= 4.03, \text{ with } 2(5 - 1) = 8 \text{ df, } p = 0.85. \quad (21)$$

We conclude that the five samples have aluminum diecastings of equal tensile strength and hardness.

Wilks's Λ is the oldest and most widely used criterion for comparing groups, but several others have been proposed. Of these, the two most widely used are Hotelling's [9] trace condition and Roy's [20] largest-root criterion. Both of these are functions of the roots $\lambda_1, \lambda_2, \dots, \lambda_r$ of $\mathbf{BW}^{-1}$. Hotelling trace criterion is defined as

$$T = \sum_r \lambda_r, \quad (22)$$

and Roy's largest-root criterion is

$$\theta = \frac{(\lambda_{\max})}{(1 + \lambda_{\max})}. \quad (23)$$

Multivariate Analysis of Covariance (MANCOVA)

Just as ANOVA can be extended to the **analysis of covariance (ANCOVA)**, MANOVA can be extended to testing the equality of group means after their dependence on other variables has been removed by regression. In the multivariate analysis of covariance (MANCOVA), we eliminate the effects of one or more confounding variables (covariates) by regressing the set of dependent variables on them; group differences are then evaluated on the set of residualized means.

Bartlett [3] reported the first published example of a MANCOVA. The paper described an experiment to examine the effect of fertilizers on grain in which eight treatments were applied in each of eight blocks. On each plot of grain two observations were made, the yield of straw (x_1) and the yield of grain (x_2). The results obtained are shown in Table 2.

Differences among the eight blocks were of no interest and, therefore, variability due to blocks was

Table 2 Results from MANOVA examining the effect of fertilizer treatment on straw and grain yield

Source	df	SS_{x_1}	$CP_{x_1 x_2}$	SS_{x_2}
Blocks	7	86 045.8	56 073.6	75 841.5
Treatments	7	12 496.8	-6786.6	32 985.0
Residual	49	136 972.6	58 549.0	71 496.1
Total	63	235 515.2	107 836.0	180 322.6

Table 3 Results from MANCOVA examining the effect of fertilizer treatment on straw and grain yield

Source	df	SS_{x_1}	$CP_{x_1x_2}$	SS_{x_2}
Total	58	149 469.4	51 762.4	104 481.1

removed from the total variability, resulting in a new total (Table 3).

The multivariate null hypothesis of equality among the eight fertilizers was tested using Wilks's Λ :

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{\begin{vmatrix} 136972.6 & 58549.0 \\ 58549.0 & 71496.1 \end{vmatrix}}{\begin{vmatrix} 149469.4 & 51762.4 \\ 51762.4 & 104481.1 \end{vmatrix}} = \frac{7113660653}{7649097476} = 0.49. \quad (24)$$

Next the chi-square approximation to Λ was computed:

$$\begin{aligned} \chi^2 &= -[(56 - 1) - 0.5(2 + 8)] \ln(0.49) \\ &= -50.0(-0.31) \\ &= 15.5, \text{ with } 2(8 - 1) = 14 \text{ df, } p = 0.34. \end{aligned} \quad (25)$$

The conclusion was that the eight fertilizer treatments yielded equal amounts of straw and grain.

References

- [1] Anderson, T.W. (1984). *Introduction to Multivariate Statistical Analysis*, 2nd Edition, Wiley, New York.
- [2] Bartlett, M.S. (1939). A note on tests of significance in multivariate analysis, *Proceedings of the Cambridge Philosophical Society* **35**, 180–185.
- [3] Bartlett, M.S. (1947). Multivariate analysis, *Journal of the Royal Statistical Society B* **9**, 176–197.
- [4] Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika* **36**, 317–346.
- [5] Fisher, R.A. (1912). On the absolute criterion for fitting frequency curves, *Messenger of Mathematics* **41**, 155–160.
- [6] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London Series A* **222**, 309–368.
- [7] Fisher, R.A. (1939). The sampling distribution of some statistics obtained from nonlinear equations, *Annals of Eugenics* **9**, 238–249.
- [8] Girshick, M.A. (1939). On the sampling theory of roots of determinantal equations, *Annals of Mathematical Statistics* **10**, 203–224.
- [9] Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics*, University of California Press, Berkeley, pp. 23–41.
- [10] Hsu, P.L. (1939). On the distribution of roots of certain determinantal equations, *Annals of Eugenics* **9**, 250–258.
- [11] Lee, Y.S. (1972). Some results on the distribution of Wilks's likelihood-ratio criterion, *Biometrika* **59**, 649–664.
- [12] Neyman, J. & Pearson, E. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20**, 175–240, 263–294.
- [13] Pearson, E.S. & Wilks, S.S. (1933). Methods of statistical analysis appropriate for K samples of two variables, *Biometrika* **25**, 353–378.
- [14] Pillai, K.C.S. & Gupta, A.K. (1968). On the non-central distribution of the second elementary symmetric function of the roots of a matrix, *Annals of Mathematical Statistics* **39**, 833–839.
- [15] Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*, Wiley, New York.
- [16] Roy, S.N. (1939). P -statistics or some generalizations in analysis of variance appropriate to multivariate problems, *Sankhyā* **4**, 381–396.
- [17] Roy, S.N. (1942a). Analysis of variance for multivariate normal populations: the sampling distribution of the requisite p -statistics on the null and non-null hypothesis, *Sankhyā* **6**, 35–50.
- [18] Roy, S.N. (1942b). The sampling distribution of p -statistics and certain allied statistics on the non-null hypothesis, *Sankhyā* **6**, 15–34.
- [19] Roy, S.N. (1946). Multivariate analysis of variance: the sampling distribution of the numerically largest of the p -statistics on the non-null hypotheses, *Sankhyā* **8**, 15–52.
- [20] Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*, Wiley, New York.
- [21] Schatzoff, M. (1966). Exact distributions of Wilks' likelihood ratio criterion, *Biometrika*, **53**, 34–358.
- [22] Wilks, S.S. (1932). Certain generalizations in the analysis of variance, *Biometrika* **24**, 471–494.

Further Reading

Wilks, S.S. (1962). *Mathematical Statistics*, Wiley, New York.

SCOTT L. HERSHBERGER

History of Path Analysis

STANLEY A. MULAİK

Volume 2, pp. 869–875

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Path Analysis

History of Path Analysis and Structural Equation Modeling

At the end of the 19th Century British empiricism was the dominant philosophy in Great Britain, having been developed by such philosophers as John Locke (1632–1704), George Berkeley (1685–1753), David Hume (1711, 1776), James Mill (1773–1836) and his son, John Stuart Mill (1806–1873). **Karl Pearson** (1857–1936), trained as a physicist but renowned today as one of the founders of modern, multivariate statistics, with the chi-squared goodness of fit test, the Pearson product moment correlation coefficient, multiple correlation and regression, and with **G. U. Yule**, partial correlation as specific contributions, was also a highly influential empiricist philosopher of science. His *Grammar of Science* [25], published in a series of editions from 1892 to the 1930s popularized and amplified upon the empiricist philosophy of the Austrian physicist Ernst Mach, and was highly influential for a whole generation of scientists. Pearson held that concepts are not about an independent reality but rather are ‘ideal limits’ created by the mind in averaging experience. Scientific laws are but summaries of average results, curves fit to scattered data points, and useful fictions for dealing with experience [20], [22]. Pearson particularly thought of causation as but association through time and regarded statistical correlation as a way of measuring the degree of that association. Deterministic causation was merely one extreme, that of a perfect correlation. Zero correlation was a lack of association and absence of causation. Aware of a shift in physical thought from determinism to probabilistic theories, Pearson further thought of correlation as the proper way to represent probabilistic relationships in science. Pearson also echoed Mach’s skepticism about the reality of atoms and even questioned the new ideas in biology of the gene because these are not given directly in experience. This view was later reinforced by the views of the Austrian physicist, and founding member of the *Vienna Circle* of logical empiricists, Morris Schlick, who declared causality was an outmoded concept in modern physics, a relic of Newtonian determinism, which was giving way to a probabilistic quantum theory.

Pearson’s method of doing research was guided by his belief that this was to be done by forming associations from data. Correlations between variables and multiple correlation were ways of establishing associations between events in experience (*see* **Partial Correlation Coefficients; Multiple Linear Regression**). Pearson did not begin with substantive hypotheses and seek to test these, other than with the assumption that nothing was associated with anything unless shown to be so. So, tests of zero correlation, of independence, were the principal statistical tools, although he could test whether data conformed to a specific probability distribution or not. After he had demonstrated an association in a nonzero correlation, he would then seek to interpret this, but more in a descriptive manner, which summarized the results.

So it was in the face of this empiricist skeptical attitude toward causality that Sewell Wright [33], a young American agricultural geneticist, presented his new statistical methodology of path analysis for the study of causation in agriculture and genetics. He argued that computing correlations between variables does not represent the actual causal nature of the relationship between variables. Causes, he said, are unidirectional, whereas correlations do not represent direction of influence. On the other hand, he held that it was possible to understand correlations between variables in terms of causal relationships between the variables. He then introduced path diagrams to represent these causal relationships. He effectively developed the graphical conventions still used today: (observed) variables are represented by rectangles. Arrows between variables indicate unidirectional causal pathways between them. Curves with arrows at each end between variables indicate correlation between the variables. He envisaged chains of causes, and even considered the possibility of interactive and nonlinear relationships between variables. However, he confined path analysis, as he called his method, to linear causal relationships between variables, although this did not preclude nonlinear causal relationships in general. He then showed how correlations between variables could be shown to arise from common causes between variables. This led to the consideration of systems of correlated causes. Within such systems a variable that is the effect of two causal variables would be represented by an equation $X = M + N$.

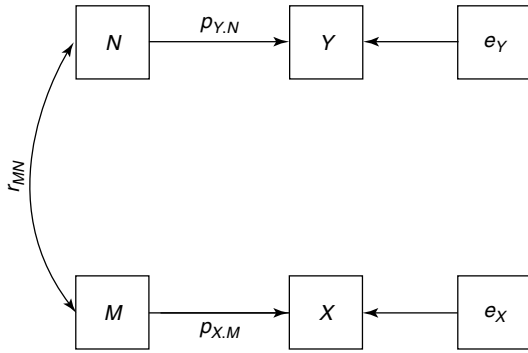
2 History of Path Analysis

Then the variance of X would be given by the equation $\sigma_X^2 = \sigma_M^2 + \sigma_N^2 + 2\sigma_M\sigma_N r_{MN}$, where σ_X^2 is the variance of X , σ_M^2 is the variance of M , σ_N^2 the variance of N , σ_M and σ_N , the standard deviations of M and N , respectively, and r_{MN} the correlation between M and N . He then defined $\sigma_{X \cdot M}$ to be the standard deviation of X when all variables other than X and M are held constant. Holding N constant makes its variance and standard deviation go to zero. Hence the variance in X due to M alone is simply σ_M^2 . Hence $\sigma_{X \cdot M} = \sigma_M$ in this case. Next he defined the quantity $p_{X \cdot M} = \sigma_M / \sigma_X$. This is known as a ‘path coefficient’. From the fact, in this case that $\sigma_X^2 / \sigma_X^2 = \sigma_M^2 / \sigma_X^2 + \sigma_N^2 / \sigma_X^2 + (2 \times \sigma_M / \sigma_X \times \sigma_N / \sigma_X \times r_{MN})$, we may arrive at the total variance of X in standardized score form as $1 = p_{X \cdot M}^2 + p_{X \cdot N}^2 + 2p_{X \cdot M}p_{X \cdot N}r_{MN}$.

On the other hand, suppose we have two effect variables X and Y . Suppose, by way of a simplified representation of Wright’s exposition, there is an equation for each effect variable:

$$\begin{aligned} X &= p_{X \cdot M}M + e_X \\ Y &= p_{Y \cdot N}N + e_Y. \end{aligned} \quad (1)$$

These equations could be represented by the path diagram:



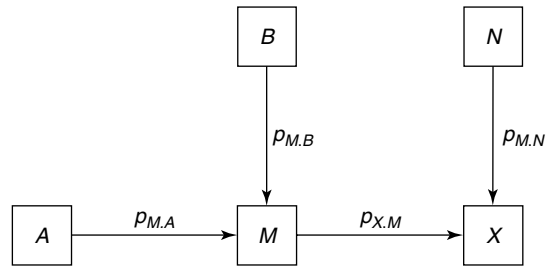
Further, suppose X, Y, M and N have unit variances and that we observe the correlations r_{XY} , r_{XM} , r_{XN} , r_{YM} , r_{YN} , and r_{MN} . From these equations and assuming e_X, e_Y are mutually uncorrelated and further each are uncorrelated with M , and N , then we can derive the following:

$$\begin{aligned} r_{XY} &= p_{X \cdot M}p_{Y \cdot N}r_{MN} \\ r_{XM} &= p_{X \cdot M} \end{aligned}$$

$$\begin{aligned} r_{XN} &= p_{X \cdot M}r_{MN} \\ r_{YM} &= p_{Y \cdot N}r_{MN} \\ r_{YN} &= p_{Y \cdot N}. \end{aligned} \quad (2)$$

Immediately we see that the parameters $p_{X \cdot M} = r_{XM}$ and $p_{Y \cdot N} = r_{YN}$ are given in terms of observable quantities. Since all other correlations are given in terms of these same parameters and the observed r_{MN} , we now have a way to test whether these equations represent the causal relationships between the variables by comparing them to the actual observed correlations. We see, for example, that $r_{XY} = r_{XM}r_{YN}r_{MN}$. If this is not so, then something is wrong with this representation of the causes of X and Y . Again, we should observe that $r_{XN} = r_{XM}r_{MN}$ and $r_{YM} = r_{YN}r_{MN}$. If none of these is the case, something is wrong with these equations as a causal model for these variables. Effectively, then, Wright saw that specifying a causal structure for the variables determined their intercorrelations, and further that the causal parameters of the structure could be estimated from the correlations among the variables. Finally, if certain correlations between the variables did not equal the values predicted by the equations and the parameter estimates, then this was evidence against the structure. So, his method of path analysis would allow one to test one’s prior knowledge or beliefs about the causal structure among the observed variables.

Consider another example involving a simple chain of causes.



The path diagram shows that B and N are uncorrelated. Furthermore, A is uncorrelated with B and M is uncorrelated with N . We will again assume that the variances of all variables are unity and have zero means. This gives rise to the equations for the effect variables:

$$\begin{aligned} X &= p_{X \cdot M}M + p_{X \cdot N}N \\ M &= p_{M \cdot A}A + p_{M \cdot B}B. \end{aligned} \quad (3)$$

Wright declared that in this case the effect of the more remote variable A on X was given by $p_{X \cdot A} = p_{M \cdot A}p_{X \cdot M}$, the product of the path coefficients along the path connecting A to X . We also can see how this model of the causes could be tested. From the two equations, we can further derive the hypothetical correlations among these variables as

$$\begin{aligned} r_{XM} &= p_{X \cdot M} & r_{AB} &= 0 \\ r_{AM} &= p_{M \cdot A} & r_{AN} &= 0 \\ r_{AX} &= p_{M \cdot A}p_{X \cdot M} & r_{MN} &= 0 \\ r_{BM} &= p_{M \cdot B} & r_{BN} &= 0 \\ r_{XN} &= p_{X \cdot N} & r_{BX} &= p_{M \cdot B}p_{X \cdot M} \end{aligned} \quad (4)$$

We see now that if this model holds, then $r_{AX} = r_{XM}r_{AM}$ must hold. This provides a test of the model when observed correlations are used. The variables B and N may also be unobserved error variables, and so the only observed correlations are between the variables A , M , and X . The remaining correlations are then given only by hypothesis.

Wright also showed remarkable prescience in considering the analysis of direct and indirect effects, of common causes, and common effects of several causes, and the effect of unmeasured ‘relevant’ causes of an effect variable that were also correlated with its other causes that were the focus of study.

When Wright’s article was published it was subsequently followed by a critical article by Niles [23] who quoted extensively Pearson’s *The Grammar of Science* (1900). According to Pearson, Niles held, correlation *is* causation. To contrast causation with correlation is unwarranted. There is no philosophical basis on which to extend to the concept of cause a wider meaning than partial or absolute association. Furthermore Niles could not see how you could study the causes of any specific variables, for, according to Pearson the causes of any part of the universe lead inevitably to the history of the universe as a whole. In addition there is, he held, no way to specify *a priori* the true system of causes among variables, that to do so implied that causation is a necessary connection between things and further that it is different from correlation. Furthermore even if a hypothesized system conforms to the observed correlations, this did not imply that it is the true system, for there could be infinitely many different equivalent systems created *a priori* to fit the same variables.

Wright responded with a rebuttal [34], arguing that his method was not a deduction of causes from correlations, but the other way around. Furthermore, he did not claim “that ‘finding the logical consequences’ of a hypothesis in regard to the causal relations depended on any prior assumption that the hypothesis is correct” (p. 241). If the hypothesized consequences do not conform to the observed correlations, then this allows us to regard the hypothesized system as untenable and in need of modification. If the hypothesized consequences correspond to independently obtained results, then this demonstrates the ‘truth’ of the hypothesis ‘in the only sense which can be ascribed to the truth of a natural law’ (p. 241). Niles followed with another attempted rebuttal (1923). But Wright went on to develop the method extensively in studying models of the heredity of traits as gene transfers between parents and offspring that manifest themselves in correlations between parents, offspring, and relatives [16, 35].

Path analysis at this point was not taken up by the behavioral or social sciences. Factor analysis at this point was a competing, well-established methodology for working with correlations, and was being used to study intelligence [29], [30] and personality. But even though it was a structural model, the exploratory factor analysis model then in use was applied in a manner that regarded all correlations as due just to common factors. But as all beginning statistics students are told, a correlation between two variables X and Y can be due to X being a cause of Y , Y being a cause of X , or there being a third variable Z that is a common cause of both X and Y . So, factor analysis only considers one of these as the causal structure for explaining all correlations. And in a purely exploratory mode it was used often by researchers without much prior consideration of what even the common causes might be. So, it was closer in research style to Pearson’s descriptive use of regression, where one automatically applied the model, and then described and summarized the results rather than as a model-testing method. But the path analysis models of Wright could consider each kind of causation in formulating causal models of correlations between variables, but these models were formulated prior to the analysis and the causal structures were given by hypothesis rather than summaries of results. The important thing was whether the pattern of correlations predicted by the model conformed to the observed pattern of correlations

among the variables. So, path analysis was a model-testing method.

Another reason that likely retarded the uptake of path analysis into the behavioral and social sciences at the outset was that this was a method used in genetic research, published in genetics and biological journals, so the technique was little known to researchers in the behavioral and social sciences for many years.

By a different route, the econometricians began implementing regression models and then extended these to a method mathematically equivalent to path analysis known as structural equation modelling. This was initially stimulated by such mathematical models of the economy as formulated by John Maynard Keynes [17], which used sets of simultaneous linear equations to specify relations between variables in the economy. The econometricians distinguished *exogenous variables* (inputs into a system of variables) from *endogenous variables* (variables that are dependent on other variables in the system). Econometricians also used matrix algebra to express their model equations. They sought further to solve several problems, such as determining the conditions under which the free parameters of their models would be identified, that is, determinable uniquely from the observed data [18]. They showed how the endogenous variables could ultimately be made to be just effects of the exogenous variables, given in the 'reduced equations'. They developed several new methods of parameter estimation such as two-stage [32] and three-stage least squares [36]. They developed both Full Information Maximum Likelihood (FIML) and Limited Information Maximum likelihood (LIML) estimates of unspecified parameters (*see Maximum Likelihood Estimation*). However, generally their models involved only measured observed variables.

Although in the 1950s logical empiricism reigned still as the dominant philosophy of science and continued to issue skeptical critiques of the idea of causation as an out-dated remnant of determinism, or to be replaced by a form of logical implication, several philosophers sought to restore causality as a central idea of science. Bunge [7] issued a significant book on causality. Simon [26] argued that causality is to be understood as a functional relation between variables, not a relation between individual events, like logical implication. This laid the groundwork for what followed in sociology and, later, psychology.

Blalock [3], a sociologist who had been originally trained in mathematics and physics, authored a highly influential book in sociology that drew upon the method of path analysis of Wright [35]. Blalock [4] also edited a collection of key articles in the study of causation in the social sciences, which was highly influential in the treatment of causality, its detection, and in providing research examples. A second edition [5] also provided newer material. This was also accompanied by a second volume [6] devoted to issues of detecting causation with experimental and panel designs. Duncan [8] wrote an influential introductory textbook on structural equation models for sociologists. Heise [10] also authored an important text on how to study causes with flowgraph analysis, a variant of path analysis.

A highly important development began in the latter half of the 1960s in psychology. Bock and Bargmann [2] described a new way of testing hypotheses about linear functional relations known as 'analysis of covariance structures'. This was followed up in the work of Karl Jöreskog, a Swedish mathematical statistician, who came to Educational Testing Service to work on problems of factor analysis. After solving the problem of finding a full information maximum likelihood estimation method for exploratory common factor analysis [12] (*see Factor Analysis: Exploratory*), Jöreskog turned his attention to solving a similar problem for confirmatory factor analysis [13] (*see Factor Analysis: Confirmatory*), which prior to that time had received little attention among factor analysts. This was followed by an even more general model that he called 'analysis of covariance structures' [14]. Collaboration with Arthur S. Goldberger led Karl Jöreskog to produce an algorithm for estimating parameters and testing the fit of a structural equation model with latent variables [15], which combined concepts from factor analysis with those of structural equations modeling. He was also able to provide for a distinction between free, fixed, and constrained parameters in his models. But of greatest importance for the diffusion of his methods was his making available computer programs for implementing the algorithms described in his papers. By showing that confirmatory factor analysis, analysis of covariance structures, and structural equation modeling could all be accomplished with a single computer program, this provided researchers with a general, highly flexible method for studying a great variety of

linear causal structures. He called this program 'LISREL' for 'linear structural relations'. It has gone through numerous revisions. But his program was shortly followed by others, which sought to simplify the representation of **structural equation models**, such as COSAN, [19], EQS [1] and several others (see **Structural Equation Modeling: Software**). Numerous texts, too many to mention here, followed, based on Jöreskog's breakthrough. A new journal, *Structural Equation Modeling* appeared in 1994.

The availability of easy-to-use computer programs for doing structural equation modeling in the 1980s and 1990s produced almost a paradigm shift in correlational psychological research from descriptive studies to testing causal models and renewed investigations of the concept of causality and the conditions under which it may be inferred. James, Mulaik, and Brett [11] sought to remind psychological researchers that structural equation modeling is not exploratory research, and that, in designing their studies, they needed to focus on establishing certain conditions that facilitated inferences of causation as opposed to spurious causes. Among these was the need to make a formal statement of the substantive theory underlying a model, to provide a theoretical rationale for causal hypotheses, to specify a causal order of variables, to establish self-contained systems of structural equations representing all relevant causes in the phenomenon, to specify boundaries such as the populations and environments to which the model applies, to establish that the phenomenon had reached an equilibrium condition when measurements were taken, to properly operationalize the variables in terms of conditions of measurement, to confirm empirically support for the functional equations in the model, and to confirm the model empirically in terms of its overall fit to data.

Mulaik [21] provided an amplified account first suggested by Simon in 1953 [28] of how one might generalize the concept of causation as a functional relation between variables to the probabilistic case. Simon had written "...we can replace the causal ordering of the variables in the deterministic model by the assumption that the realized values of certain variables at one point or period in time determines the probability distribution of certain variables at later points in time" [27, 1977, p. 54]. This allows one to join linear structural equation modeling with other, nonlinear forms of probabilistic causation, such as item-response theory.

Four philosophers of science [9] put forth a description of a method for discovering causal structure in correlations based on an artificial intelligence algorithm that implemented heuristic searches for certain zero partial correlations between variables and/or zero tetrad differences among correlations [29] that implied certain causal path structures. Their approach combined graph theory with artificial intelligence search algorithms and statistical tests of vanishing partial correlations and vanishing tetrad differences. They also produced a computer program for accomplishing these searches known as Tetrad. In a brief history of heuristic search in applied statistics, they argued that researchers had abandoned an optimal approach to testing causal theories and discovering causal structure first suggested by Spearman's (1904) use of tetrad difference tests, by turning to a less optimal approach in factor analysis. The key idea was that instead of estimating parameters and then checking the fit of the reproduced covariance matrix to the observed covariance matrix, and then, if the fit was poor, taking another factor with associated loadings to estimate, as in factor analysis, Spearman had identified constraints implied by a causal model on the elements of the covariance matrix, and sought to test these constraints directly. Generalizing from this, Glymour et al. [9] showed how one could search for those causal structures having the greatest number of constraints implying vanishing partial correlations and vanishing tetrad differences on the population covariance matrix for the variables that would be most consistent with the sample covariance matrix. The aim was to find a causal structure that would apply regardless of the values of the model parameters.

Spirtes, Glymour, and Scheines [31] followed the previous work with a book that went into considerable detail to show how probability could be connected with causal graphs. To do this, they considered that three conditions were needed for this: the Causal Markov Condition, the Causal Minimality Condition, and the Faithfulness Condition. Kinship metaphors were used to identify certain sets of variables. For example, the parents of a variable V would be all those variables that are immediate causes of the variable V represented by 'directed edges' of a graph leading from these variables to the variable in question. The descendants would be all those variables that are in directed paths from V . A directed acyclic

graph for a set of variables $\mathbf{V}$ and a probability distribution would be said to satisfy the Markov Condition if and only if for every variable W in $\mathbf{V}$, W is independent of all variables in $\mathbf{V}$ that are neither parents nor descendants of W conditional on the parents of W . Satisfying the Markov Condition allows one to specify conditional independence to occur between certain sets of variables that could be represented by vanishing partial correlations between the variables in question, conditional on their parents. This gives one way to perform tests on the causal structure without estimating model parameters. Spirtes, Glymour, and Scheines [31] showed how from these assumptions one could develop discovery algorithms for causally sufficient structures. Their book was full of research examples and advice on how to design empirical studies. A somewhat similar book by [24], because Spirtes, Glymour, and Scheines [31] drew upon many of Pearl's earlier works, attempted to restore the study of causation to a prominent place in scientific thought by laying out the conditions by which causal relations could be and not be established between variables. Both of these works differ in emphasizing tests of conditional independence implied by a causal structure rather than tests of fit of an estimated model to the data in evaluating the model.

References

- [1] Bentler, P.M. & Weeks, D.G. (1980). Linear structural equations with latent variables, *Psychometrika* **45**, 289–308.
- [2] Bock, R.D. & Bargmann, R.E. (1966). Analysis of covariance structures, *Psychometrika* **31**, 507–534.
- [3] Blalock Jr, H.M. (1961). *Causal Inferences in Nonexperimental Research*, University of North Carolina Press, Chapel Hill.
- [4] Blalock Jr, H.M. (1971). *Causal Models in the Social Sciences*, Aldine-Atherton, Chicago.
- [5] Blalock Jr, H.M. (1985a). *Causal Models in the Social Sciences*, 2nd Edition, Aldine-Atherton, Chicago.
- [6] Blalock, H.M. (ed.) (1985b). *Causal Models in panel and experimental designs*, Aldine, New York.
- [7] Bunge, M. (1959). *Causality*, Harvard University Press, Cambridge.
- [8] Duncan, O.D. (1975). *Introduction to Structural Equation Models*, Academic Press, New York.
- [9] Glymour, C., Scheines, R., Spirtes, P. & Kelly, K. (1987). *Discovering Causal Structure*, Academic Press, Orlando.
- [10] Heise, D.R. (1975). *Causal Analysis*, Wiley, New York.
- [11] James, L.R., Mulaik, S.A. & Brett, J.M. (1982). *Causal Analysis: Assumptions, Models, and Data*, Sage Publications and the American Psychological Association Division 14, Beverly Hills.
- [12] Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis, *Psychometrika* **32**, 443–482.
- [13] Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis, *Psychometrika* **34**, 183–202.
- [14] Jöreskog, K.G. (1970). A general method for the analysis of covariance structures, *Biometrika* **57**, 239–251.
- [15] Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, Eds, Seminar Press, New York, pp. 85–112.
- [16] Kempthorne, O. (1969). *An Introduction to Genetic Statistics*, Iowa State University Press, Ames.
- [17] Keynes, J.M. (1936). *The General Theory of Employment, Interest and Money*, Macmillan, London.
- [18] Koopmans, T.C. (1953). Identification problems in economic model construction, in *Studies in Econometric Method*, Cowles Commission Monograph 14, W.C. Hood & T.C. Koopmans, eds, Wiley, New York.
- [19] McDonald, R.P. (1978). A simple comprehensive model for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **31**, 59–72.
- [20] Mulaik, S.A. (1985). Exploratory statistics and empiricism, *Philosophy of Science* **52**, 410–430.
- [21] Mulaik, S.A. (1986). Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept, *Philosophy of Science* **53**, 313–332.
- [22] Mulaik, S.A. (1987). A brief history of the philosophical foundations of exploratory factor analysis, *Multivariate Behavioral Research* **22**, 267–305.
- [23] Niles, H.E. (1922). Correlation, causation and Wright's theory of 'path coefficients', *Genetics* **7**, 258–273.
- [24] Pearl, J. (2000). *Causality: Models, reasoning and inference*, Cambridge University Press, Cambridge.
- [25] Pearson, K. (1892). *The Grammar of Science*, Adam & Charles Black, London.
- [26] Simon, H.A. (1952). On the definition of the causal relation, *Journal of Philosophy* **49**, 517–528.
- [27] Simon, H.A. (1953). Causal ordering and identifiability, in *Studies in Econometric Method*, W.C. Hood & T.C. Koopmans, eds, Wiley, New York, pp. 49–74.
- [28] Simon, H.A. (1977). *Models of Discovery*, R. Reidel, Dordrecht, Holland.
- [29] Spearman, C. (1904). General intelligence objectively determined and measured, *British Journal of Psychology* **15**, 201–293.
- [30] Spearman, C. (1927). *The Abilities of Man*, MacMillan, New York.
- [31] Spirtes, P., Glymour, C. & Scheines, R. (1993). *Causation, Prediction and Search*, Springer-Verlag, New York.
- [32] Theil, H. (1953). *Estimation and Simultaneous Correlation in Complete Equation Systems*, Central Planbureau (mimeographed), The Hague.

-
- [33] Wright, S. (1921). Correlation and causation, *Journal of Agricultural Research* **20**, 557–585.
- [34] Wright, S. (1923). The theory of path coefficients: A reply to Niles' criticism, *Genetics* **8**, 239–255.
- [35] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.
- [36] Zellner, A. & Theil, H. (1962). Three-stage least squares: simultaneous estimation of simultaneous equations, *Econometrica* **30**, 54–78.

Further Reading

- Wright, S. (1931). Statistical methods in biology, *Journal of the American Statistical Association* **26**, 155–163.

STANLEY A. MULAİK

History of Psychometrics

RODERICK D. BUCHANAN AND SUSAN J. FINCH

Volume 2, pp. 875–878

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Psychometrics

Introduction

Psychometrics can be described as the science of measuring psychological abilities, attributes, and characteristics. Such a ubiquitous and hybridized set of techniques has been said, not surprisingly, to have many protoscientific and professional antecedents, some dating back to antiquity. Modern psychometrics is embodied by standardized psychological tests. American psychometrician Lee Cronbach famously remarked in the 1960s, ‘the general mental test ... stands today as the most important single contribution of psychology to the practical guidance of human affairs’ [16, p. 113]. However, psychometrics has come to mean more than just the tests themselves; it also encompasses the mathematical, statistical, and professional protocols that underpin tests – how tests are constructed and used, and indeed, how they are evaluated.

Early Precedents

Historians have noted the examples of ‘mental testing’ in ancient China and other non-Western civilizations where forms of proficiency assessment were used to grade or place personnel. However, the most obvious template for psychometric assessment, with a more direct lineage to modern scientific manners, was the university and school examination. Universities in Europe first started giving formal oral assessments to students in the thirteenth century. With the invention of paper, the Jesuits introduced written examinations during the sixteenth century. In England, competitive university examinations began in Oxbridge institutions in the early 1800s [14]. By the end of the nineteenth century, compulsory forms of education had spread throughout much of the Western world. Greater social mobility and vocational streaming set the scene for practical forms of assessment as governments, schools, and businesses of industrialized nations began to replace their reliance in personal judgment with a trust in the impartial authority of numbers [12].

Enter Darwin

Darwinian thought was a key example of the challenge of scientific materialism in the nineteenth century. If humans were a part of nature, then they were subject to natural law. The notion of continuous variation was central to the new evolutionary thought. Coupled with an emerging notion of personhood as a relatively stable, skin-bound entity standing apart from professional function and social worth, Darwin’s ideas paved the way for measurement-based psychology. Late in the nineteenth century, Darwin’s cousin **Francis Galton** articulated key ideas for modern psychometrics, particularly the focus on human variation. The distribution of many physical attributes (e.g., height) had already been shown by **Quetelet** to approximate a Gaussian curve (*see Catalogue of Probability Density Functions*). Galton suggested that many psychological characteristics would show similar distributional properties. As early as 1816, Bessel had described ‘personal equations’ of systematic individual differences in astronomical observations. In contrast, some of the early psychologists of the modern era chose to ignore these types of differences. For instance, Wundt focused on common or fundamental mechanisms by studying a small number of subjects in-depth. Galton shifted psychologists’ attention to how individuals differed and by how much [8, 14].

Mental Testing Pioneers

Galton’s work was motivated by his obsession with eugenics. Widespread interest in the riddles of **heredity** provided considerable impetus to the development of psychometric testing. If many psychological properties were at least partly innate and inherited, then, arguably, it was even more important and useful to measure them. Galton was especially interested in intellectual functioning. By the mid-1880s, he had developed a diverse range of what (today) seem like primitive measures: tests of physical strength and swiftness, visual acuity and memory of forms. Galton was interested in how these measures related to each other, whether scores taken at an early age might predict later scientific or professional eminence, and whether eminence passed from one generation to the next. These were questions of agreement that were never going to be perfect. Galton needed an index

to calibrate the probabilistic rather than the deterministic relationship between two variables. He used **scatterplots** and noticed how scores on one variable were useful for predicting the scores on another and developed a measure of the ‘correlation’ of two sets of scores. His colleague, the biometric statistician **Karl Pearson** formalized and extended this work. Using the terms ‘normal curve’ and ‘standard deviation’ from the mean, Pearson developed what would become the statistical building blocks for modern psychometrics (e.g., the product-moment correlation (*see* **Pearson Product Moment Correlation**), multiple correlation (*see* **Multiple Linear Regression**), biserial correlation (*see* **Point Biserial Correlation**) [8, 13].

By the turn of the twentieth century, James Cattell and a number of American psychologists had developed a more elaborate set of anthropometric measures, including tests of reaction time and sensory acuity. Cattell was reluctant to measure higher mental processes, arguing these were a result of more basic faculties that could be measured more precisely. However, Cattell’s tests did not show consistent relationships with outcomes they were expected to, like school grades and later professional achievements. Pearson’s colleague and rival **Charles Spearman** argued this may have been due to the inherent unreliability of the various measures Cattell and others used. Spearman reasoned that any test would inevitably contain measurement error, and any correlation with other equally error-prone tests would underestimate the true correlation. According to Spearman, one way of estimating the measurement error of a particular test was to correlate the results of successive administrations. Spearman provided a calculation that corrected for this ‘attenuation’ due to ‘accidental error,’ as did William Brown independently, and both gave proofs they attributed to **Yule**. Calibrating measurement error in this way proved foundational. Spearman’s expression of the correlation of two composite measures in terms of their variance and covariance later became known as the ‘index of reliability’ [9].

Practical Measures

The first mental testers lacked effective means for assessing the qualities they were interested in. In France in 1905, Binet introduced a scale that provided a different kind of measurement. Binet did not

attempt to characterize intellectual processes; instead he assumed that performance on a uniform set of tasks would constitute a basis for a meaningful ranking of school children’s ability. Binet thought it necessary to sample complex mental functions, since these most resembled the tasks faced at school and provided for a maximum spread of scores [15].

Binet did not interpret his scale as a measure of innate intelligence; he insisted it was only a screening device for children with special needs. However, Goddard and many other American psychologists thought Binet’s test reflected a general factor in intellectual functioning and also assumed this was largely hereditary. Terman revised the Binet test just prior to World War II, paying attention to relevant cultural content and documenting the score profiles of various American age groups of children. But Terman’s revision (called the *Stanford–Binet*) remained an age-referenced scale, with sets of problems or ‘items’ grouped according to age appropriate difficulty, yielding an intelligence quotient mental age/chronological age (IQ) score.

Widespread use of Binet-style tests in the US army during World War I helped streamline the testing process and standardize its procedures. It was the first large-scale deployment of group testing and multiple-choice response formats with standardized tests [6, 16].

Branching Out

In the 1920s, criticism of interpretation of the Army test data – that the average mental age of soldiers, a large sample of the US population, was ‘below average’ – drew attention to the problem of appropriate ‘normative’ samples that gave meaning to test scores. The innovations of the subsequent Wechsler intelligence scales – with test results compared to a representative sample of adult scores – could be seen as a response to the limitations of younger age-referenced Binet tests. The interwar period also saw the gradual emergence of the concept of ‘validity,’ that is, whether the test measured what it was supposed to. Proponents of Binet-style tests wriggled out of the validity question with a tautology: intelligence was what intelligence tests measured. However, this stance was developed more formally as operationism, a stopgap or creative solution (depending on your point of view) to the problem of quantitative ontology. In the mid-1930s, **S. S. Stevens** argued that

the theoretical meaning of a psychological concept could be defined by the operations used to measure it, which usually involved the systematic assignment of number to quality. For many psychologists, the operations necessary to transform a concept into something measurable were taken as *producing* the concept itself [11, 14, 18].

The practical success of intelligence scales allowed psychologists to extend operationism to various interest, attitude, and personality measures. While pencil-and-paper questionnaires dated back to at least Galton's time, the new branch of testing appearing after World War I took up the standardization and group comparison techniques of intelligence scales. Psychologists took to measuring what were assumed to be dispositional properties that differed from individual to individual not so much in quality but in amount. New tests of personal characteristics contained short question items sampling seemingly relevant content. Questions usually had fixed response formats, with response scores combined to form additive, linear scales. Scale totals were interpreted as a quantitative index of the concept being measured, calibrated through comparisons with the distribution of scores of normative groups. Unlike intelligence scales, responses to interest, attitude, or personality inventory items were not thought of as unambiguously right or wrong – although different response options usually reflected an underlying psychosocial ordering. Ambiguous item content and poor relationships with other measures saw the first generation of personality and interest tests replaced by instruments where definitions of what was to be measured were largely determined by reference to external criteria. For example, items on the Minnesota Multiphasic Personality Inventory were selected by contrasting the responses of normal and psychiatric subject groups [3, 4].

Grafting on Theoretical Respectability

In the post World War II era, psychologists subtly modified their operationist approach to measurement. Existing approaches were extended and given theoretical rationalizations. The factor analytic techniques (*see* **Factor Analysis: Exploratory**) that Spearman, Thurstone, and others had developed and refined became a mathematical means to derive

latent concepts (*see* **Latent Variable**) from more directly measured variables [1, 10]. They also played a role in guaranteeing both the validity and reliability of tests, especially in the construction phase. Items could be selected that apparently measured the same underlying variable. Several key personality and attitude scales, such as the **R.B. Cattell's** 16 PF and Eysenck's personality questionnaires, were developed primarily using factor analysis. **Thurstone** used factor analysis to question the unitary concept of intelligence. New forms of item analyses and scaling (e.g., indices of item difficulty, discrimination, and consistency) also served to guide the construction of reliable and valid tests.

In the mid-1950s, the American Psychological Association stepped in to upgrade all aspects of testing, spelling out the empirical requirements of a 'good' test, as well as extending publishing and distribution regulations. They also introduced the concept of 'construct validity,' the test's conceptual integrity borne out by its theoretically expected relationships with other measures. Stung by damaging social critiques of cultural or social bias in the 1960s, testers further revived the importance of theory to a historically pragmatic field. Representative content coverage, relevant, and appropriate predictive criteria, all became keystones for fair and valid tests [5, 14].

The implications of Spearman's foundational work were finally formalized by Gulliksen in 1950, who spelt out the assumptions the classical 'true score model' required. The true score model was given a probabilistic interpretation by Lord and Novick in 1968 [17]. More recently, psychometricians have extended item level analyses to formulate generalized response models. Proponents of item response theory claim it enables the estimation of latent aptitudes or attributes free from the constraints imposed by particular populations and item sets [2, 7].

References

- [1] Bartholomew, D.J. (1995). Spearman and the origin and development of factor analysis, *British Journal of Mathematical and Statistical Psychology* **48**, 211–220.
- [2] Bock, R.D. (1997). A brief history of item response theory, *Educational Measurement: Issues and Practice* **16**, 21–33.
- [3] Buchanan, R.D. (1994). The development of the MMPI, *Journal of the History of the Behavioral Sciences* **30**, 148–161.

4 History of Psychometrics

- [4] Buchanan, R.D. (1997). Ink blots or profile plots: the Rorschach versus the MMPI as the right tool for a science-based profession, *Science, Technology and Human Values* **21**, 168–206.
- [5] Buchanan, R.D. (2002). On *not* ‘Giving Psychology Away’: the MMPI and public controversy over testing in the 1960s, *History of Psychology* **5**, 284–309.
- [6] Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*, Cambridge University Press, Cambridge.
- [7] Embretson, S.E. (1996). The new rules of measurement, *Psychological Assessment* **8**, 341–349.
- [8] Gillham, N.W. (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*, Oxford University Press, New York.
- [9] Levy, P. (1995). Charles Spearman’s contributions to test theory, *British Journal of Mathematical and Statistical Psychology* **48**, 221–235.
- [10] Lovie, A.D. & Lovie, P. (1993). Charles Spearman, Cyril Burt, and the origins of factor analysis, *Journal of the History of the Behavioral Sciences* **29**, 308–321.
- [11] Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept*, Cambridge University Press, Cambridge.
- [12] Porter, T.M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton.
- [13] Porter, T.M. (2004). *The Scientific Life in a Statistical Age*, Princeton University Press, Princeton.
- [14] Rogers, T.B. (1995). *The Psychological Testing Enterprise: An Introduction*, Brooks/Cole, Pacific Grove.
- [15] Rose, N. (1979). The psychological complex: mental measurement and social administration, *Ideology and Consciousness* **5**, 5–68.
- [16] Sokal, M.M. (1987). *Psychological Testing and American Society, 1890–1930*, Rutgers University Press, New Brunswick.
- [17] Traub, R.E. (1997). Classical test theory in historical perspective, *Educational Measurement: Issues and Practice* **16**, 8–14.
- [18] Wright, B.D. (1997). A history of social science measurement, *Educational Measurement: Issues and Practice* **16**, 33–52.

(See also **Measurement: Overview**)

RODERICK D. BUCHANAN AND SUSAN
J. FINCH

History of Surveys of Sexual Behavior

BRIAN S. EVERITT

Volume 2, pp. 878–887

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

History of Surveys of Sexual Behavior

My own belief is that there is hardly anyone whose sexual life, if it were broadcast, would not fill the world at large with surprise and horror.

Somerset Maugham

Survey research (*see* **Survey Questionnaire Design**) is largely a product of the twentieth century, although there are some notable exceptions. In the last decade of the nineteenth century, for example, Charles Booth, a successful businessman and dedicated conservative, sought accurate data on the poor of London after becoming disturbed by a socialist claim that a third of the people in the city were living in poverty. But it is only in the past 70 to 80 years that survey research has become firmly established, particularly as market research, opinion polling, and election research. Among the factors that brought surveys into favour was the change from speculation to empiricism in social psychology and sociology – the demand that ‘hunches’ must be backed by numerical evidence, that is data.

Sample surveys provide a flexible and powerful approach to gathering information, but careful consideration needs to be given to various aspects of the survey if the information collected is to be accurate, particularly when dealing with a sensitive topic such as sexual behavior. If such surveys *are* to be taken seriously as a source of believable material a number of issues must be addressed, including;

- Having a sample that is truly representative of the population of interest. Can the sample be regarded as providing the basis for inferences about the target population? A biased selection process may produce deceptive results.
- Having a large enough sample to produce reasonably precise estimates of the prevalence of possibly relatively rare behaviors,
- Minimizing nonresponse. Nonresponse can be a thorny problem for survey researchers. After carefully designing a study, deciding on an appropriate sampling scheme, and devising an acceptable questionnaire, researchers often quickly discover that human beings can be cranky creatures; many of the potential respondents will not be at home (even after making an appointment for a specified

time), or will not answer the telephone, or have moved away, or refuse to reply to mail shots, and so generally make the researcher’s life difficult. In many large-scale surveys, it may take considerable effort and resources to achieve a response rate even as high as 50%. And nonresponse often leads to biased estimates.

- The questions asked. Do the questions illicit accurate responses? Asking questions that appear judgmental can affect the way people answer. The wording of questions by the interviewer or on the questionnaire is critical. Everyday English, as used in colloquial speech, is often ambiguous. For surveys, definitions of terms need to be precise to measure phenomena accurately. At the same time, the terms should be easily understood – technical terms should be avoided. This is not always easy because there are few terms that are universally understood. This is particularly true in surveys of sexual behavior. The meaning of terms such as ‘vaginal sex’, ‘oral sex’, ‘penetrative sex’ and ‘heterosexual’, for example, is taken for granted in much health education literature, but there is evidence that much misunderstanding of such terms exists in the general public.
- Are people likely to be truthful in their answers? Systematic distortion of the respondent’s true status clearly jeopardizes the validity of survey measurements. This problem has been shown even in surveys of relatively innocuous subject matter, owing in part to a respondent’s perceptions and needs that emerge during the data collection process. Consequently the potential for distortion to cause problems in surveys of sensitive information is likely to be considerable due to heightened respondent concern over anonymity. Of course, a person’s sex life is very likely to be a particularly sensitive issue. The respondents need to be assured about confidentiality and in face-to-face interviews the behavior of the interviewer might be critical.

In the end the varying tendencies among respondents to cooperate in surveys (particularly sex surveys), or to under-report/overreport if they respond, can easily lead to wildly inaccurate estimates of the extent of sensitive phenomena. There *are* techniques to collect sensitive information that largely remove the problem of under or over reporting by introducing an element of chance into the responses. These

2 History of Surveys of Sexual Behavior

techniques disguise the true response yet allow the researcher sufficient data for analysis. The most common of these techniques is the **randomized response** approach but there is little evidence of its use in the vast majority of investigations into human sexual behavior.

Surveys of Sexual Behavior

The possibility that women might enjoy sex was not considered by the majority of our Victorian ancestors. The general Victorian view was that women should show no interest in sex and preferably be ignorant of its existence unless married; then they must submit to their husbands without giving any sign of pleasure. A lady was not even supposed to be interested in sex, much less have a sexual response. (A Victorian physician, Dr. Acton, even went as far as to claim ‘It is a vile aspersion to say that women were ever capable of sexual feelings’.) Women were urged to be shy, blushing, and genteel. As Mary Shelley wrote in the early 1800s, ‘Coarseness is completely out of fashion.’ (Such attitudes might, partially at least, help explain both the increased interest in pornography amongst Victorian men and the parallel growth in the scale of prostitution.)

But in a remarkable document written in the 1890s by Clelia Mosher, such generalizations about the attitudes of Victorian women to matters sexual are thrown into some doubt, at least for a minority of women. The document, *Study of the Physiology and Hygiene of Marriage*, opens with the following introduction;

In 1892, while a student in biology at the University of Wisconsin, I was asked to discuss the marital relation in a Mother’s Club composed largely of college women. The discussion was based on replies given by members to a questionnaire.

Mosher probed the sex lives of 45 Victorian women by asking them whether they liked intercourse, how often they had intercourse, and how often they wanted to have intercourse. She compiled approximately 650 pages of spidery handwritten questionnaires but did not have the courage to publish, instead depositing the material in Stanford University Archives. Publication had to await the heroic efforts of James MaHood and his colleagues who collated and edited the questionnaires, leading in 1980 to their book, *The Mosher Survey* [9].

Clelia Mosher’s study, whilst not satisfactory from a sampling point-of-view because the results can in no way be generalized (the 45 women interviewed were, after all, mature, married, experienced, largely college-educated American women) remains a primary historical document of premodern sex and marriage in America. The reasons are clearly identified in [9];

... it contains statements of great rarity directly from Victorian women, whose lips previously had been sealed on the intimate questions of their private lives and cravings. Although one day it may come to light, we know of no other sex survey of Victorian women, in fact no earlier American sex survey of any kind, and certainly no earlier survey conducted by a woman sex researcher.

Two of the most dramatic findings of the Mosher survey are

- The Victorian women interviewed by Mosher appeared to relish sex, and claimed higher rates of orgasm than those reported in far more recent surveys.
- They practised effective birth-control techniques beyond merely abstinence or withdrawal.

For these experienced, college-educated women at least, the material collected by Mosher produced little evidence of Victorian prudery.

Nearly 40 years on from Mosher’s survey, Katharine Davis studied the sex lives of 2200 upper-middle class married and single women. The results of Davis’s survey are described in her book, *Factors in The Sex Life of Twenty Two Hundred Women*, published in 1929 [2]. Her stated aim was to gather data as to ‘normal experiences of sex on which to base educational programs’. Davis considered such normal sexual experiences to be, to a great extent, scientifically unexplored country. Unfortunately, the manner in which the eponymous women were selected for her study probably meant that these experiences were to remain so for some time to come.

Initially a letter asking for cooperation was sent to 10000 married women in all parts of the United States. Half of the addresses were furnished by a ‘large national organization’ (not identified by Davis). Recipients were asked to submit names of normal married women – ‘that is, women of good standing in the community, with no known physical, mental, or moral handicap, of sufficient

intelligence and education to understand and answer in writing a rather exhaustive set of questions as to sex experience'. (The questionnaire was eight pages long.)

Another 5000 names were selected from published membership lists of clubs belonging to the *General Federation of Women's Clubs*, or from the alumnae registers of women's colleges and coeducational universities.

In each letter was a return card and envelope. The women were asked to indicate on the card whether they would cooperate by filling out the questionnaire, which was sent only to women requesting it. This led to returned questionnaires from 1000 married women.

The unmarried women in the study were those five years out of a college education; again 10000 such women were sent a letter asking whether or not they would be willing to fill out, in their case, a 12-page questionnaire. This resulted in the remaining 1200 women in the study.

Every aspect of the selection of the 2200 women in Dr Davis's study is open to statistical criticism. The respondents were an unrepresentative sample, of volunteers who were educationally far above average and only about 10% of those contacted ever returned a questionnaire. The results are certainly not generalizable to any recognisable population of more universal interest. But despite its flaws a number of the charts and tables in the report retain a degree of fascination. Part of the questionnaire, for example, dealt with the use of methods of contraception. At the time, contraceptive information was categorized as obscene literature under federal law. Despite this, 730 of the 1000 married women who filled out questionnaires had used some form of contraceptive measure. Where did they receive their advice about these measures? Davis's report gives the sources shown in Table 1.

Davis along with most organizers of sex surveys also compiled figures on frequency of sex; these are shown in Table 2.

Davis's rationale for compiling the figures in Table 2 was to investigate the frequency of intercourse as a possible factor in sterility and for this purpose she breaks down the results in a number of ways. She found no evidence to suggest a relationship between marked frequency of intercourse and sterility – indeed she suggests that her results indicate the reverse.

Table 1 Sources of information about contraceptive measures (from [2])

Physicians	370
Married women friends	174
Husband	139
Mother	42
Friend of husband	39
Books	33
Birth-control circulars	31
'Common knowledge'	27
Nurse	15
Medical studies	9
'Various'	8
'Drug-store man'	6
The Bible	2
A servant	1
A psychoanalyst	1

Table 2 Frequency of intercourse of married women (from [2])

Answer	Number	Percent
More than once a day	19	2.0
Once a day	71	7.6
Over twice, less than seven times a week	305	31.3
Once or twice a week	391	40.0
One to three times a month	125	12.8
'Often' or 'frequently'	22	2.4
'Seldom' or 'infrequently'	38	3.9
Total answers to frequency questions	971	100
None in early years	8	
Unanswered (No answer)	21	
Total group	1000	

From a methodological point-of-view, one of the most interesting aspects of the Davis report is her attempt to compare the answers of women who responded by both interview and questionnaire. Only a relatively small number of women (50) participated in this comparison but in general there was a considerably higher incidence of 'sex practices' reported on the questionnaire. Davis makes the following argument as to why she considers the questionnaire results to be more likely to be closer to the truth;

In the evolutionary process civilization, for its own protection, has had to build up certain restraints on sexual instincts which, for the most part, have been in sense of shame, especially for sex outside of the legal sanction of marriage. Since sex practices prior to marriage have not the general approval of society, and since the desire for social approval is one of the

4 History of Surveys of Sexual Behavior

fundamental motives in human behavior, admitting such a practice constitutes a detrimental confession on the part of the individual and is more likely to be true than a denial of it. In other words, the group admitting the larger number of sex practices is assumed to contain the greater number of honest replies [2].

The argument is not wholly convincing, and would certainly not be one that could be made about the respondents in contemporary surveys of sexual behavior.

Perhaps the most famous sex survey ever conducted was the one by Kinsey and his colleagues in the 1940s. Alfred Charles Kinsey was undoubtedly the most famous American student of human sexual behavior in the first half of the twentieth century. He was born in 1894 and had a strict Methodist upbringing. Originally a biologist who studied *Cynipidae* (gall wasps), Kinsey was a professor of zoology, who never thought to study human sexuality until 1938, when he was asked to teach the sexuality section of a course on marriage. In preparing his lectures, he discovered that there was almost no information on the subject. Initially, and without assistance, he gathered sex histories on weekend field trips to nearby cities. Gradually this work involved a number of research assistants and was supported by grants from Indiana University and the Rockefeller Foundation.

Until Kinsey's work (and despite the earlier investigations of people like Mosher and Davis) most of what was known about human sexual behavior was based on what biologists knew about animal sex, what anthropologists knew about sex among natives in Non-Western, nonindustrialized societies, or what Freud and others learnt about sexuality from emotionally disturbed patients. Kinsey and his colleagues were the first psychological researchers to interview volunteers in depth about their sexual behavior. The research was often hampered by political investigations and threats of legal action. But in spite of such harassment, the first Kinsey report, *Sexual Behavior in the Human Male*, appeared in 1948 [7], and the second, *Sexual Behavior in the Human Female*, in 1953 [8]. It is no exaggeration to say that both caused a sensation and had massive impact. *Sexual Behavior in the Human Male*, quickly became a bestseller, despite its apparent drawbacks of stacks of tables, graphs, bibliography, and a 'scholarly' text that it is

kind to label as merely monotonous. The report certainly does not make for lively reading. Nevertheless, six months after its publication it still held second place on the list of nonfiction bestsellers in the USA. The first report proved of interest not only to the general public, but to psychiatrists, clergymen, lawyers, anthropologists, and even home economists. Reaction to it ranged all the way from extremely favourable to extremely unfavourable – here are some examples of both:

- The Kinsey Report has done for sex what Columbus did for geography,
- ... a revolutionary scientific classic, ranking with such pioneer books as Darwin's *Origin of the Species*, Freud's and Copernicus' original works,
- ... it is an assault on the family as the basic unit of society, a negation of moral law, and a celebration of licentiousness,
- there should be a law against doing research dealing exclusively with sex.

What made the first Kinsey report the talk of every town in the USA lies largely in the following summary of its main findings:

Of American males,

- 86% have premarital intercourse by the age of 30,
- 37%, at some time in their lives, engaged in homosexual activity climaxed by orgasm,
- 70% have, at some time, intercourse with prostitutes,
- 97% engage in forms of sexual activity, at some time in their lives, that are punishable as crimes under the law,
- of American married males, 40% have been involved in extramarital relations,
- of American farm boys, 16% have sexual contacts with animals.

These figures shocked because they suggested that there was much more sex, and much more variety of sexual behavior amongst American men than was suspected.

But we need to take only a brief look at some of the details of Kinsey's study to see that the figures above and the many others given in the report hardly stand up to statistical scrutiny.

Although well aware of the scientific principles of sampling, Kinsey based all his tables, charts, and so on, on a total of 5300 interviews with volunteers. He knew that the ideal situation would have

been to select people at random, but he did not think it possible to coax a randomly selected group of American males to answer truthfully when asked deeply personal questions about their sex lives. Kinsey sought volunteers from a diversity of sources so that all types would be sampled. The work was, for example, carried on in every state of the Union, and individuals from various educational groups were interviewed. But the 'diversification' was rather haphazard and the proportion of respondents in each cell did not reflect the United States population data. So the study begins with the disadvantage of volunteers and without a representative sample in any sense. The potential for introducing bias seems to loom large since, for example, those who volunteer to take part in a sex survey might very well have different behavior, different experiences, and different attitudes towards sex than the general population. In fact, recent studies show that people who volunteer to take part in surveys about sexual behavior are likely to be more sexually experienced and also more interested in sexual variety than those who do not volunteer.

A number of procedures were used by Kinsey to obtain interviews and to reduce refusals. Contacts were made through organizations and institutions that in turn persuaded their members to volunteer. In addition, public appeals were made and often one respondent would recommend another. Occasionally, payments were given as incentives. The investigators attempted to get an unbiased selection by seeking all kinds of histories and by long attempts to persuade those who were initially hostile to come into the sample. In a two-hour interview, Kinsey's investigators covered from 300 to 500 items about the respondent's sexual history, but no sample questionnaire is provided in the published report. The definition of each item in the survey was standard, but the wording of the questions and the order in which they were given were varied for each respondent. In many instances leading questions were asked such as, '*When did you last . . .*' or '*When was the first time you . . .*', thereby placing the onus of denial on the respondent. The use of leading questions is generally thought to lead to the overreporting of an activity. Kinsey's aim was to provide the ideal setting for each individual interview whilst retaining an equivalence in the interviews administered to all respondents. So the objective conditions of the interview were not uniform and variation in sexual behavior between

individuals might be confounded with differences in question wording and order.

The interview data in the Kinsey survey were recorded in the respondent's presence by a system of coding that was consigned to memory by all six interviewers during the year-long training that preceded data collection. Coding in the field has several advantages such as speed and the possibility of clarifying ambiguous answers; memory was used in preference to a written version of the code to preserve the confidence of the interviewee. But the usual code ranged from six to twenty categories for each of the maximum of 521 items that could be covered in the interview, so prodigious feats of memory were called for. One can only marvel at the feat. Unfortunately, although field coding was continually checked, no specific data on the reliability of coding are presented and there has to be some suspicion that occasionally, at least, the interviewer made coding mistakes.

Memory certainly also played a role in the accuracy of respondent's answers to questions about events which might have happened long ago. It's difficult to believe, for example, that many people can remember details of frequency of orgasm per week, per five-year period, but this is how these frequencies are presented. Many of the interviews in the first Kinsey report were obtained through the cooperation of key individuals in a community who recommended friends and acquaintances, and through the process of developing a real friendship with the prospective respondent before starting the interview as the following quotation from the report indicates:

We go with them to dinner, to concerts, to nightclubs, to the theatre, we become acquainted with them at community dances and in poolrooms and taverns, and in other places which they frequent. They in turn invite us to meet friends in their homes, at teas, at dinners, at other social events [7, p. 40].

This all sounds very pleasant both for the respondents and the interviewers but is it good survey research practice? Probably not, since experience suggests that the 'sociological stranger' gets the more accurate information in a sensitive survey, because the respondent is wary about revealing his most private behavior to a friend or acquaintance. And assuming that all the interviewers were white males the question arises as to how this affected interviews with say, African-American respondents (and in the second report, with women)?

Finally there are some more direct statistical criticisms that can be levelled at the first Kinsey report. There is, for example, often a peculiar variation in the number of cases in a given cell, from table to table. A particular group will be reported on one type of sexual behavior, and this same group may be of slightly different size in another table. The most likely explanation is that the differences are due to loss of information through 'Don't know' responses or omissions of various items, but the discrepancies are left unexplained in the report. And Kinsey seems shaky on the definition of terms such as median although this statistic is often used to summarize findings. Likewise he uses the sample **range** as a measure of how much particular measurements varied amongst his respondents rather than the preferable standard deviation statistic.

Kinsey addressed the possibility of bias in his study of male sexual behavior and somewhat surprisingly suggested that any lack of validity in the reports he obtained would be in the direction of concealment or understatement. Kinsey gives little credence to the possibility of overstatement:

Cover-up is more easily accomplished than exaggeration in giving a history [7, p. 54].

Kinsey thought that the interview approach provided considerable protection against exaggeration but not so much against understatement. But given all the points made earlier this claim is not convincing, and it is not borne out by later, better-designed studies, which generally report lower levels of sexual activity than Kinsey. For example, the 'Sex in America' survey [10] was based on a representative sample of Americans and it showed that individuals were more monogamous and more sexually conservative than had been reported previously.

Kinsey concludes his first report with the following.

We have performed our function when we have published the record of what we have found the human male doing sexually, as far as we have been able to ascertain the facts.

Unfortunately, the 'facts' arrived at by Kinsey and his colleagues may have been distorted in a variety of ways because of the many flaws in the study. But despite the many methodological errors, Kinsey's studies remain gallant attempts to survey the approximate range and norms of sexual behavior.

The Kinsey report did have the very positive effect of encouraging others to take up the challenge of investigating human sexual behavior in a scientific and objective manner. In the United Kingdom, for example, an organization known as *Mass-Observation* carried out a sex survey in 1949 that was directly inspired by Kinsey's first study. In fact it became generally known as 'Little Kinsey' [3]. Composed of three related surveys, 'Little Kinsey' was actually very different methodologically from its American predecessor. The three components of the study were as follows:

1. A 'street sample' survey of over 2000 people selected by random sampling methods carried out in a wide cross section of cities, towns and villages in Britain.
2. A postal survey of about 1000 each of three groups of 'opinion leaders': clergymen, teachers, and doctors.
3. A set of interrelated questions sent to members of Mass-Observation's National Panel, which produced responses from around 450 members.

The report's author, Tom Harrison, was eager to get to the human content lying behind the line-up of percentages and numbers central to the Kinsey report proper, and he suggested that the Mass-Observation study was both 'something less and something more than Kinsey'. It tapped into 'more of the actuality, the real life, the personal stuff of the problem'. He tried to achieve these aims by including in each chapter some very basic tables of responses, along with large numbers of comments from respondents to particular questions. Unfortunately this idiosyncratic approach meant that the study largely failed to have any lasting impact, although later authors, for example, Liz Stanley in *Sex Surveyed 1949-1994* [11], claim it was of pioneering importance and was remarkable for pinpointing areas of behavioral and attitudinal change. It does appear to be one of the earliest surveys of sex that used random sampling. Here are some of the figures and comments from Chapter 7 of the report, *Sex Outside Marriage*.

The percentages who disapproved of extramarital relations were

- 24% on the National Panel,
- 63% of the street sample,
- 65% of doctors,
- 75% of teachers,

- 90% of clergymen.

Amongst the street sample the following percentages were given for those opposed to extramarital relations:

- 73% of all weekly churchgoers,
- 54% of all non-churchgoers,
- 64% of people leaving school up to and including 15 years,
- 50% of all leaving school after 16,
- 68% of all living in rural areas,
- 50% of all Londoners,
- 67% of all women,
- 57% of all men,
- 64% of all married people over 30,
- 48% of all single people over 30.

The Kinsey report, 'Little Kinsey2', and the surveys of Clelia Mosher and Katherine Davis, represent, despite their flaws, genuine attempts at taking an objective, scientific approach to information about sexual behavior. But sex, being such a fascinating topic also attracts the more sensational commercial 'pseudosurveys' like those regularly conducted amongst the readership of magazines such as *Playboy* and *Cosmopolitan*. Here the questions asked are generally a distinctly 'racier' variety than in more serious surveys. Here is just one example:

- When making love, which of the following do you like? (check all that apply)
 1. Have your man undress you
 2. Pinch, bite, slap him
 3. Be pinched, bitten, slapped
 4. Have someone beat you
 5. Pretend to fight physically with the man or try to get away.

The aims of these surveys are to show that the readership of the magazine enjoys sexually exciting lives, to celebrate their reader's 'sexual liberation' and to make the rest of us green eyed with envy (or red faced with shame). The results are generally presented in the form of tabloid type headlines, for example;

French have more sex than Englishmen.

Such surveys are, essentially, simply sources of fun, fantasy, and profit and can, of course, be easily dismissed from serious consideration because of their obvious biases, clear lack of objectivity, poor sampling methods and shoddy questionnaire design.

Unfortunately, there have been several surveys of sexual behavior that demand to be taken seriously, but to which the same criticisms can be applied, and where, in addition, attempts to interpret the findings of the survey may have been colored by the likely *a priori* prejudices of the survey's instigator. One such example is the basis of that 1976 bestseller, *The Hite Report on Female Sexuality* [6].

Shire Hite is a member of the *National Organization of Women* and an active feminist. When she undertook her study in the 1970s, the aim of which she stated as 'to define or discover the physical nature of [women's] sexuality', she clearly had a feminist political axe to grind. — 'Most sex surveys have been done by men' she said and nobody had asked women the right questions. She wanted 'women to be experts and to say what female sexuality was about'. However, Dr Hite often appeared to have a strong prior inkling of what her respondents would tell her and such clear expectations of results are a matter of concern. First, we consider the methodology underlying the Hite report.

Hite sent questionnaires to 'consciousness-raising', abortion rights, and other women's groups and also advertised for respondents in newspapers and magazines, including *Ms.*, *Mademoiselle* and *Brides*. Of the 100 000 questionnaires distributed, Hite received somewhat more than 3000 responses, a response rate, she claimed, that was standard for surveys of this type. However, most serious survey researchers would regard 3% as very low. So the survey begins with an extremely biased sample and a very low response rate.

A further problem was that the questionnaire used in the study was hard to complete. Each question contained multiple subquestions, never a good idea in any survey. In addition, the survey began with numerous questions about orgasm rather than with more innocuous questions. Many questions called for 'essay-like' responses and others asked for seemingly impossible details from past events. Here are some examples:

- Do you have orgasms? If not, what do you think would contribute to your having them?
- Do you always have orgasms during the following (please indicate whether always, usually, sometimes, rarely, or never):
 1. Masturbation,
 2. Intercourse (vaginal penetration),

3. Manual clitoral stimulation by partner,
 4. Oral stimulation by a partner,
 5. Intercourse plus manual clitoral stimulation,
 6. Never have orgasms.
- Also indicate above how many orgasms you usually have during each activity, and how long you usually take.
 - Please give a graphic description of how your body could best be stimulated to orgasm.

Hite's questionnaire began with items about orgasm and much of her book dwells on her interpretation of the results from these items. She concludes that women can reach orgasm easily through masturbation but far less easily, if at all, through intercourse with their male partners. Indeed, one of her main messages is that intercourse is less satisfying to women than masturbation. She goes on to blame what she sees as the sorry state of female sexual pleasure in patriarchal societies, such as the United States, that glorify intercourse. Critics pointed out that there may be something in all of this, but that Hite was being less than honest to suppose that her views were an inescapable conclusion from the results of her survey. As the historian Linda Gordon pointed out [5], the Hite report was orientated towards young, attractive, autonomous career women, who were focused on pleasure and unencumbered by children. These women could purchase vibrators, read the text, and undergo the self-improvement necessary for one-person sexual bliss.

The Hite report has severe methodological flaws and these are compounded by the suspicion that its writer is hardly objective about the issues under investigation. The numbers are neither likely to have accurately reflected the facts, nor to have been value-free.

(It is not, of course, feminist theory that is at fault in the Hite report, as the comprehensive study of sex survey research given in [4], demonstrates; these two authors combine feminist theory with a critical analysis of survey research to produce a well-balanced and informative account.)

If the Hite Report was largely a flash in the media pan (Sheer Hype perhaps?), the survey on sexual attitudes and lifestyles undertaken in the UK in the late 1980s and early 1990s by Kaye Wellings and her coworkers [12] acts as a model of excellence for survey research in such a sensitive area. The impetus for the survey was the emergence of the HIV pandemic, and the attendant effort to assess and

control its spread. The emergence in the 1980s of a lethal epidemic of sexually transmitted infection focused attention on the profound ignorance that still remained about many aspects of sexual behavior, despite Kinsey and others. The collaboration of epidemiologists, statisticians, and survey researchers produced a plan and a survey about sex in which all the many problems with such surveys identified earlier were largely overcome.

A feasibility study assessed the acceptability of the survey, the extent to which it would produce valid and reliable results, and the sample size needed to produce statistically acceptable accuracy in estimates of minority behavior. The results of the feasibility study guided the design of the final questionnaire that was used in obtaining results from a carefully selected random sample of individuals representative of the general population. Of the 20 000 planned interviews 18876 were completed. Nonresponse rates were generally low. The results provided by the survey give a convincing account of sexual lifestyle in Britain at the end of the twentieth century. For interest one of the tables from the survey is reproduced in Table 3. The impact of AIDS has also been responsible for an increasing number of surveys about sexual behavior in the developing world, particularly in parts of Africa. A comprehensive account of such surveys is given in [1].

Summary

Since 1892 when a biology student, Clelia Mosher, questioned 45 upper middle-class married Victorian women about their sex lives, survey researchers have asked thousands of people about their sexual behavior. According to Julia Ericksen [4] in *Kiss and Tell*, 'Sexual behavior is a volatile and sensitive topic, and surveys designed to reveal it have great power and great limits'. Their power has been to help change, radically change in particular aspects, attitudes about sex compared to 50 years ago. Their limits have often been their methodological flaws. And, of course, when it comes to finding out about their sexual behavior, people may not want to tell, and even if they agree to be interviewed they may not be entirely truthful. But despite these caveats the information from many of the surveys of human sexual behavior has probably helped remove the conspiracy of silence about sex that existed in society, which condemned

Table 3 Number and percent of respondents taking part in different sexual practices in the last year and ever, by social class (from [12])

	Vaginal intercourse			Oral sex		
	Last year (%)	Ever (%)	Number of respondents	Last year (%)	Ever (%)	Number of respondents
Men						
Social Class						
I, II	91.5	97.7	2757	67.9	84.3	2748
III NM	90.3	95.5	1486	67.9	78.2	1475
III M	86.1	95.2	2077	60.4	72.8	2058
IV, V	83.3	91.0	849	57.6	67.3	840
Other	52.9	61.6	693	40.8	50.0	686
Women						
Social Class						
I, II	91.8	98.2	3460	61.0	76.2	3413
III NM	85.9	94.3	2248	60.3	71.5	2216
III M	90.1	97.2	1857	54.5	65.5	1826
IV, V	81.9	93.6	1007	52.4	64.3	992
Other	56.7	74.5	1212	41.9	54.7	1200

NM = nonmanual workers; M = Manual workers.

many men and women to a miserable and unfulfilling sex life. The results have challenged views of the past 100 years that sex was not central to a happy marriage and that sex, as a pleasure for its own sake, debased the marital relationship. Sex as pleasure is no longer regarded by most people as a danger likely to overwhelm the supposedly more spiritual bond between a man and a woman thought by some to be achieved when sex occurs solely for the purposes of reproduction. Overall the information about human sexual behavior gathered from sex surveys has helped to promote, all be it in a modest way, a healthier attitude toward sexual matters and perhaps a more enjoyable sex life for many people.

References

- [1] Cleland, J. & Ferry, B. (1995). *Sexual Behavior in the Developing World*, Taylor & Francis, Bristol.
- [2] Davis, K. (1929). *Factors in the Sex Life of Twenty-Two Hundred Women*, Harper and Brothers, New York.
- [3] England, L.R. (1949). "Little Kinsey": an outline of sex attitudes in Britain, *Public Opinion* **13**, 587–600.
- [4] Ericksen, J.A. & Steffen, S.A. (1999). *Kiss and Tell: Surveying Sex in the Twentieth Century*, Harvard University Press, Cambridge.
- [5] Gordon, L. (2002). *The Moral Property of Women: A History of Birth Control Politics in America*, 3rd Edition, University of Illinois Press, Champaign.
- [6] Hite, S. (1976). *The Hite Report: A Nationwide Study on Female Sexuality*, Macmillan, New York.
- [7] Kinsey, A.C., Wardell, B.P. & Martin, C.E. (1948). *Sexual Behavior in the Human Male*, Saunders, Philadelphia.
- [8] Kinsey, A.C., Wardell, B.P., Martin, C.E. & Gebhard, P.H. (1953). *Sexual Behavior in the Human Female*, Saunders, Philadelphia.
- [9] MaHood, J. & Wenburg, K. (1980). in *The Mosher Survey*, C.D. Mosher, ed., Arno Press, New York.
- [10] Michael, R.T., Gagnon, J., Lauman, E.O. & Kolata, G. (1994). *Sex in America: A definitive Survey*, Little, Brown, Boston.
- [11] Stanley, L. (1995). *Sex Surveyed 1949-1994*, Taylor & Francis, London.
- [12] Wellings, K., Field, J., Johnson, A. & Wadsworth, J. (1994). *Sexual Behavior in Britain*, Penguin Books, London.

BRIAN S. EVERITT

Hodges–Lehman Estimator

CLIFFORD E. LUNNEBORG

Volume 2, pp. 887–888

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hodges–Lehman Estimator

The Hodges–Lehmann one-sample estimator provides a valuable alternative to the sample mean or median as an estimator of the location of the center of a distribution. While the median is to be preferred to the mean with nonsymmetric populations, it requires far more observations than the mean to obtain the same level of precision. The median’s **asymptotic relative efficiency** with respect to the mean for data from a normal population is only 0.64. By contrast, while the Hodges–Lehmann estimator offers the same advantages as the median, its asymptotic relative efficiency with respect to the mean is 0.96 for similar data [2].

The one-sample estimator, based on a random sample of n observations, is linked with the **Wilcoxon signed-rank test** [1] and is defined as the median of the set of $[n(n+1)/2]$ **Walsh averages**. Each Walsh average is the arithmetic average of a pair of observations, including observations paired with themselves.

The sample (2, 5, 7, 11), for example, gives rise to the 10 Walsh averages tabled below (Table 1).

The median of the set of Walsh averages is the one-sample Hodges–Lehmann estimator. For our example, with ten Walsh averages, the median estimate is the average of the fifth and sixth smallest Walsh averages, $(6 + 6.5)/2 = 6.25$.

The Hodges–Lehmann two-sample estimator provides an alternative to the difference in sample means when estimating a shift in location between two population distributions. The location shift model associated with the Wilcoxon Mann-Whitney test postulates that, while the two distributions are alike in shape and variability, the Y distribution is shifted upward or downward by an amount Δ relative to the X distribution [1]. The model is usually stated in terms of the relation between the cumulative probabilities for the two distributions: $Prob[Y \leq z] = Prob[X \leq (z - \Delta)]$, for all values of z . Positive values of Δ are associated with larger values in the Y than in the X distribution.

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_m)$ be independent random samples from the X and Y distributions, respectively. The Hodges–Lehmann estimate of Δ is the median of the $(n \times m)$ pairwise differences, $(y_j - x_k)$, $j = 1, \dots, m$, $k = 1, \dots, n$.

For the two samples $x = (2, 5, 7, 11)$ and $y = (3, 4, 8, 20)$, the relevant pairwise differences are displayed in the table below (Table 2).

The median of the 16 pairwise differences is the average of the eighth and ninth smallest differences. As both of these differences are 1, the Hodges–Lehmann estimate is 1.0. The difference-in-means estimate of Δ , by comparison, is $(35/4) - (25/4) = 2.5$. The latter estimator is much more sensitive to outliers.

The author gratefully acknowledges the assistance of Phillip Good in the preparation of this article.

Table 1 Computation of one-sample Walsh averages

	2	5	7	11
2	$(2 + 2)/2 = 2$	$(2 + 5)/2 = 3.5$	$(2 + 7)/2 = 4.5$	$(2 + 11)/2 = 6.5$
5		$(5 + 5)/2 = 5$	$(5 + 7)/2 = 6$	$(5 + 11)/2 = 8$
7			$(7 + 7)/2 = 7$	$(7 + 11)/2 = 9$
11				$(11 + 11)/2 = 11$

Table 2 Computations of two-sample Walsh averages

$x \setminus y$	3	4	8	20
2	$(3 - 2) = 1$	$(4 - 2) = 2$	$(8 - 2) = 6$	$(20 - 2) = 18$
5	$(3 - 5) = -2$	$(4 - 5) = -1$	$(8 - 5) = 3$	$(20 - 5) = 15$
7	$(3 - 7) = -4$	$(4 - 7) = -3$	$(8 - 7) = 1$	$(20 - 7) = 13$
11	$(3 - 11) = -8$	$(4 - 11) = -7$	$(8 - 11) = -3$	$(20 - 11) = 9$

2 Hodges–Lehman Estimator

References

- [1] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.
- [2] Lehmann, E.L. (1999). *Elements of Large Sample Theory*, Springer, New York.

CLIFFORD E. LUNNEBORG

Horseshoe Pattern

BRIAN S. EVERITT

Volume 2, pp. 889–889

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Horseshoe Pattern

Horseshoe Effect

All forms of **multidimensional scaling** have as their aim the low-dimensional representation of a set of proximity data (*see Proximity Measures*). A classic example is the recreation of a map from a matrix of say intercity road distances in a country. Often, such a map can be successfully recreated if only the ranking of the distances is given (*see [2]*). With such data, the underlying structure is essentially two-dimensional, and so can be represented with little distortion in a two-dimensional scaling solution. But when the observed data have a one-dimensional structure, for example, in a chronological study, representing the observed proximities in a two-dimensional scaling solution often gives rise to what is commonly referred to as the *horseshoe effect*. This effect appears to have been first identified in [2] and can be illustrated by the following example:

Consider 51 objects, $O_1, O_2, \dots, O_{51}$ assumed to be arranged along a straight line with the j th object being located at the point with coordinate j . Define the similarity, s_{ij} , between object i and object j , as, follows:

$$s_{ij} = \begin{cases} 9 & \text{if } i = j, \\ 8 & \text{if } 1 \leq |i - j| \leq 3, \\ 7 & \text{if } 4 \leq |i - j| \leq 6, \\ \vdots & \\ 1 & \text{if } 22 \leq |i - j| \leq 24 \\ 0 & \text{if } |i - j| \geq 25 \end{cases} \quad (1)$$

Next, convert these similarities into dissimilarities, δ_{ij} , using $\delta_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{1/2}$ and then apply *classical multidimensional scaling* (*see Multidimensional Scaling*) to the resulting dissimilarity matrix. The two-dimensional solution is shown in Figure 1. The original order has been reconstructed very well, but the plot shows the characteristic horseshoe shape,

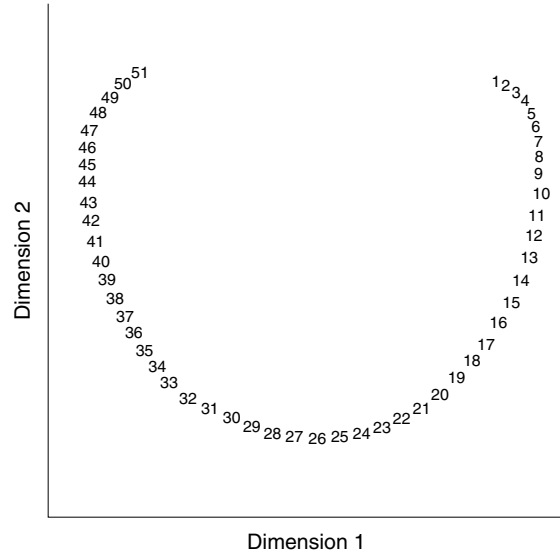


Figure 1 An example of the horseshoe effect

which is a consequence of the ‘blurring’ of the large distances and is characteristic of such situations.

Further discussion of the horseshoe effect is given in [3] and some examples of its appearance in practice are described in [1].

References

- [1] Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*, Arnold, London.
- [2] Kendall, D.G. (1971). A mathematical approach to seriation, *Philosophical Transactions of the Royal Society of London*, **A269**, 125–135.
- [3] Podani, J. & Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis, *Ecology* **83**, 3331–3343.

BRIAN S. EVERITT

Hotelling, Howard

SCOTT L. HERSHBERGER

Volume 2, pp. 889–891

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hotelling, Howard

Born: September 29, 1895, Minnesota, USA.

Died: December 26, 1973, North Carolina, USA.

Although Howard Hotelling was born in Fulda, Minnesota in 1895, he spent most of his childhood and young adulthood in Seattle, Washington. He attended the University of Washington, graduating in 1919 with a bachelor's degree in journalism. As an undergraduate student, Hotelling worked for several local newspapers. While still an undergraduate, he took a mathematics class taught by Eric Temple Bell. Bell recognized Hotelling's unusual analytic capabilities and encouraged him to do graduate work in mathematics. Hotelling earned a master's degree in mathematics from the University of Washington in 1921 and a Ph.D. in mathematics from Princeton University in 1924. In 1925, he published an article, 'Three-dimensional Manifolds of States of Motion', based on his dissertation in the *Transactions of the American Mathematical Society* [1].

From 1925 to 1931, Hotelling taught probability and statistics at Stanford University. During these years, Hotelling applied newly developed statistical techniques to areas as diverse as journalism, political science, and economics. His interests in statistics led him in 1929 to study with **R. A. Fisher** at the Rothamsted Experimental Station, an agricultural research institute in Hertfordshire, UK.

In 1931, Hotelling went to Columbia University as a professor of economics, where he stayed until 1946. During the 15 years Hotelling was at Columbia, the statistical research group he founded lent statistical assistance to the United States' military efforts in WWII. This statistical research group, which counted Abraham Wald among its many prominent members, introduced and developed sequential analysis. Sequential analysis proved to be so useful to the US military that the technique itself was considered to be classified information until the end of WWII.

In 1946, Hotelling moved from Columbia University to the University of North Carolina at Chapel Hill, where he founded a department of statistics. Hotelling remained at the University of North Carolina for the remainder of his professional life. He received a number of honors, including an honorary LL.D. from the University of Chicago, and

an honorary D.Sc. from the University of Rochester. He was an honorary fellow of the Royal Statistical Society and a distinguished fellow of the American Economic Association. He was president of the Econometric Society from 1936 to 1937 and of the Institute of Mathematical Statistics in 1941. The National Academy of Sciences elected him as a member in 1972. In May of 1972, he experienced a severe stroke; his death following in December of 1973.

Hotelling's greatest contributions to statistics were in the general areas of econometrics and multivariate analysis. In econometrics, Hotelling's 1929 paper on the stability of competition introduced the idea of spatial competition, known as 'Hotelling's model', when there are only two sellers competing in a market [2]. The solution to this problem was an early statement of the **game theory** concept of the 'Subgame-perfect equilibrium', although Hotelling did not refer to it as such. Hotelling introduced the calculus of *variations* to economics in 1931 as a method of analyzing resource exhaustion [4]. Although Hicks and Allen are generally credited in 1934 with explaining the downward slope of demand curves, Hotelling, in fact, also in that year, had independently derived an identical solution. (The paper describing his solution did not, however, appear until 1935, [7]). In 1938, Hotelling introduced the concept of 'marginal cost pricing' equilibrium: Economic efficiency is achieved if every good is produced and priced at marginal cost. At the same time, he introduced the 'two-part' tariff as a solution to situations of natural monopoly [9].

Hotelling's contributions to multivariate analysis were many, among the more important of which was his multivariate generalization of Student's t Test in 1931 [5] – now known as Hotelling's T^2 . Hotelling also proposed the methods of **principal component analysis** [6] and **canonical correlations** [8]. His paper (cowritten by Working) on the interpretation of trends contains the first example of a confidence region, as well as multiple comparisons [11]. His contributions (with Richards) to the theory of rank statistics (*see Rank Based Inference*) were also highly influential [10]. This work grew out of his interest in specifying the precise conditions for the consistency and asymptotic normality of **maximum likelihood estimates** [3]. Hotelling was always clear that he was a statistician first, econometrician second.

References

- [1] Hotelling, H. (1925). Three-dimensional manifolds of states of motion, *Transactions of the American Mathematical Society* **27**, 329–344.
- [2] Hotelling, H. (1929). Stability in competition, *Economic Journal* **39**, 41–57.
- [3] Hotelling, H. (1930). The consistency and ultimate distribution of optimum statistics, *Transactions of the American Mathematical Society* **32**, 847–859.
- [4] Hotelling, H. (1931). The economics of exhaustible resources, *Journal of Political Economy* **39**, 137–175.
- [5] Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.
- [6] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.
- [7] Hotelling, H. (1935). Demand functions with limited budgets, *Econometrica* **3**, 66–78.
- [8] Hotelling, H. (1936). Relations between two sets of variates, *Biometrika* **28**, 321–377.
- [9] Hotelling, H. (1938). The general welfare in relation in problems of taxation and of railway and utility rates, *Econometrica* **6**, 242–269.
- [10] Hotelling, H. & Pabst, M.R. (1936). Rank correlation and test of significance involving no assumption of normality, *Annals of Mathematical Statistics* **7**, 29–43.
- [11] Working, H. & Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends, *Journal of the American Statistical Association* **24**, 73–85.

SCOTT L. HERSHBERGER

Hull, Clark L

SANDY LOVIE

Volume 2, pp. 891–892

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Hull, Clark L

Born: May 24, 1884, in New York State.

Died: May 10, 1952, in Connecticut.

Clark L. Hull embraced behaviorism from the start of his academic career in psychology: his Ph.D. in 1918, for example, offered a behavioral recasting of concept acquisition as a *quantitative* form of learning, in contrast to the heavy emphasis on introspection and philosophy, which characterized contemporary work on thinking and reasoning. He also took up classical conditioning after the American publication of Pavlov's *Conditioned Reflexes* in 1927, followed by Thorndike's law of effect, thus melding reinforcement to associationism as the basis of all learning. But his most important contribution to psychology was to pioneer mathematical models of learning. It is, of course, the case that there had been some attempts to formally model behavior before Hull (see Miller [3], for an elementary and now classic introduction to the field), but his was the most ambitious attempt prior to the advent of today's more sophisticated forms of mathematical psychology.

His early days had been very difficult, not only because of the poverty of his parents and the hard work involved in helping them run their farm but also because he had suffered a debilitating bout of typhoid fever in 1905, having just graduated from Alma College with the aim of eventually becoming a mining engineer. Even worse was the attack of poliomyelitis at the age of 24, which left him with a paralyzed leg and a lifelong fear both of illness and that he might not live long enough to accomplish his goals. Clearly, a career in mining was now out of question, so Hull opted for a degree in psychology at the University of Michigan, graduating after two years study with an A.B. and a Teachers diploma. After a year teaching at a Normal School in Kentucky, which had been set up to educate teachers, Hull was finally accepted for graduate work at the University of Wisconsin under Joseph Jastrow. An M.A. was quickly obtained in 1915, followed by a Ph.D. in 1918, the latter appearing as a monograph in 1920. Hull's career was now, after many setbacks, on an upward trajectory: he became a full professor at Wisconsin in 1925, for instance, and had published his first book (on aptitude measurement) in 1928.

More important, however, was his move to Yale's Institute of Psychology in 1929 as an expert on test theory, although he had quickly abandoned this research area in favor of learning and conditioning, although not without some internal bickering. The effective incorporation of the Institute of Psychology into the much larger and more generously funded Institute of Human Relations at Yale, together with a more sympathetic head of the Institute, meant that Hull could now indulge in his lifelong passion. This resulted in a flow of experimental and theoretical papers in the 1930s, which pushed the study of learning to new heights of sophistication, with Hull and his occasional rival Edward C. Tolman being increasingly cited as the only game in town.

What Hull did was essentially to model the so-called learning curve, that is, the growth in learning or extinction over trials. The models used to capture such curves were almost exclusively exponential in form, although Hull was not averse to employing more than one exponential function to model a particular data set. Underlying the models was an elaborate internal mechanism designed to reflect the effects of such factors as reinforcement, stimulus and response generalization, excitation and inhibition (of the Pavlovian variety), and behavioral fluctuation or oscillation (to use his term) on learning. This turned into a considerable program of research that was supported by a large number of graduate students and Institute staff, and was organized on an almost industrial or military scale. Hull was also happy to use data produced by other laboratories provided that the quality of their work satisfied Hull's own exacting standards. He also ransacked the journals for good quality data. Consequently, his *magnum opus* of 1943, *The Principles of Behavior* [1], mixed mathematics and data not just from the Yale laboratory but from an international collection of research sources. This book had followed a much more indigestible monograph on modeling rote learning that had appeared in mimeograph form in 1940 and in which the full rigor of fitting exponential functions to learning was laid out for the first time.

The two books also revealed the weaknesses in Hull's approach in that the fitting process did not use contemporary statistical methods like **least squares**, and the models themselves were historically somewhat antiquated. Thus, Hull fitted the data by eye and by reiterated fitting alone, after having selected the exponential function on highly informal grounds.

The structure of the models was also essentially deterministic in that the only random variable was error, that is, behavioral oscillation, which was added to the fixed effects of reinforcement, stimulus generalization, and so on, in order to account for the results.

The last years of Hull's life were dogged by increasing ill health, and he was only able to publish the updated and expanded version of the *Principles, A Behavior System* [2], in 1962, the year of his death. Hull fought against considerable personal and bureaucratic odds to demonstrate that learning, the most self-consciously *scientific* area of all experimental psychology, could be moved that much closer to

respectability through the application of mathematics to data.

References

- [1] Hull, C.L. (1943). *Principles of Behavior*, Appleton-Century, New York.
- [2] Hull, C.L. (1962). *A Behavior System*, Yale University Press, New Haven.
- [3] Miller, G.A. (1964). *Mathematics and Psychology*, Wiley, New York.

SANDY LOVIE

Identification

DAVID KAPLAN

Volume 2, pp. 893–896

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Identification

A prerequisite for the estimation of the parameters of any model is to establish whether the parameters of the model are *identified*. Identification refers to whether the parameters of the model can be uniquely determined from the data. If the parameters of the model are not identified, estimation of the parameters is not possible. Although the problem of identification is present in almost all parametric statistical models, the role of identification is perhaps clearest in the context of covariance structure models. Here, we will define the problem of identification from the **covariance structure modeling** perspective. Later, we will introduce the problem of identification from the simultaneous equation modeling perspective when considering some simple rules for establishing identification. As simultaneous equation modeling can be seen as a special case of covariance structure modeling, this discussion is quite general.

Definition of Identification

We begin with a definition of *identification* from the perspective of covariance structure modeling. The advantage of this perspective is that covariance structure modeling includes, as a special case, the simple linear regression model, and, therefore, we can understand the role of identification even in this simple case. First, arrange the unknown parameters of the model in the vector Ω . Consider next a population covariance matrix Σ whose elements are population variances and covariances (*see Correlation and Covariance Matrices*). It is assumed that a substantive model can be specified to explain the variances and covariances contained in Σ . Such a substantive model can be as simple as a two-variable linear regression model or as complicated as a simultaneous equation model. We know that the variances and covariances contained in Σ can be estimated by their sample counterparts in the sample covariance matrix S using straightforward formulae for the calculation of sample variances and covariances. Thus, the parameters in Σ are identified.

Having established that the elements in Σ are identified from their sample counterparts, what we need to establish in order to permit estimation of the model parameters is whether the model parameters

are identified. By definition, we say that the elements in Ω are *identified* if they can be expressed uniquely in terms of the elements of the covariance matrix Σ . If all elements in Ω are identified, we say that the model is identified.

Some Common Identification Rules

Let us now consider the identification of the parameters of a general simultaneous equation model that can be written as

$$y = By + \Gamma x + \zeta, \quad (1)$$

where y is a vector endogenous variable that the model is specified to explain, x is a vector exogenous variable that is purported to explain y but whose behavior is not explained, ζ is a vector of disturbance terms, and B and Γ are coefficient matrices. Note that when $B = 0$, we have the **multivariate multiple regression** model

$$y = \Gamma x + \zeta. \quad (2)$$

When the vector of endogenous variables contains only one column (i.e., only one explanatory variable is considered), then we have the case of simple linear regression.

To begin, we note that there exists an initial set of restrictions that must be imposed even for simple regression models. The first restriction, referred to as *normalization*, requires that we set the diagonal elements of B to zero, such that an endogenous variable cannot have a direct effect on itself.

The second requirement concerns the vector of disturbance terms ζ . Note that the disturbances for each equation are unobserved, and, hence, have no inherent metric. The most common way to set the metric of ζ , and the one used in simple regression modeling, is to fix the coefficient relating the endogenous variables to the disturbance terms to 1.0. An inspection of (2) reveals that ζ is actually being multiplied by the scaling factor 1.0. Thus, the disturbance terms are in the same scale as their relevant endogenous variables.

With the normalization rule in place and the metric of ζ fixed, we can now discuss some common rules for the identification of simultaneous equation model parameters. Recall again that we wish to know whether the variances and covariances of the

2 Identification

exogenous variables (contained in Φ), the variances and covariances of the disturbance terms (contained in Ψ), and the regression coefficients (contained in B and Γ) can be solved in terms of the variances and covariances contained in Σ .

Two classical approaches to identification can be distinguished in terms of whether identification is evaluated on the model as a whole, or whether identification is evaluated on each equation comprising the system of simultaneous equations. The former approach is generally associated with social science applications of simultaneous equation modeling, while the latter approach appears to be favored in the econometrics field applied mainly to simultaneous (i.e., nonrecursive) models. Nevertheless, they both provide a consistent picture of identification in that if any equation is not identified, the model as a whole is not identified.

The first, and perhaps simplest, method for ascertaining the identification of the model parameters is referred to as the *counting rule*. Let $s = p + q$ be the total number of endogenous and exogenous variables, respectively. Then the number of nonredundant elements in Σ is equal to $1/2 s(s + 1)$. Let t be the total number of parameters in the model that are to be estimated (i.e., the free parameters). Then, the counting rule states that a necessary condition for identification is that $t \leq 1/2 s(s + 1)$. If the equality holds, then we say that the model may be *just identified*. If t is strictly less than $1/2 s(s + 1)$, then we say that the model may be *overidentified*. If t is greater than $1/2 s(s + 1)$, then the model may be *not identified*.

Clearly, the advantage to the counting rule is its simplicity. However, the counting rule is a necessary but not sufficient rule. We can, however, provide rules for identification that are sufficient, but that pertain only to recursive models, or special cases of recursive models. Specifically, a sufficient condition for identification is that B is triangular and that Ψ is a diagonal matrix. However, this is the same as saying that recursive models are identified. Indeed, this is the case, and [1] refers to this rule as the *recursive rule* of identification. In combination with the counting rule above, recursive models can be either just identified or overidentified.

A special case of the recursive rule concerns the situation where $B = 0$ and Ψ again a diagonal matrix. Under this condition, the model in (1) reduces to the model in (2). Here too, we can utilize the

counting rule to show that regression models are also just identified.

Note that recursive models place restrictions on the form of B and Ψ and that the identification conditions stated above are directly related to these types of restrictions. Nonrecursive models, however, do not restrict B and Ψ in the same way. Thus, we need to consider identification rules that are relevant to nonrecursive models.

As noted above, the approach to identification arising out of econometrics (see [2]), considers one equation at a time. The concern is whether a true simultaneous equation can be distinguished from a false one formed by a linear combination of the other equations in the model (see [3]). In complex systems of equations, trying to determine linear combinations of equations is a tedious process. One approach would be to evaluate the rank of a given matrix, because if a given matrix is not of full rank, then it means that there exist columns (or rows) that are linear combinations of each other. This leads to developing a *rank condition* for identification.

To motivate the rank and order conditions, consider the simultaneous equation model represented in path analytic form shown in Figure 1. Let p be the number of endogenous variables and let q be the number of exogenous variables. We can write this model as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ 0 & 0 & \gamma_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}. \quad (3)$$

In this example, $p = 2$ and $q = 3$. As a useful device for assessing the rank and order condition, we can

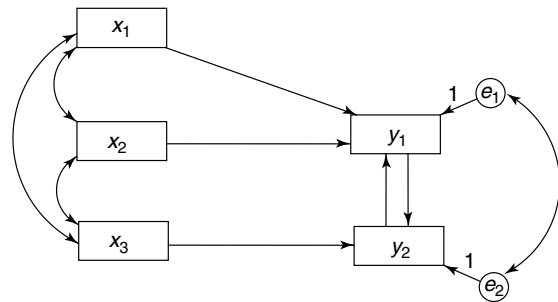


Figure 1 Prototype nonrecursive path model

arrange the structural coefficients in a partitioned matrix A of dimension $p \times s$ as

$$\begin{aligned} \mathbf{A} &= [(\mathbf{I} - \mathbf{B}) | -\Gamma], \\ &= \begin{bmatrix} 1 & -\beta_{12} & -\gamma_{11} & -\gamma_{12} & 0 \\ -\beta_{21} & 1 & 0 & 0 & -\gamma_{23} \end{bmatrix}, \end{aligned} \quad (4)$$

where $s = p + q$. Note that the zeros placed in (4) represent paths that have been excluded (restricted) from the model based on *a priori* model specification. We can represent the restrictions in the first equation of A , say A_1 , as $A_1\phi_1 = 0$, where ϕ_1 is a column vector whose h th element ($h = 1, \dots, s$) is unity and the remaining elements are zero. Thus, ϕ_1 selects the particular element of A_1 for restriction. A similar equality can be formed for A_2 , the second equation in the system. The rank condition states that a necessary and sufficient condition for the identifiability of the first equation is that the rank of $A\phi_1$ must be at least equal to $p - 1$. Similarly for the second equation. The proof of the rank condition is given in [2]. If the rank is less than $p - 1$, then the parameters of the equation are not identified. If the rank is exactly equal to $p - 1$, then the parameters of the equation in question are just identified. If the rank is greater than $p - 1$, then the parameters of the equation are overidentified.

The rank condition can be easily implemented as follows. Delete the columns containing nonzero elements in the row corresponding to the equation of interest. Next, check the rank of the resulting submatrix. If the rank is $p - 1$, then the equation is identified. To take the above example, consider the identification status of the first equation. Recall that for this example, $p - 1 = 1$. According to the procedure just described, the resulting submatrix is

$$\begin{bmatrix} 0 \\ -\gamma_{23} \end{bmatrix}.$$

With the first row zero, the rank of this matrix is one, and, hence, the first equation is identified. Considering the second equation, the resulting submatrix is

$$\begin{bmatrix} -\gamma_{11} & -\gamma_{12} \\ 0 & 0 \end{bmatrix}.$$

Again, because of the zeros in the second row, the rank of this submatrix is 1 and we conclude that the second equation is identified.

A corollary of the rank condition is referred to as the *order condition*. The order condition states that the number of variables (exogenous and endogenous) excluded (restricted) from any of the equations in the model must be at least $p - 1$ [2]. Despite the simplicity of the order condition, it is only a necessary condition for the identification of an equation of the model. Thus, the order condition guarantees that there is a solution to the equation, but it does not guarantee that the solution is unique. A unique solution is guaranteed by the rank condition.

As an example of the order condition, we observe that the first equation has one restriction and the second equation has two restrictions as required by the condition that the number of restrictions must be at least $p - 1$ (here, equal to one). It may be of interest to modify the model slightly to demonstrate how the first equation of the model would not be identified according to the order condition. Referring to Figure 1, imagine a path from x_3 to y_1 . Then the zero in the first row of A would be replaced by $-\gamma_{13}$. Using the simple approach for determining the order condition, we find that there are no restrictions in the first equation and, therefore, the first equation is not identified. Similarly, the first equation fails the rank condition of identification.

This chapter considered identification for recursive and nonrecursive simultaneous equation models. A much more detailed exposition of identification can be found in [2]. It should be pointed out that the above discussion of identification is model-specific and the data play no role. Problems of identification can arise from specific aspects of the data. This is referred to as *empirical identification* and the problem is most closely associated with issues of colinearity.

Briefly, consider a simple linear regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \zeta. \quad (5)$$

If x_1 and x_2 were perfectly collinear, then $x_1 = x_2 = x$, and equation (5) can be rewritten as

$$\begin{aligned} y &= \beta_1 x + \beta_2 x + \zeta, \\ &= (\beta_1 + \beta_2)x + \zeta. \end{aligned} \quad (6)$$

It can be seen from application of the counting rule that (5) is identified, whereas (6) is not. Therefore, the problem of colinearity can induce empirical nonidentification.

4 Identification

References

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, John Wiley & Sons, New York.
- [2] Fisher, F. (1966). *The Identification Problem in Econometrics*, McGraw-Hill, New York.
- [3] Land, K.C. (1973). Identification, parameter estimation, and hypothesis testing in recursive sociological models, in *Structural Equation Models in the Social Sciences*,

A.S. Goldberger & O.D. Duncan, eds, Seminar Press, New York.

(See also **Residuals in Structural Equation, Factor Analysis, and Path Analysis Models; Structural Equation Modeling: Software**)

DAVID KAPLAN

Inbred Strain Study

CLAIRE WADE

Volume 2, pp. 896–898

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Inbred Strain Study

Genetic contributions in complex trait analyses can be controlled or removed by the use of inbred strains as the test subjects. To be regarded as inbred strains, the mammals concerned must have been subjected to at least 20 generations of full-sib mating (the expected inbreeding coefficient will be 98.6% or greater) [1], resulting in near total genetic similarity among strain members. Such levels of inbreeding are most easily achieved with smaller laboratory species such as the mouse. More than 100 inbred strains of mouse are commercially available through the Jackson Laboratories (www.jax.org) alone. Other species (rat, hamster, guinea pig, rabbit, chicken, amphibian, and fish) are less widely represented by commercially available inbred strains.

Variation in the test scores observed within inbred strains will be of either environmental (treatment) origin or otherwise caused by random experimental error, since genetic variation is largely absent. When the same inbred strain is tested across multiple environments or treatments, the source and extent of the environmental effects can be readily ascertained.

The fixed effect of genotype within environment (treatment) can be measured reliably if there is replication of inbred strain and gender within strain when multiple strains are compared.

By combining the resources of strain and treatment in a model in which other environmental variables are controlled, the assessment of genotype by treatment interaction (*see* **Gene-Environment Interaction**) is enabled. The advantage of using different inbred strains over out-bred individuals in such a study is that genotype can be rigorously fitted as a fixed rather than random effect (*see* **Fixed and Random Effects**) in the analysis, and replication within genotype is possible leading to reduced random genetic errors.

Analytical methods commonly applied to the testing of the relative significance of genotypic and treatment effects on behavioural traits include the **analysis of variance** and **multivariate analysis of variance**, **principal component analysis** or principal factor analysis (*see* **Factor Analysis: Exploratory**), **discriminant analysis** (which has now been largely replaced by **logistic regression**), with ANOVA being the most commonly employed analytical method.

The genetic stability of the inbred strains makes them amenable to crossbreeding for Linkage Analysis.

Ascertainment of genetic distinction in measurement between inbred mouse strains can be used to localize regions of the genome responsible for the trait in question using haplotype analysis. Recent studies have shown that many of the commonly used inbred mice share common ancestry over around 67% of their genomes in any pairwise comparison of strains [2]. Should this commonality of ancestry be generally true and should the ancestral regions be randomly placed, then, as more inbred strains are observed for the trait, the genomic regions that fit the phenotypic results for all strains would be reduced in size. Simple observation of inbred strain phenotype in combination with knowledge of ancestral haplotype might reduce the search space for an inferred genomic region by a substantial proportion. This method is most easily applied when a broad region of likelihood of causation is first identified by either linkage analysis or a genome-wide SNP scan. For accurate resolution, SNPs need to be ultimately observed at intervals of 50 kilobases or less. The method is applicable in the inbred mouse because of the nature of its breeding history and is unlikely to be applicable to other inbred species at this time.

Example

To ascertain whether behavioral test standardization was sufficient to prevent inter-laboratory differences in the results of behavioral testing in mice, Wahlsten et al. [3], studied the impact of three laboratory environments on six behavioral test results in the mouse. Testing four mice from each of eight strains, two sexes, and two shipping conditions (locally bred or shipped), 128 mice per laboratory were tested. Factorial analysis of variance (ANOVA) (*see* **Factorial Designs**) was used in the final analysis of the data. Degrees of freedom for individual effects are simply $(g-1)$, where g is the number of genotypes and $(s-1)$, where s is the number of sites (laboratories) tested. The interaction (genotype by site) degrees of freedom is the product of the degrees of freedom for each contributing main effect.

The Table 1 below gives the omega-squared (ω^2) values (*see* **Effect Size Measures**) for three of treatment effects analyzed (that is, those pertaining

2 Inbred Strain Study

Table 1 Results of ANOVA for the elevated plus maze

Variable	<i>N</i>	Genotype (<i>df</i> = 7)	Site (<i>df</i> = 2)	Genotype × site (<i>df</i> = 14)	Multiple <i>R</i> ²
Time in center	379	0.302	0.180	0.134	0.523
Total arms entries	379	0.389	0.327	0.217	0.660
Percent time in open arms	379	0.050 ^a	0.265	N.S.	0.445

^a $p < .01$; all other effects significant at $p < .001$. N.S. indicates $p > .01$.

Values for specific effects are partial omega-squared coefficients.

Source: Table derived from Wahlsten, D., et al. (2003). Different data from different labs: lessons from studies of gene-environment interaction, *Journal of Neurobiology* **54**(1), 283–311.

to the Elevated Plus Maze). Omega-squared is an estimate of the dependent variance accounted for by the independent variable in the population for a fixed effects model, and so is a measure of the importance of that treatment effect in relation to all effects in the model.

$$\omega^2 = \frac{(SS_{\text{effect}} - (DF_{\text{effect}})(MS_{\text{error}}))}{MS_{\text{error}} + SS_{\text{total}}} \quad (1)$$

where SS is the Sum of Squares, DF is the degrees of freedom and MS is the measured Mean Square (SS/DF).

The multiple R-squared (R^2) (see **Multiple Linear Regression**) describes the proportion of all variance accounted for by the corrected model. It is calculated as the sum of squares for the fitted model divided by the total sum of squares.

The researchers concluded that, while there were significant interactions between laboratories and genotypes for the observed effects, the magnitude of the interactions depended upon the measurements in question. The results suggested that test standardization alone is unlikely to completely overcome the influences of different laboratory environments. Most of the larger differences between

inbred strains were able to be successfully replicated across the labs in the study, though strain differences of moderate effects size were less likely to be resolved.

References

- [1] Festing, F.W. (1979). *Inbred Strains in Biomedical Research*, Oxford University Press, New York.
- [2] Wade, C.M., Kulbokas, E.J., Kirby, A.W., Zody, M.C., Mullikin, J.C., Lander, E.S., Lindblad-Toh, K., Daly, M.J. (2002). The mosaic structure of variation in the laboratory mouse genome, *Nature* **420**(6915), 574–8.
- [3] Wahlsten, D., Metten, P., Phillips, T.J., Boehm, S.L. II, Burkhart-Kasch, S., Dorow, J., Doerksen, S., Downing, C., Fogarty, J., Rodd-Henricks, K., Hen, R., McKinnon, C.S., Merrill, C.M., Nolte, C., Schalomon, M., Schlumbohm, J.P., Sibert, J.R., Wenger, C.D., Dudek, B.C. & Crabbe, J.C. (2003). Different data from different labs: lessons from studies of gene-environment interaction, *Journal of Neurobiology* **54**(1), 283–311.

(See also **Selection Study (Mouse Genetics)**)

CLAIRE WADE

Incidence

HANS GRIETENS

Volume 2, pp. 898–899

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Incidence

Incidence is defined as the number of new cases of a disease in a population at a specified interval of time. For instance, the incidence rate per 1000 is calculated as:

$$1000 \left(\frac{\text{Number of new cases beginning during a defined period of time}}{\text{Averaged number in a defined population exposed to risk during that time}} \right)$$

Cumulative incidence refers to the proportion of the population becoming diseased during a specified period of time. Just like **prevalence** rates, incidence rates can be studied in general or in clinical populations. They are useful morbidity rates that can be considered as baseline data or base rates in case they are derived from a general population study. For instance, one can calculate the incidence of new cases of influenza, tuberculosis, or AIDS per 100 000 individuals per year. Studying the incidence of diseases is a main aim of epidemiology [3, 4]. In psychiatry and behavioral science, incidence refers to the number of newly appearing mental disorders (e.g., schizophrenia) or behavioral problems (e.g., hyperactivity) during a certain time period (e.g., a month, a year). It is very difficult to compute incidence rates of mental disorders or behavioral problems, since often it is unclear when symptoms appeared for the first time. For this reason, most epidemiological studies in psychiatry and behavioral science present period prevalence rates – for instance, the studies by Verhulst [6] on childhood psychopathology in the Netherlands, and the Epidemiological Catchment Area Study [5] and the National Comorbidity Survey [1], both conducted in the United States.

There is a dynamic relationship between prevalence and incidence. This relationship can be presented as follows:

$$\begin{aligned} \text{Point prevalence rate} &= \text{Incidence rate} \\ &\times \text{Average duration} \end{aligned} \quad (1)$$

Underlying this formula is the assumption that the average duration of a disease and the incidence are both fairly stable over time. If this is not the case, the relationship becomes much more complex [3, 4].

An incidence study allows the measurement of the rate at which new cases are added to the population of individuals with a certain disease. It is also possible to examine how a disease develops in a population, to check differences in development between populations and between time periods, and to examine the influence of etiological factors. Incidence studies are to be preferred over prevalence studies if one is interested to identify individuals at risk for a disease or risk factors, since prevalence rates are determined by the incidence and the duration of a disease [2]. Treatment characteristics, preventive measures, and factors affecting mortality may affect the duration of a disease or a problem, and so indirectly influence prevalence rates.

References

- [1] Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U. & Kendler, K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey, *Archives of General Psychiatry* **51**, 8–19.
- [2] Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S. & Kupfer, D.J. (1997). Coming to terms with the terms of risk, *Archives of General Psychiatry* **54**, 337–343.
- [3] Lilienfeld, M. & Stolley, P.D. (1994). *Foundations of Epidemiology*, Oxford University Press, New York.
- [4] MacMahon, B. & Trichopoulos, D. (1996). *Epidemiology Principles and Methods*, Little, Brown, Boston.
- [5] Robins, L.N. & Regier, D.A., eds (1991). *Psychiatric Disorders in America*, Free Press, New York.
- [6] Verhulst, F.C. (1999). Fifteen years of research in child psychiatric epidemiology, in *Child Psychiatric Epidemiology. Accomplishments and Future Directions*, H.M. Koot, A.A.M. Crijnen & R.F. Ferdinand, eds, Van Gorcum, Assen, pp. 16–38.

HANS GRIETENS

Incomplete Contingency Tables

SCOTT L. HERSHBERGER

Volume 2, pp. 899–900

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Incomplete Contingency Tables

Whether by design or accident, there may be no observations in one or more cells of a **contingency table**. We refer to such contingency tables as *incomplete* and the data as *sparse*. We can distinguish between two situations in which incomplete tables can be expected [12]:

1. *Structural zeros*. On the basis of our knowledge of the population, we do *not* expect one or more combinations of the factor levels to be observed in a sample. By design, we have one or more empty cells.
2. *Sampling zeros*. Although in the population all possible combinations of factor levels occur, we do not observe one or more of these combinations in our sample. By accident, we have one or more empty cells.

While sampling zeros occur from deficient sample sizes, too many factors, or too many factor levels, structural zeros occur when it is theoretically impossible for a cell to have any observations. For example, let us assume we have two factors, sex (*male, female*) and breast cancer (*yes, no*). While it is medically possible to have observations in the cell representing males who have breast cancer (*male × yes*), the rareness of males who have breast cancer in the population may result in no such cases appearing in our sample. On the other hand, let us say we sample both sex and the frequency of different types of cancers. While the cell representing males who have prostate cancer will have observations, it is impossible to have any observations in the cell representing females who have prostate cancer. Sampling and structural zeros should not be analytically treated the same. While sampling zeros should contribute to the estimation of the model parameters, structural zeros should not.

Invariably, the presence of structural zeros will directly prevent the estimation of certain parameters. Thus, one consequence of fitting **loglinear models** with structural zeros is the necessity for correcting the model degrees of freedom to accurately reflect the number of cells contributing to the analysis and the actual number of parameters estimated

from these cells, not only those representing theoretically impossible combinations (e.g., females with prostate cancer) but also those indirectly preventing the estimation. Typically, ensuring that the degrees of freedom of the model is correct is the most serious problem caused by structural zeros.

In contrast, fitting loglinear models to tables with sampling zeros can be more problematic, because of the infinite parameter estimates that may arise if the tables have margins with zero entries. In addition, the failure to satisfy the large sample assumptions may mean that the actual null distributions of the generalized likelihood ratio (G^2) or the chi-squared (χ^2) test approximations to the true chi-squared (X^2) distribution are far from the intended asymptotic approximations, leading to mistakes in model selection. Owing to inaccuracies in their approximation, G^2 and χ^2 used as goodness-of-fit statistics can mislead as to which of a series of hierarchically nested models is best. Several investigators (e.g., [2, 3, 5, 6], and [8]) have studied the effects of empty cells on G^2 and χ^2 . The basic findings may be summarized as follows: (a) the correctness of the approximations is largely a function of the ratio n/N , where n = the number of cells and N = the total sample size, (b) as n/N becomes smaller, the approximations become less accurate, and (c) the chi-squared approximation to χ^2 is more valid than G^2 for testing models when n/N is small, particularly when $n/N < 5$. However, the maximum value of n/N that is permissible for G^2 and χ^2 to be accurate approximations undoubtedly varies from situation to situation.

If in a particular case n/N is judged too large, there are various strategies available to ameliorate the situation. If sensible theoretically to do so, a simple but often effective strategy is to combine categories together to prevent a cell from having no observations. Exact methods (see **Exact Methods for Categorical Data**) can also be used if the loglinear model is not too large, obviating the need for approximations to X^2 altogether [7, 11]. When exact methods are not feasible, resampling methods such as **bootstrapping** can provide a good approximation to exact distributions [10]. In addition, test statistics from asymptotic approximations that are more accurate when n/N is small can be used instead of the traditional G^2 and χ^2 approximations [4]. Some test statistics are based on refinements to G^2 and χ^2 , whereas others are entirely new, such as the power divergence statistics (λ) introduced by Read and Cressie [9].

2 Incomplete Contingency Tables

One strategy that is often used inadvisably is to add a small constant, such as $1/2$, to cells counts – previously empty cells are no longer empty. The problem with this strategy is that it tends to increase the apparent equality of the cells' frequencies, resulting in a loss of **power** for finding significant effects. If the strategy of adding a constant is adopted, an extremely small constant should be used, much smaller than $1/2$. Agresti [1] recommends a constant on the order of 10^{-8} . He also recommends conducting a **sensitivity analysis** in which the analysis is repeated using different constants, in order to evaluate the relative effects of the constants on parameter estimation and model testing.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, New York.
- [2] Berry, K.J. & Mielke, P.W. (1986). R by C chi-square analyses with small expected cell frequencies, *Educational & Psychological Measurement* **46**, 169–173.
- [3] Berry, K.J. & Mielke, P.W. (1988). Monte Carlo comparisons of the asymptotic chi-square and likelihood ratio tests with the nonasymptotic chi-square tests for sparse $r \times c$ tables, *Psychological Bulletin* **103**, 256–264.
- [4] Greenwood, P.E. & Nikulin, M.S. (1996). *A Guide to Chi-squared Testing*, John Wiley & Sons, New York.
- [5] Koehler, K. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables, *Journal of the American Statistical Association* **81**, 483–493.
- [6] Koehler, K. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials, *Journal of the American Statistical Association* **75**, 336–344.
- [7] Mielke, P.W. & Berry, K.J. (1988). Cumulant methods for analyzing independence of r-way contingency tables and goodness-of-fit frequency data, *Biometrika* **75**, 790–793.
- [8] Mielke, P.W., Berry, K.J. & Johnston, J.E. (2004). Asymptotic log-linear analysis: some cautions concerning sparse frequency tables, *Psychological Reports* **94**, 19–32.
- [9] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.
- [10] von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: results of a Monte Carlo study, *Methods of Psychological Research* **2**, 29–48.
- [11] Whittaker, J. (1990). *Graphical Methods in Applied Mathematical Multivariate Statistics*, John Wiley & Sons, New York.
- [12] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum Associates, Hillsdale.

SCOTT L. HERSHBERGER

Incompleteness of Probability Models

RANALD R. MACDONALD

Volume 2, pp. 900–902

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Incompleteness of Probability Models

This article concerns the claim that probabilities cannot take account of everything that could be relevant to a particular uncertainty. It has implications for the extent to which probabilities can set standards for people's judgments and for theories concerning the nature of probability. These issues are dealt with in (see **Probability: Foundations of**).

Aristotle held that what we take to be the likelihood of an event occurring depends both on the frequency of occurrence of the type of event we take it to be, and on how representative it is of that sort of event. '... if the thing in question *both* happens *oftener* as we represent it *and* happens more *as* we represent it, the probability is particularly great' [12, p. 1402a]. Thus Aristotle supposed that an event's likelihood is related to how it is represented. Probabilities operate in the same way. A probability describes the uncertainty of an occasion where one of a number of mutually exclusive types of outcome may occur. Where a probability is assigned a value, someone has had to define the outcome types and determine the probability value or at least the means of determining that value. The assignment may seem obvious in games of chance, but even here there are possibilities of bias, cheating, or player incompetence that could be taken into account. However, most uncertainties do not relate to explicit games of chance, and here the assignment of probabilities is more obviously subjective.

The inherent subjectivity in the origins of probabilities means that probabilities are less than ideal because they take no account of self-referential uncertainty. The problem arises because a probability is a description of an uncertainty, and there is always uncertainty about the accuracy of any description that the description itself does not incorporate. Both **Fisher** [5] and Reichenbach [11] postulated complex systems of hierarchical probabilities to deal with this problem. However, although the uncertainty associated with a probability distribution can be addressed by postulating a higher-order probability distribution, there is uncertainty associated with the new characterization of uncertainty, and so an unlimited number of higher-order probability

distributions could be postulated without the problem being resolved.

A second limitation of probability models is that the outcomes they model are either finite or involve continuous variables that can be approximated by finite models to whatever accuracy is desired. States of the world are more complex than this. Cantor showed that there are many higher-order infinities of possibilities that can be imagined. In particular, there is more than a single infinity of points in a continuum [8]. It follows that probability models are not complex enough to capture the uncertainty regarding an outcome that lies at a point on a continuum, and there is no reason why outcomes should not be more complex, for example, the shape of an object (see [10] for the complexities of specifying the outline of an object). Probability models cannot conceivably incorporate all the possible distinctions between what might be regarded as outcomes, let alone characterize the uncertainties associated with all distinguishable outcomes [9].

Lest the above comments be seen to be mere mathematical technicalities that in practice can be ignored, it is worth thinking about the problem psychologically. What people perceive as happenings in the world have more substance and are less well-defined than the events in probability models. Consider a murder – what constitutes the event of the killing? When does it start – with the planning, the acquiring of the means, the execution of the act of murder, or when the victim is fatally injured? Where does it take place – where the murderer is, where the victim is, or both, and does it move as the killing takes place, for example, does it follow the mortally wounded victim until death occurs? There is even an issue about exactly when a person dies [14]. A little thought shows that events that take place in people's lives (births, deaths, marriages, etc.) do not occur instantaneously. They take place over time and space and can be incorporated into larger events or broken down into component events without any apparent limit. A complete characterization of the uncertainty associated with a particular happening would seem to require that the happening itself be fully specified, yet such a specification seems inconceivable. To apply a probability model, there has to be a limit to the distinctions that can be made between events. Happenings characterized by probabilities cannot be extended or broken down to an unlimited extent. Anscombe [2] put it nicely when

2 Incompleteness of Probability Models

she said that events have to be ‘cut and dried’ in order to be assigned probabilities. Because of this, a probability may always be improved by making more distinctions between events regardless of however well an associated probability model may appear to be consistent with the available evidence.

Fisher [4] considered what an ideal probability model would be like starting with a ‘cut and dried’ outcome. Such a model would classify events into sets containing no identifiable subsets – that is to say, sets that could not be further broken down into subsets where the outcome in question was associated with different probabilities. This involves postulating sets of equivalent or exchangeable events, which is how **de Finetti** conceived of ideal probability models [3]. Probabilities assigned to events in such sets could not be improved on by taking additional variables into account. Such probabilities are important as the laws of large numbers ensure that, in this case, each probability has a ‘correct’ value – the limit of the relative frequency of equivalent events as n increases. Betting that one of the equivalent events will occur taking the odds to be $p:(1-p)$, where p is this ‘correct’ probability, will do better than any other value over the long term (Dutch book theorem: [7]). Unfortunately, one can never know that any two events are equivalent let alone that any set of events has no identifiable subsets.

If all the uncertainty concerning happenings in the world could be captured by a probability, then it would have a correct value, and optimal solutions to probability problems regarding real-life events would exist. As things stand, however, probabilities regarding real-life events are based on analogies between the uncertainties in the world and models originating in people’s heads. Furthermore, probability models do not explain how these analogies come to be formed [13]. As Aristotle foresaw around 2000 years before probability was invented, what probability should be assigned to an event depends on how it is characterized, and that is a matter for reasoned argument. Inferences from probabilities, be they relative frequencies, estimates of people’s beliefs, the P values in statistical tests, or posterior probabilities in **Bayesian statistics** are necessarily subject to revision when it is argued that the events being modeled should be characterized in more detail or that

the probability models should be modified to take account of possibilities their inventors had not foreseen. Unforeseen occurrences can also pose problems for theories that suppose that probabilities are the limits of relative frequencies because the relative frequency of events that have never occurred is zero, but clearly novel events occur [6]. Probabilities and their interpretation should be seen as matters for debate rather than as the necessary consequences of applying the correct analysis to a particular problem [1].

References

- [1] Abelson, R.P. (1995). *Statistics as Principled Argument*, Lawrence Erlbaum Associates, Hillsdale.
- [2] Anscombe, G.E.M. (1979). Under a description, *Noûs* **13**, 219–233.
- [3] de Finetti, B. (1972). *Probability, Induction & Statistics: The Art of Guessing*, Wiley, London.
- [4] Fisher, R.A. (1957/1973). *Statistical Methods and Scientific Inference*, 3rd Edition, Hafner, New York.
- [5] Fisher, R.A. (1957). The underworld of probability, *Sankhya* **18**, 201–210.
- [6] Franklin, J. (2001). *The Science of Conjecture: Evidence and Probability before Pascal*, John Hopkins University Press, Baltimore.
- [7] Howson, C. & Urbach, P. (1993). *Scientific Reasoning: the Bayesian Approach*, 2nd Edition, Open Court, Peru, Illinois.
- [8] Luchins, A.S. & Luchins, E.H. (1965). *Logical Foundations of Mathematics for Behavioral Scientists*, Holt Rinehart & Winston, New York.
- [9] Macdonald, R.R. (2000). The limits of probability modelling: A serendipitous tale of goldfish, transfinite numbers and pieces of string, *Mind and Society* **1**(part 2), 17–38.
- [10] Mandelbrot, B. (1982). *The Fractal Geometry of Nature*, Freeman, New York.
- [11] Reichenbach, H. (1970). *The Theory of Probability*, 2nd Edition, University of California Press, Berkeley.
- [12] Ross, W.D. (1966). *The Works of Aristotle Translated into English*, Clarendon, Oxford.
- [13] Searle, J.R. (2001). *Rationality in Action*, MIT Press, Massachusetts.
- [14] Thompson, J.J. (1971). The time of a killing, *Journal of Philosophy* **68**, 115–132.

RANALD R. MACDONALD

Independence: Chi-square and Likelihood Ratio Tests

BRUCE L. BROWN AND KENT A. HENDRIX

Volume 2, pp. 902–907

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Independence: Chi-square and Likelihood Ratio Tests

The simplest and most common test of independence is the chi-square test. It is a *nonparametric* test of significance that is used to analyze *categorical variables*. Categorical variables can often be on a nominal scale, indicating that the categories may have no ordering. Chi-square tests of independence can therefore be applied to a very broad range of data, qualitative as well as quantitative, anything that one can categorize and count (*see Contingency Tables; Goodness of Fit for Categorical Variables*).

Suppose that occupational data and musical preference data are gathered from a random sample of 1292 citizens of an Alaskan community. For each respondent, the data will consist of two qualitative categories, one for occupation and one for musical preference data. The primary qualitative data could look like this:

Respondent 1: lumberjack, preference for The Eagles
Respondent 2: merchant, preference for Led Zeppelin
Respondent 3: lumberjack, preference for Led Zeppelin
Respondent 4: fisherman, preference for Jethro Tull
Respondent 5: fisherman, preference for Led Zeppelin.

Data of this kind for all 1292 respondents can be collected into a **contingency table**, as shown below in Table 1:

These data are *frequency* data – they indicate how frequently each combination of categories in the 3×3 table occurs. The chi-square test of independence may be used to analyze this contingency table to determine whether there is a relationship between occupation and musical preference. The null

hypothesis holds that the two are not related – that they are independent.

The Chi-square Test of Independence

The 3×3 contingency table given above for the occupation and musical preference study (minus the row and column marginal totals) is the *observed* frequency matrix, symbolized as O :

$$O = \begin{bmatrix} 230 & 246 & 9 \\ 130 & 478 & 47 \\ 15 & 83 & 54 \end{bmatrix} \quad (1)$$

The matrix of expected values (the frequency values that would be expected if the two variables of occupation and musical preference were independent) is found by obtaining the product of the *row marginal total* and the *column marginal total* corresponding to each cell, and dividing this product by the *grand total*. This is done for each cell as shown below:

$$E = \begin{bmatrix} \left(\frac{485 \times 375}{1292}\right) & \left(\frac{485 \times 807}{1292}\right) & \left(\frac{485 \times 110}{1292}\right) \\ \left(\frac{655 \times 375}{1292}\right) & \left(\frac{655 \times 807}{1292}\right) & \left(\frac{655 \times 110}{1292}\right) \\ \left(\frac{152 \times 375}{1292}\right) & \left(\frac{152 \times 807}{1292}\right) & \left(\frac{152 \times 110}{1292}\right) \end{bmatrix} \\ = \begin{bmatrix} 140.77 & 302.94 & 41.29 \\ 190.11 & 409.12 & 55.77 \\ 44.12 & 94.94 & 12.94 \end{bmatrix} \quad (2)$$

The *observed minus expected* matrix is

$$O - E = \begin{bmatrix} 89.23 & -56.94 & -32.29 \\ -60.11 & 68.88 & -8.77 \\ -29.12 & -11.94 & 41.06 \end{bmatrix} \quad (3)$$

The fundamental equation for chi-square is to obtain the *squared difference between the observed value (O) and the corresponding expected value (E)*,

Table 1 Contingency table relating musical preference to occupation

	The Eagles	Led Zeppelin	Jethro Tull	Row totals
Lumberjacks	230	246	9	485
Fishermen	130	478	47	655
Merchants	15	83	54	152
Column totals	375	807	110	1292

2 Independence: Chi-square and Likelihood Ratio Tests

divided by the expected value (E) for each cell, and then to sum all of these:

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{89.23^2}{140.77} + \frac{(-56.94)^2}{302.94} + \frac{(-32.29)^2}{41.29} \\ &\quad + \frac{(-60.11)^2}{190.11} + \frac{68.88^2}{409.12} + \frac{(-8.77)^2}{55.77} \\ &\quad + \frac{(-29.12)^2}{44.12} + \frac{(-11.94)^2}{94.94} + \frac{41.06^2}{12.94} \\ &= 275.48\end{aligned}\quad (4)$$

The degrees of freedom value for this chi-square test is the product $(R - 1) \times (C - 1)$, where R is the number of rows in the contingency table and C is the number of columns. For this 3×3 contingency table $(R - 1) \times (C - 1) = (3 - 1) \times (3 - 1) = 4$ df. In a table of critical values for the chi-square distribution, the value needed to reject the null hypothesis at the 0.001 level for 4 degrees of freedom is found to be 18.467. The obtained value of 275.48 exceeds this, so the null hypothesis of independence can be rejected at the 0.001 level. That is, these data give substantial evidence that occupation and musical preference are systematically related, and therefore not independent. Although this example is for a 3×3 contingency table, the chi-square test of independence may be used for two-way tables with any number of rows and any number of columns.

Expanding Chi-square to Other Kinds of Hypotheses

Other chi-square tests are possible besides the test of independence. Suppose in the example just given that the 1292 respondents were a random sample of workers in a particular Alaskan city, and that you wish to test workers' relative preferences for these three musical groups. In other words, you wish to test the null hypothesis that workers in the population from which this random sample is obtained are equally distributed in their preferences for the three musical groups. The *observed* (O) matrix is the column marginal totals:

$$O = [375 \quad 807 \quad 110] \quad (5)$$

The null hypothesis in this case is that the three column categories have equal frequencies in the population, so the *expected* (E) matrix consists of the total 1292 divided by 3 (which gives 430.67), with this entry in each of the three positions. The $O - E$ matrix is therefore

$$\begin{aligned}O - E &= [375 \ 807 \ 110] - [430.67 \ 430.67 \ 430.67] \\ &= [-55.67 \ 376.33 \ -320.67]\end{aligned}\quad (6)$$

The obtained chi-square statistic is

$$\chi_{\text{columns}}^2 = \sum \frac{(O - E)^2}{E} = 574.81 \quad (7)$$

This chi-square test has 2 degrees of freedom, three columns minus one ($C - 1$), for which the critical ratio at the 0.001 level is 13.816. The null hypothesis of equal preferences for the three groups can therefore be rejected at the 0.001 level.

So far, two chi-square statistics have been calculated on this set of data, a test of independence and a test of equality of proportions across columns (musical preference). The third chi-square test to be computed is the test for row effects (occupation). The *observed* (O) matrix is the row marginal totals:

$$O = \begin{bmatrix} 485 \\ 655 \\ 152 \end{bmatrix} \quad (8)$$

The null hypothesis in this case is that the three row categories (occupations) have equal frequencies within the population. Therefore, the *expected* (E) matrix consists of the total 1292 divided by 3 (which gives 430.67), with this same entry in each of the three positions. The obtained chi-square statistic is

$$\chi_{\text{rows}}^2 = \sum \frac{(O - E)^2}{E} = 304.01 \quad (9)$$

With 2 degrees of freedom, the critical ratio for significance at the 0.001 level is 13.816, so the null hypothesis of equal occupational distribution can be rejected at the 0.001 level.

The fourth and final chi-square statistic to be computed is that for the total matrix. The *observed* (O) matrix is again the 3×3 matrix of observed frequencies, just as it was for the $R \times C$ test of independence given above. However, even though the observed matrix is the same, the null hypothesis (and therefore the expected matrix) is very different.

This test is an omnibus test of whether there is any significance in the matrix considered as a whole (row, column, or $R \times C$ interaction), and the null hypothesis is therefore the hypothesis that all of the cells are equal. All nine entries in the *expected* (E) matrix are 143.56, which is one-ninth of 1292. The $O - E$ matrix is therefore

$$\begin{aligned} O - E &= \begin{bmatrix} 230 & 246 & 9 \\ 130 & 478 & 47 \\ 15 & 83 & 54 \end{bmatrix} \\ &\quad - \begin{bmatrix} 143.56 & 143.56 & 143.56 \\ 143.56 & 143.56 & 143.56 \\ 143.56 & 143.56 & 143.56 \end{bmatrix} \\ &= \begin{bmatrix} 86.44 & 102.44 & -134.56 \\ -13.56 & 334.44 & -96.56 \\ -128.56 & -60.56 & -89.56 \end{bmatrix} \quad (10) \end{aligned}$$

The obtained chi-square statistic is

$$\chi_{\text{total}}^2 = \sum \frac{(O - E)^2}{E} = 1293.20 \quad (11)$$

This chi-square has 8 degrees of freedom, the total number of cells minus one ($RC - 1$), with a critical ratio, at the 0.001 level, of 26.125. The null hypothesis of no differences of any kind within the entire data matrix can therefore be rejected at the 0.001 level.

These four kinds of information can be obtained from a contingency table using chi-square. However, there is a problem in using chi-square in this way. The total analysis should have a value that is equal to the sum of the other three analyses that make it up. The values should be additive, but this is not the case as shown with the example data:

$$\begin{aligned} \chi_{\text{total}}^2 &= 1293.20 \neq 1154.30 \\ &= 304.01 + 574.81 + 275.48 \\ &= \chi_{\text{rows}}^2 + \chi_{\text{columns}}^2 + \chi_{R \times C}^2 \quad (12) \end{aligned}$$

As will be shown in the next section, this additivity property *does* hold for the likelihood ratio G^2 statistic of log-linear analysis. Log-linear analysis is supported by a more coherent mathematical theory than chi-square that enables this additivity property to hold, and also enables one to use the full power of linear models applied to categorical data (see **Log-linear Models**).

Multiplicative Models, Additive Models, and the Rationale for Log-linear

The procedure given above for obtaining the matrix of expected frequencies is a direct application of the *multiplication rule of probabilities for independent joint events*:

$$P(A \text{ and } B) = P(A) \times P(B) \quad (13)$$

For example, for the Alaskan sample described above, the probability of a respondent being a lumberjack is

$$P(A) = \frac{\text{frequency of lumberjacks}}{\text{total frequency}} = \frac{485}{1292} = 0.375 \quad (14)$$

Similarly, the probability of a respondent preferring Jethro Tull music is

$$\begin{aligned} P(B) &= \frac{\text{frequency of Jethro Tull preference}}{\text{total frequency}} \\ &= \frac{110}{1292} = 0.085 \quad (15) \end{aligned}$$

If these two characteristics were independent of one another, the joint probability of a respondent being a lumberjack *and* also preferring Jethro Tull music would (by the multiplication rule for joint events) be

$$\begin{aligned} P(A \text{ and } B) &= P(A) \times P(B) \\ &= (0.375) \cdot (0.085) = 0.032 \quad (16) \end{aligned}$$

Multiplying this probability by 1292, the number in the sample, gives 41.3, which is (to one decimal place) the value of the expected frequency of the lumberjack/Jethro Tull cell.

Interaction terms in **analysis of variance** (ANOVA) use a similar kind of ‘observed minus expected’ logic and an analogous method for obtaining expected values. The simple definitional formula for the sum of squares of a two-way interaction is

$$\begin{aligned} SS(AB) &= n \sum \sum (\bar{X}_{ab} - \bar{X}_{a.} - \bar{X}_{.b} + \bar{X}_{..})^2 \\ &= n \sum \sum (\bar{X}_{ab} - \bar{E}_{ab})^2 \quad (17) \end{aligned}$$

This sum can be decomposed as the multiplicative factor n (cell frequency) times the ‘sum of squared

4 Independence: Chi-square and Likelihood Ratio Tests

deviations from additivity', where 'deviations from additivity' refers to the deviations of the observed means (O) from the expected means (E), those that would be expected if an additive model were true. This is because the additive model for creating means is given by

$$\bar{E}_{ab} = \bar{X}_{a.} + \bar{X}_{.b} - \bar{X}_{..} \quad (18)$$

The two processes are analogous. To obtain the expected cell means for ANOVA, one sums the marginal row mean and the marginal column mean and subtracts the grand mean. To obtain expected cell frequencies in a contingency table, one multiplies marginal row frequencies by marginal column frequencies and divides by the total frequency. By taking logarithms of the frequency values, one converts the multiplicative computations of contingency table expected values into the additive computations of ANOVA, thus making frequency data amenable to linear models analysis. This log-linear approach comes with a number of advantages: it enables one to test three-way and higher contingency tables, to additively decompose test statistics, and in general to apply powerful general linear models analysis to categorical data. The log-linear model will be briefly demonstrated for two-way tables.

Log-linear Models

The observed and expected matrices for each of the four tests demonstrated above are the same for log-linear analysis as for the chi-square analysis. All that differs is the formula for calculating the likelihood ratio statistic G^2 . It is given as two times the sum over all the cells of the following quantity: *the observed value times the natural logarithm of the ratio of the observed value to the expected value*.

$$G^2_{\text{total}} = 2 \sum O \log_e \left(\frac{O}{E} \right) \quad (19)$$

Four likelihood ratio statistics will be demonstrated, corresponding to the four chi-square statistics just calculated. The first is the test of the row by column interaction. This is called *the likelihood ratio test for independence*. The likelihood ratio statistic for this test (using O and E values obtained above) is calculated as

$$G^2_{RC} = 2 \sum O \log_e \left(\frac{O}{E} \right)$$

$$= 2 \times \left[\begin{aligned} &230 \times \log_e \left(\frac{230}{140.77} \right) \\ &+ 246 \times \log_e \left(\frac{246}{302.94} \right) \\ &+ 9 \times \log_e \left(\frac{9}{41.29} \right) \\ &+ 130 \times \log_e \left(\frac{130}{190.11} \right) \\ &+ 478 \times \log_e \left(\frac{478}{409.12} \right) \\ &+ 47 \times \log_e \left(\frac{47}{55.77} \right) \\ &+ 15 \times \log_e \left(\frac{15}{44.12} \right) \\ &+ 83 \times \log_e \left(\frac{83}{94.94} \right) \\ &+ 54 \times \log_e \left(\frac{54}{12.94} \right) \end{aligned} \right] \quad (20)$$

$$= 229.45$$

This likelihood ratio has the same degrees of freedom as the corresponding chi-square, $(R - 1) \times (C - 1) = 4$, and it is also looked up in an ordinary chi-square table. The critical ratio for significance at the 0.001 level with 4 degrees of freedom is 18.467. The null hypothesis of independence can therefore be rejected at the 0.001 level.

The likelihood ratio statistic for rows is

$$\begin{aligned} G^2_{\text{rows}} &= 2 \sum O \log_e \left(\frac{O}{E} \right) \\ &= 2 \times \left[\begin{aligned} &485 \times \log_e \left(\frac{485}{430.67} \right) \\ &+ 655 \times \log_e \left(\frac{655}{430.67} \right) \\ &+ 152 \times \log_e \left(\frac{152}{430.67} \right) \end{aligned} \right] \\ &= 347.93 \end{aligned} \quad (21)$$

This test has a 0.001 critical ratio of 13.816 with $2(R - 1)$ degrees of freedom, so the null hypothesis can be rejected at the 0.001 level.

The third likelihood ratio statistic to be calculated is that for columns:

$$G^2_{\text{columns}} = 2 \sum O \log_e \left(\frac{O}{E} \right) = 609.50 \quad (22)$$

which is also significant at the 0.001 level.

The likelihood ratio for the total matrix is

$$G^2_{\text{total}} = 2 \sum O \log_e \left(\frac{O}{E} \right) = 1186.88 \quad (23)$$

With 8 degrees of freedom and a critical ratio of 26.125, this test is also significant at the 0.001 level.

The additivity property holds with these four likelihood ratio statistics. That is, the sum of the obtained G^2 values for rows, columns, and $R \times C$ interaction is equal to the obtained G^2 value for the total matrix:

$$\begin{aligned} G^2_{\text{total}} &= 1186.88 = 347.93 + 609.50 + 229.45 \\ &= G^2_{\text{rows}} + G^2_{\text{columns}} + G^2_{R \times C} \end{aligned} \quad (24)$$

The history of log-linear models for categorical data is given by Imrey, Koch, and Stokes [3], and detailed accounts of the mathematical development are given by Agresti [1], and by Imrey, Koch, and Stokes [4]. Marascuilo and Levin [6] give a particularly readable account of how the logarithmic transformation enables one to analyze categorical data with the general linear model, and Brown, Hendrix, and Hendrix [2] demonstrate the convergence of chi-square and ANOVA through log-linear models with simplest case data.

The calculations involved in both the chi-square analysis and also the log-linear analysis are simple enough to be easily accomplished using a spreadsheet, such as Microsoft Excel, Quattro Pro, ClarisWorks, and so on. They can also be accomplished

using computer statistical packages such as SPSS and SAS. Chi-square analysis can be accomplished using the FREQ procedure of SAS, and log-linear analysis can be accomplished in SAS using CATMOD (see **Software for Statistical Analyses**). Landau and Everitt [5] demonstrates (in Chapter 3) how to use SPSS to do chi-square analysis and also how to do cross-tabulation of categorical and continuous data.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, Hoboken.
- [2] Brown, B.L., Hendrix, S.B. & Hendrix, K.A. (in preparation). *Multivariate for the Masses*, Prentice-Hall, Upper Saddle River.
- [3] Imrey, P.B., Koch, G.G. & Stokes, M.E. (1981). Categorical data analysis: some reflections on the log-linear model and logistic regression. Part I: historical and methodological overview, *International Statistics Review* **49**, 265–283.
- [4] Imrey, P.B., Koch, G.G. & Stokes, M.E. (1982). Categorical data analysis: some reflections on the log-linear model and logistic regression. Part II: data analysis, *International Statistics Review* **50**, 35–63.
- [5] Landau, S. & Everitt, B.S. (2004). *A Handbook of Statistical Analyses Using SPSS*, Chapman & Hall, Boca Raton.
- [6] Marascuilo, L.A. & Levin, J.R. (1983). *Multivariate Statistics in the Social Sciences: A Researcher's Guide*, Brooks-Cole, Monterey.

BRUCE L. BROWN AND KENT A. HENDRIX

Independent Component Analysis

JAMES V. STONE

Volume 2, pp. 907–912

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Independent Component Analysis

Introduction

‘I shouldn’t be surprised if it hailed a good deal tomorrow’, Eeyore was saying ... ‘Being fine to-day doesn’t Mean Anything. It has no sig – what’s that word? Well, it has none of that.’

– The House at Pooh Corner, AA Milne, 1928.

Most measured quantities are actually mixtures of other quantities. Typical examples are (a) sound signals in a room with several people talking simultaneously, (b) an electroencephalogram (EEG) signal, which contains contributions from many different brain regions, and (c) a person’s height, which is determined by contributions from many different genetic and environmental factors. Science is, to a large extent, concerned with establishing the precise nature of the component processes responsible for a given set of measured quantities, whether these involve height, EEG signals, or even IQ. Under certain conditions, the signals underlying measured quantities can be recovered by making use of ICA. ICA is a member of a class of *blind source separation* (BSS) methods.

The success of ICA depends on one key assumption regarding the nature of the physical world. This assumption is that independent variables or signals¹ are generated by different underlying physical processes. If two signals are independent, then the value of one signal cannot be used to predict anything about the corresponding value of the other signal. As it is not usually possible to measure the output of a single physical process, it follows that most measured signals must be mixtures of independent signals. Given such a set of measured signals (i.e., mixtures), ICA works by finding a transformation of those mixtures, which produces independent signal components, on the assumption that each of these independent component signals is associated with a different physical process. In the language of ICA, the measured signals are known as *signal mixtures*, and the required independent signals are known as *source signals*.

ICA has been applied to separation of different speech signals [1]², analysis of EEG data [6], functional magnetic resonance imaging (fMRI) data [7], image processing [2], and as a model of biological image processing [10].

Before embarking on an account of the mathematical details of ICA, a simple, intuitive example of how ICA could separate two speech signals is given. However, it should be noted that this example could equally well apply to *any physically measured set of signals, and to any number of signals* (e.g., images, biomedical data, or stock prices).

Applying ICA to Speech Data

Consider two people speaking at the same time in a room containing two microphones, as depicted in Figure 1. If each voice signal is examined at a fine time scale, then it is apparent that the amplitude of one voice at any given point in time is unrelated to the amplitude of the other voice at that time. The reason that the amplitudes of two voices are unrelated is that they are generated by two unrelated physical processes (i.e., by two different people). If we know that the voices are unrelated, then one key strategy for

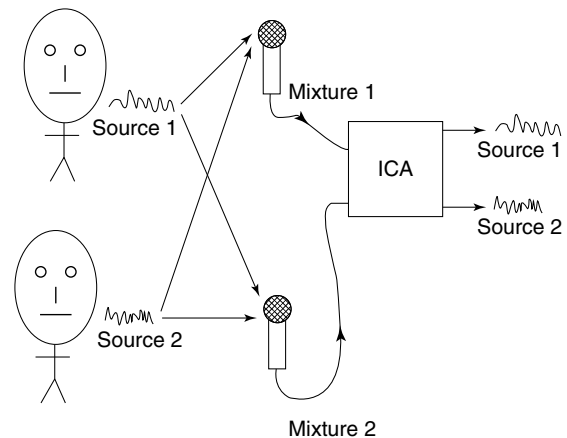


Figure 1 ICA in a nutshell: If two people speak at the same time in a room containing two microphones, then the output of each microphone is a *mixture* of two voice signals. Given these two *signal mixtures*, ICA can recover the two original voices or *source signals*. This example uses speech, but ICA can extract source signals from any set of two or more measured signal mixtures, where each signal mixture is assumed to consist of a linear mixture of source signals (see section ‘Mixing Signals’)

separating voice mixtures (e.g., microphone outputs) into their constituent voice components is to extract unrelated time-varying signals from these mixtures. The property of being unrelated is of fundamental importance.

While it is true that two voice signals are unrelated, this informal notion can be captured in terms of *statistical independence* (see **Probability: An Introduction**), which is often truncated to *independence*. If two or more signals are statistically independent of each other, then the value of one signal provides no information regarding the value of the other signals.

The Number of Sources and Mixtures

One important fact about ICA is often not appreciated. Basically, there must usually be at least as many different mixtures of a set of source signals as there are source signals (but see [9]). For the example of speech signals, this implies that there must be at least as many microphones (different voice mixtures) as there are voices (source signals).

Effects of Mixing Signals

When a set of two or more independent source signals are mixed to make a corresponding set of signal mixtures, as shown in Figure 1, three effects follow.

- *Independence*. Whereas source signals are independent, their signal mixtures are not. This is because each source signal contributes to every mixture, and the mixtures cannot, therefore, be independent.
- *Normality*. The **central limit theorem** ensures that a signal mixture that is the sum of almost any signals yields a bell-shaped, *normal* or *Gaussian* histogram. In contrast, the **histogram** of a typical source signal has a non-Gaussian structure (see Figure 2).
- *Complexity*. The complexity of any mixture is greater than (or equal to) that of its simplest (i.e., least complex) constituent source signal. This ensures that extracting the least complex

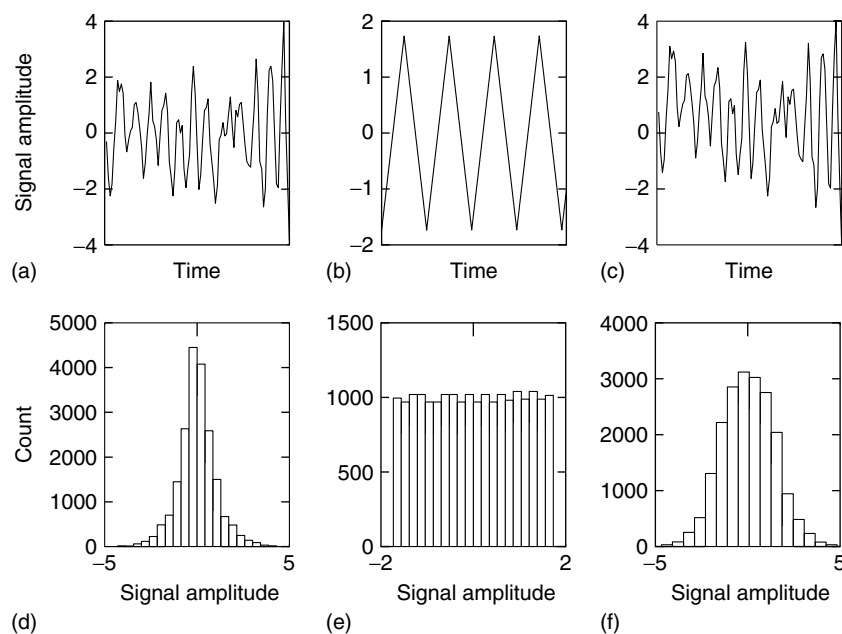


Figure 2 Signal mixtures have Gaussian or normal histograms. Signals (top row) and corresponding histograms of signal values (bottom row), where each histogram approximates the *probability density function* (pdf) of one signal. The top panels display only a small segment of the signals used to construct displayed histograms. A speech source signal (a), and a histogram of amplitude values in that signal (d). A sawtooth source signal (b), and its histogram (e). A signal mixture (c), which is the sum of the source signals on the left and middle, and its bell-shaped histogram (f)

signal from a set of signal mixtures yields a source signal [9].

These three effects can be used either on their own or in combination to extract source signals from signal mixtures. The effects labeled *normality* and *complexity* are used in **projection pursuit** [5] and *complexity pursuit* [4, 8], respectively, and the effects labeled *independence* and *normality* are used together in ICA (also see [9]).

Representing Multiple Signals

A speech source signal s_1 is represented as $s_1 = (s_1^1, s_1^2, \dots, s_1^N)$, where s_1 adopts amplitudes s_1^1 , then s_1^2 , and so on; superscripts specify time and subscripts specify signal identity (e.g., speaker identity). We will be considering how to mix and unmix a *set* of two or more signals, and we define a specific set of two time-varying speech signals s_1 and s_2 in order to provide a concrete example. Now, the amplitudes of both signals can be written as a *vector variable* $\mathbf{s}$, which can be rewritten in one of several mathematically equivalent forms:

$$\mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \quad (1)$$

$$= \begin{pmatrix} (s_1^1, s_1^2, \dots, s_1^N) \\ (s_2^1, s_2^2, \dots, s_2^N) \end{pmatrix}. \quad (2)$$

We introduce the transpose operator, which simply transforms rows into columns (or vice versa), and is defined by $\mathbf{s} = (s_1, s_2)^T$.

Mixing Signals

The different distance of each source (i.e., person) from a microphone ensures that each source contributes a different amount to the microphone's output. The microphone's output is, therefore, a *linear* mixture x_1 that consists of a weighted sum of the two source signals $x_1 = as_1 + bs_2$, where the *mixing coefficients* a and b are determined by the distance of each source from each microphone. As we are concerned here with unmixing a set of two signal mixtures (see Figure 1), we need another microphone in a different location from the first. In this case, the microphone's output x_2 is $x_2 = cs_1 + ds_2$, where the mixing coefficients are c and d .

Unmixing Signals

Generating mixtures from source signals in this linear manner ensures that each source signal can be recovered by a linearly recombining signal mixtures. The precise nature of this recombination is determined by a set of *unmixing coefficients* $(\alpha, \beta, \gamma, \delta)$, such that $s_1 = \alpha x_1 + \beta x_2$ and $s_2 = \gamma x_1 + \delta x_2$. Thus, the problem solved by ICA, and by all other BSS methods, consists of finding values for these unmixing coefficients.

The Mixing and Unmixing Matrices

The set of mixtures defines a vector variable $\mathbf{x} = (x_1, x_2)^T$, and the transformation from $\mathbf{s}$ to $\mathbf{x}$ defines a *mixing matrix* $\mathbf{A}$:

$$\begin{aligned} \mathbf{x} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} s_1^1 & s_1^2 & \dots & s_1^N \\ s_2^1 & s_2^2 & \dots & s_2^N \end{pmatrix} \\ &= \mathbf{A}\mathbf{s}. \end{aligned} \quad (3)$$

The mapping from $\mathbf{x}$ to $\mathbf{s} = (s_1, s_2)^T$ defines an optimal *unmixing matrix* $\mathbf{W}^* = (\mathbf{w}_1, \mathbf{w}_2)^T$ with (row) weight vectors $\mathbf{w}_1^T = (\alpha, \beta)$ and $\mathbf{w}_2^T = (\gamma, \delta)$

$$\begin{aligned} \mathbf{s} &= \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^N \\ x_2^1 & x_2^2 & \dots & x_2^N \end{pmatrix} \\ &= (\mathbf{w}_1, \mathbf{w}_2)^T (x_1, x_2) \end{aligned} \quad (4)$$

$$= \mathbf{W}^* \mathbf{x}. \quad (5)$$

It can be seen that $\mathbf{W}^*$ reverses, or inverts, the effects of $\mathbf{A}$, and indeed, $\mathbf{W}^*$ could be estimated from the *matrix inverse* $\mathbf{W}^* = \mathbf{A}^{-1}$, if $\mathbf{A}$ were known³. However, as we are ultimately concerned with finding $\mathbf{W}^*$ when $\mathbf{A}$ is not known, we cannot, therefore, use $\mathbf{A}^{-1}$ to estimate $\mathbf{W}^*$. For arbitrary values of the unmixing coefficients, the unmixing matrix is suboptimal and is denoted $\mathbf{W}$. In this case, the signals extracted by $\mathbf{W}$ are not necessarily source signals, and are denoted $\mathbf{y} = \mathbf{W}\mathbf{x}$.

Maximum Likelihood ICA

In practice, it is extremely difficult to measure the independence of a set of extracted signals unless we

4 Independent Component Analysis

have some general knowledge about those signals. In fact, the observations above suggest that we do often have some knowledge of the source signals. Specifically, we know that they are non-Gaussian, and that they are independent. This knowledge can be specified in terms of a formal model, and we can then extract signals that conform to this model. More specifically, we can search for an unmixing matrix that maximizes the agreement between the model and the signals extracted by that unmixing matrix.

One common interpretation of ICA is as a *maximum likelihood* (ML) method for estimating the optimal unmixing matrix $\mathbf{W}^*$. **Maximum likelihood estimation (MLE)** is a standard statistical tool for finding parameter values (e.g., the unmixing matrix $\mathbf{W}$) that provide the best fit of some data (e.g., the signals $\mathbf{y}$ extracted by $\mathbf{W}$) to a given a model. The ICA ML ‘model’ includes the adjustable parameters in $\mathbf{W}$, and a (usually fixed) model of the source signals. However, this source signal model is quite vague because it is specified only in terms of the general shape of the histogram of source signals. The fact that the model is vague means that we do not have to know very much about the source signals.

As noted above, mixtures of source signals are almost always Gaussian (see Figure 2), and it is fairly safe to assume that non-Gaussian signals must, therefore, be source signals. The amount of ‘Gaussian-ness’ of a signal can be specified in terms of its histogram, which is an approximation to a *probability density function* (pdf) (see Figure 2). A pdf $p_s(s)$ is essentially a histogram in which bin widths Δs are extremely small. The value of the function $p_s(s')$ is the *probability density* of the signal s at the value s' , which is the probability that the signal s lies within a small range around the value⁴ s' . As a pure speech signal contains a high proportion of silence, its pdf is highly ‘peaky’ or *leptokurtotic*, with a peak around zero (see Figure 3). It, therefore, makes sense to specify a leptokurtotic function (see **Kurtosis**) as our model pdf for speech source signals.

As we know the source signals are independent, we need to incorporate this knowledge into our model. The degree of mutual independence between signals can be specified in terms of their *joint pdf* (see Figure 3). By analogy, with the pdf of a scalar signal, a joint pdf defines the probability that the values of a set of signals $\mathbf{s} = (s_1, s_2)^T$ fall within a small range around a specific set of values $\mathbf{s}' = (s'_1, s'_2)^T$. Crucially, if these signals are mutually independent,

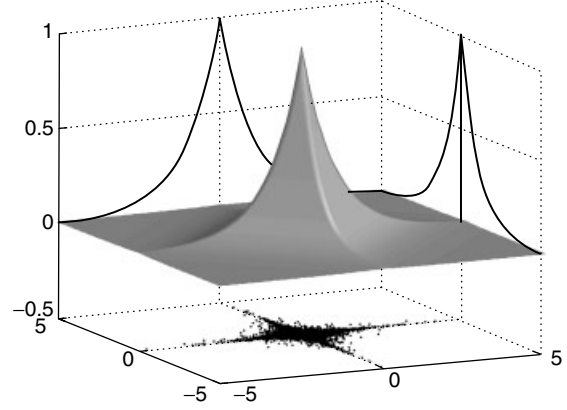


Figure 3 Marginal and joint probability density function (pdfs) of two high-kurtosis independent variables (e.g., speech signals). Given a set of signals $\mathbf{s} = (s_1, s_2)^T$, the pdf of each signal is essentially a histogram of values in that signal, as indicated by the two curves plotted along the horizontal axes. Similarly, the joint pdf p_s of two signals is essentially a two-dimensional histogram of pairs of signal values $\mathbf{s}' = (s'_1, s'_2)$ at time t . Accordingly, the joint probability of observing values $\mathbf{s}' = (s'_1, s'_2)$ is indicated by the local density of plotted points on the horizontal plane. This local density is an approximation to the joint pdf p_s , which is indicated by the height of the solid surface. The pdfs p_{s_1} and p_{s_2} of the signals s_1 and s_2 are the *marginal* pdfs of the joint pdf p_s .

then the joint pdf p_s of $\mathbf{s}$ can be expressed as the product of the pdfs (p_{s_1}, p_{s_2}) of its constituent signals s_1 and s_2 . That is, $p_s = p_{s_1} \times p_{s_2}$, where the pdfs p_{s_1} and p_{s_2} of the signals s_1 and s_2 (respectively) are known as the *marginal* pdfs of the joint pdf p_s .

Using ML ICA, the objective is to find an unmixing matrix $\mathbf{W}$ that yields extracted signals $\mathbf{y} = \mathbf{W}\mathbf{x}$, which have a joint pdf as similar as possible to the model joint pdf p_s of the unknown source signals $\mathbf{s}$. This model incorporates the assumptions that source signals are non-Gaussian (leptokurtotic, in the case of speech) and independent. Fortunately, ICA seems to be very robust with respect to differences between model pdfs and the pdfs of source signals [3]. Note that, as $\mathbf{A}$ and $\mathbf{W}$ are inverses of each other⁵, it does not matter whether the model parameters are expressed in terms of $\mathbf{A}$ or $\mathbf{W}$.

Somewhat perversely, we can consider the probability of obtaining the observed mixtures $\mathbf{x}$ in the context of such a model, where this probability is known as the *likelihood* of the mixtures. We can then pose the question: given that the source signals have

a joint pdf p_s , which particular mixing matrix $\mathbf{A}$ (and, therefore, which unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$) is most likely to have generated the observed signal mixtures $\mathbf{x}$? In other words, if the likelihood of obtaining the observed mixtures (from some unknown source signals with joint pdf p_s) were to vary with $\mathbf{A}$, then which particular $\mathbf{A}$ would maximize this likelihood?

MLE is based on the assumption that if the model joint pdf p_s and the model parameters $\mathbf{A}$ are correct, then a high probability (i.e., likelihood) should be obtained for the mixtures $\mathbf{x}$ that were actually observed. Conversely, if $\mathbf{A}$ is far from the correct parameter values, then a low probability of the observed mixtures would be expected. We will assume that all source signals have the same (leptokurtotic) pdf p_s . This may not seem much to go on, but it turns out to be perfectly adequate for extracting source signals from signal mixtures.

The Nuts and Bolts of ML ICA

Consider a (mixture) vector variable $\mathbf{x}$ with joint pdf p_x , and a (source) vector variable $\mathbf{s}$ with joint pdf p_s , such that $\mathbf{s} = \mathbf{W}^* \mathbf{x}$, where $\mathbf{W}^*$ is the optimal unmixing matrix. As noted above, the number of source signals and mixtures must be equal, which ensures that $\mathbf{W}^*$ is square. In general, the relation between the joint pdfs of $\mathbf{x}$ and $\mathbf{s}$ is

$$p_x(\mathbf{x}) = p_s(\mathbf{s}) |\mathbf{W}^*|, \quad (6)$$

where $|\mathbf{W}^*| = |\partial \mathbf{s} / \partial \mathbf{x}|$ is the *Jacobian* of $\mathbf{s}$ with respect to $\mathbf{x}$. Equation (6) defines the likelihood of the observed mixtures $\mathbf{x}$, which is the probability of $\mathbf{x}$ given $\mathbf{W}^*$.

For any nonoptimal unmixing matrix $\mathbf{W}$, the extracted signals are $\mathbf{y} = \mathbf{W} \mathbf{x}$. Making the dependence on $\mathbf{W}$ explicit, the likelihood $p_x(\mathbf{x}|\mathbf{W})$ of the signal mixtures $\mathbf{x}$ given $\mathbf{W}$ is

$$p_x(\mathbf{x}|\mathbf{W}) = p_s(\mathbf{W} \mathbf{x}) |\mathbf{W}|. \quad (7)$$

We would naturally expect $p_x(\mathbf{x}|\mathbf{W})$ to be maximal if $\mathbf{W} = \mathbf{W}^*$. Thus, (7) can be used to evaluate the quality of any putative unmixing matrix $\mathbf{W}$ in order to find that particular $\mathbf{W}$ that maximizes $p_x(\mathbf{x}|\mathbf{W})$. By convention, (7) defines a *likelihood*

function $L(\mathbf{W})$ of $\mathbf{W}$, and its logarithm defines the *log likelihood function* $\ln L(\mathbf{W})$. If the M source signals are mutually independent, so that the joint pdf p_s is the product of its M marginal pdfs, then (7) can be written

$$\ln L(\mathbf{W}) = \sum_i^M \sum_t^N \ln p_s(\mathbf{w}_i^T \mathbf{x}^t) + N \ln |\mathbf{W}|. \quad (8)$$

Note that the likelihood $L(\mathbf{W})$ is the joint pdf $p_x(\mathbf{x}|\mathbf{W})$ for $\mathbf{x}$, but using MLE, it is treated as if it were a function of the parameter $\mathbf{W}$. If we substitute a commonly used leptokurtotic model joint pdf for the source signals $p_s(\mathbf{y}) = (1 - \tanh(\mathbf{y}^2))$, then

$$\ln L(\mathbf{W}) = \sum_i^M \sum_t^N \ln(1 - \tanh(\mathbf{w}_i^T \mathbf{x}^t)^2) + N \ln |\mathbf{W}|. \quad (9)$$

The matrix $\mathbf{W}$ that maximizes this function is the *maximum likelihood estimate* of the optimal unmixing matrix $\mathbf{W}^*$. Equation (9) provides a measure of similarity between the joint pdf of the extracted signals $\mathbf{y} = \mathbf{W} \mathbf{x}$ and the joint model pdf of the source signals $\mathbf{s}$. Having such a measure permits us to use standard optimization methods to iteratively update the unmixing matrix in order to maximize this measure of similarity.

ICA, Principal Component Analysis and Factor Analysis

ICA is related to conventional methods for analyzing large data sets such as **principal component analysis (PCA)** and **factor analysis (FA)**. Whereas ICA finds a set of source signals that are mutually independent, PCA and FA find a set of signals that are mutually decorrelated (consequently, neither PCA nor FA could extract speech signals, for example). The ‘forward’ assumption that signals from different physical processes are uncorrelated still holds, but the ‘reverse’ assumption that uncorrelated signals are from different physical processes does not. This is because lack of correlation is a weaker property than independence. In summary, independence implies a lack of correlation, but a lack of correlation does not imply independence.

6 Independent Component Analysis

Notes

1. We use the term signal and variable interchangeably here.
2. This is a seminal paper, which initiated the recent interest in ICA.
3. The matrix inverse is analogous to the more familiar inverse for scalar variables, such as $x^{-1} = 1/x$.
4. For brevity, we will abuse this technically correct, but lengthy, definition by stating that $p_s(s')$ is simply the probability that s adopts the value s' .
5. Up to an irrelevant permutation of rows.

References

- [1] Bell, A.J. & Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**, 1129–1159.
- [2] Bell, A.J. & Sejnowski, T.J. (1997). The independent components of natural scenes are edge filters, *Vision Research* **37**(23), 3327–3338.
- [3] Cardoso, J. (2000). On the stability of source separation algorithms, *Journal of VLSI Signal Processing Systems* **26**(1/2), 7–14.
- [4] Hyvärinen, A. (2001). Complexity pursuit: separating interesting components from time series, *Neural Computation* **13**, 883–898.
- [5] Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, John Wiley & Sons, New York.
- [6] Makeig, S., Jung, T., Bell, A.J., Ghahremani, D. & Sejnowski, T.J. (1997). Blind separation of auditory event-related brain responses into independent components, *Proceedings National Academy of Sciences of the United States of America* **94**, 10979–10984.
- [7] McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S. & Sejnowski, T.J. (1998). Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task, *Proceedings National Academy of Sciences of the United States of America* **95**, 803–810.
- [8] Stone, J.V. (2001). Blind source separation using temporal predictability, *Neural Computation* **13**(7), 1559–1574.
- [9] Stone, J.V. (2004). *Independent Component Analysis: A Tutorial Introduction*, MIT Press, Boston.
- [10] van Hateren, J.H. & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex, *Proceedings of the Royal Society of London. Series B. Biological Sciences* **265**(7), 359–366.

JAMES V. STONE

Independent Pathway Model

FRÜHLING RIJSDIJK

Volume 2, pp. 913–914

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Independent Pathway Model

The independent pathway model, as applied to genetically sensitive data, is a multivariate model in which the relationship between a group of variables is controlled by genetic and environmental common **latent factors** [3]. The common factors each have their own free paths to the observed variables and account for the between trait covariance. In addition, a set of specific genetic and environmental factors are specified accounting for variance that is not shared with the other variables in the model (residual or unique variance).

For twin data, two identical pathway models are modeled for each twin's set of variables with the genetic and environmental factors across twins (both common and specific) connected by the expected correlations. For the genetic factors, these correlations are unity for MZ twins and 0.5 for DZ twins.

For twins reared together, correlations for the shared (family) environmental effects are 1. Unshared environmental factors are uncorrelated cross twins (see Figure 1). For the specific factors to all have free loadings, the minimal number of variables in this model is three. For example, for two variables, the independent pathway model would have specific factors which are constrained to be equal, which is then equivalent to a **Cholesky decomposition**.

Similar to the *univariate genetic model*, the MZ and DZ ratio of the cross-twin within variable correlations (e.g., Twin 1 variable 1 and Twin 2 variable 1) will indicate the relative importance of genetic and environmental variance components for each variable. On the other hand, the MZ and DZ ratio of the cross-twin cross-trait correlations (e.g., Twin 1 variable 1 and Twin 2 variable 2) will determine the relative importance of genetic and environmental factors in the covariance between variables (i.e., genetic and environmental correlations). In addition, for any two variables it is possible to derive part of the phenotypic correlation determined by common genes

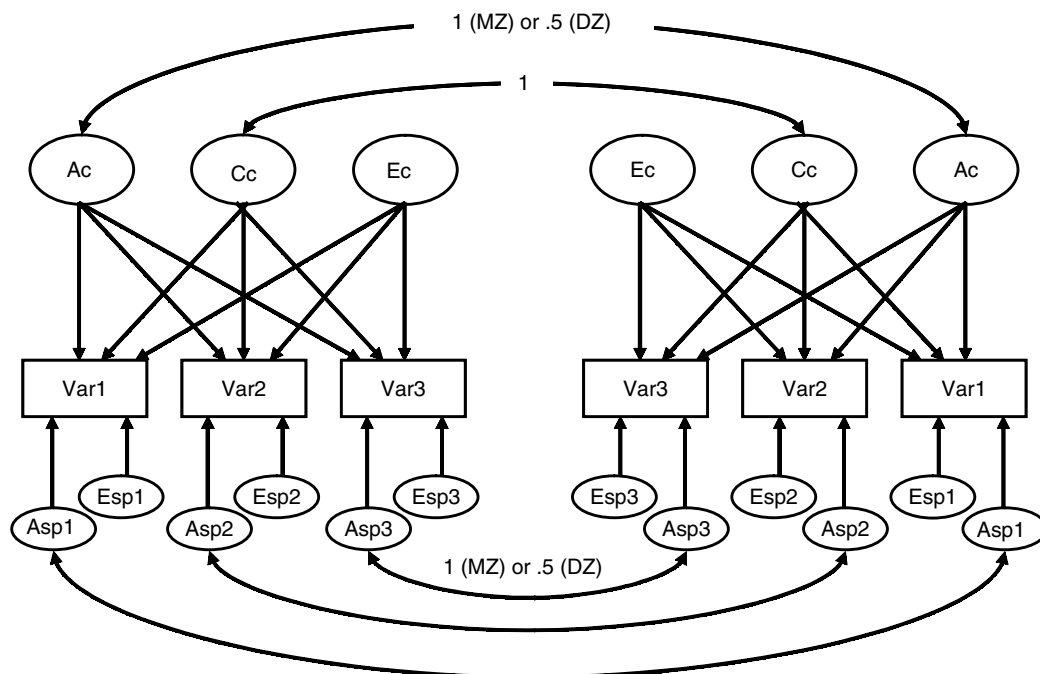


Figure 1 Independent Pathway Model for a Twin Pair: Ac, Cc, and Ec are the common additive, genetic, common shared, and common nonshared environmental factors, respectively. The factors at the bottom are estimating the variable specific A and E influences. For simplicity the specific C factors were omitted from the diagram

2 Independent Pathway Model

(which will be a function of both their h^2 (see **Heritability**) and genetic correlation) and by common shared and unique environmental effects (which will be a function of their c^2 and e^2 and the C and E correlation [see **ACE Model**]). For more information on genetic and environmental correlations between variables, see the general section on **multivariate genetic analysis**. Parameter estimates are estimated from the observed variances and covariances by **structural equation modeling**.

So what is the meaning and interpretation of this factor model? An obvious one is to examine the etiology of **comorbidity**. One of the first applications in this sense was by Kendler et al. [1], who illustrated that the tendency of self-report symptoms of anxiety and depression to form separate symptom clusters was mainly due to shared genetic influences. This means that genes act largely in a nonspecific way to influence the overall level of psychiatric symptoms. The effects of the environment were mainly specific. The conclusion was that separable anxiety and depression symptom clusters in the general population are largely the result of environmental factors.

Independent pathway models with more than one set of common genetic and environmental factors are also possible. Multivariate genetic analyses of

schizotypal personality data indicated that at least two latent factor (see **Latent Variable**) structures are required to account for the genetic covariation between the various components of schizotypy. The positive components (reality distortion, such as magical ideas, unusual perceptions, and referential ideas) and negative components (anhedonia, social isolation, and restricted affect) are relatively genetically independent, although each in turn may be related to cognitive disorganization [2].

References

- [1] Kendler, K.S., Heath, A.C., Martin, N.G. & Eaves, L.J. (1987). Symptoms of anxiety and symptoms of depression: same genes, different environment? *Archives of General Psychiatry* **44**, 451–457.
- [2] Linney, Y.M., Murray, R.M., Peters, E.R., Macdonald, A.M., Rijdsdijk, F.V. & Sham, P.C. (2003). A quantitative genetic analysis of Schizotypal personality traits, *Psychological Medicine* **33**, 803–816.
- [3] McArdle, J.J. & Goldsmith, H.H. (1990). Alternative common-factor models for multivariate biometrical analyses, *Behavior Genetics* **20**, 569–608.

FRÜHLING RIJSDIJK

Index Plots

PAT LOVIE

Volume 2, pp. 914–915

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Index Plots

An index plot is a **scatterplot** of data plotted serially against the observation (case) number within the sample. The data could consist of original observations or some derived measure, such as **residuals** or predicted values.

An index plot is a useful exploratory tool for two different situations and purposes. If the data are in serial order, for instance, because they were collected over time or systematically over a demographic area, the index plot can be an effective way of detecting patterns or trends; this version is sometimes known as a sequence plot. Furthermore, an index plot yields information about anomalous values (potential **outliers**) irrespective of whether the cases are in arbitrary or serial order.

As a simple illustration, consider the following 10 observations which are the times (in milliseconds) between reported reversals of orientation of a Necker cube obtained from a study on visual illusions: 302, 274, 334, 430, 489, 1697, 703, 978, 2745, 1656. Suppose that these represent 10 consecutive inter-reversal times for one person in a single experimental session. From Figure 1 we can see that times between reversals tend to increase the longer the person

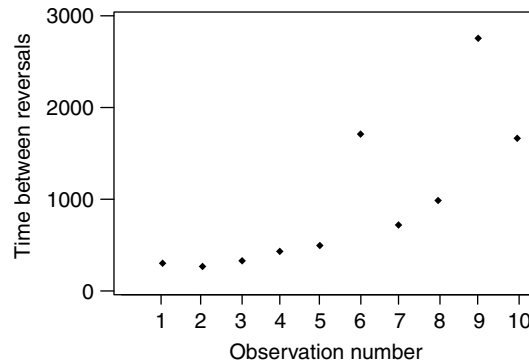


Figure 1 Index plot of times (in msecs) between reported reversals of a Necker cube

viewed the cube, and also that cases 6 and 9 are unusually large and so are possibly outliers. Interestingly, observation 10, though almost the same time as observation 6, does not seem out of line with the general trend. On the other hand, if each of these 10 observations had represented an inter-reversal time for a different person, then we would certainly have reported all three cases as anomalous.

PAT LOVIE

Industrial/Organizational Psychology

RONALD S. LANDIS AND SETH A. KAPLAN

Volume 2, pp. 915–920

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Industrial/Organizational Psychology

The field of industrial and organizational (IO) psychology has a long, rich statistical tradition. At the beginning of the twentieth century, experimental psychologists Walter Dill Scott and Hugo Münsterberg began applying psychological principles to the study of organizations, giving birth to the field of IO [2]. Around the same time, Frederick Taylor, an engineer, proposed principles of Scientific Management, which were designed to guide the selection, training, and reward of production workers [2]. From the time of Münsterberg, Scott, and Taylor, quantitative methodologies and statistical techniques have played a central role in the IO field of inquiry. This tradition continues as IO researchers apply sophisticated statistical procedures such as quantitative **meta-analysis**, **structural equation modeling**, and multilevel techniques (see **Linear Multilevel Models**). In addition, IO scholars contribute meaningfully to the improvement of existing techniques and the development of novel methodologies.

This entry, which is divided into four sections, provides a summary and description of statistical methods used in IO psychology. The first section lists the statistical techniques reported in two leading journals in the field. The second section summarizes the reasons that IO psychologists use several popular techniques. The third section illustrates contributions made by IO psychologists to the statistical analysis literature and the fourth notes the importance of statistics in IO theory development.

Statistical Methods Appearing in Two Leading IO Journals

Several reviews [1, 3, 4] of the research methods and statistical techniques used in IO and related fields appeared in the past decade or so. The purpose of this entry is to focus on statistical techniques rather than broader research design elements.

In a recent chapter, Austin et al. [1] reviewed the use of various research methods in IO for the 80-year period from 1920 to 2000 using articles from every tenth volume of *Journal of Applied Psychology* (*JAP*), which were coded for measurement, design,

and analytic issues. Our methodology was similar, but with three important differences. First, we used studies in two IO journals, *JAP* and *Personnel Psychology* (*PP*), in order to determine whether differences in statistical methods exist between publications. Second, we coded all studies in both journals for the nine-year period from 1995 to 2003 rather than sampling representative articles. Finally, we coded only statistical methods and not measurement or design issues. Our intention was to provide a thorough picture of the specific statistical methods currently used in the IO literature.

Results for *JAP* and *PP* are presented in Tables 1 and 2, respectively. These tables display the yearly percent and nine-year average for 10 techniques from 1995 to 2003. Because studies often used several statistical analyses, the percentages in each year sum to more than 100%. These data reveal that basic correlational statistics (see **Correlation Studies**), **multiple linear regression**, and **analysis of variance** (ANOVA)-based methods appear with the greatest frequency, while more sophisticated techniques such as meta-analysis, structural equation modeling, and hierarchical linear modeling (see **Hierarchical Models**) appear less often.

Two other notable features are prominent in Tables 1 and 2. First, there is consistency in the recent use of the techniques across time. Despite the fact that absolute percentage use varies across the time for particular methods (e.g., in *JAP* the percent use of multiple regression varied between 31 and 62%), the relative use appears quite stable over time. If techniques are treated as cases, the correlation between percentage use in *JAP* between 1995 and 1996 is quite large ($r = 0.96$). In fact, the average correlation of adjacent years was quite high (*JAP*, $\bar{r} = 0.97$, *PP*, $\bar{r} = 0.93$, Overall $\bar{r} = 0.95$).

Second, there has also been remarkable consistency between the journals. For example, if techniques are treated as cases, the correlation between percentage use in *JAP* and *PP* in 1995 was quite high ($r = 0.94$) with a similarly large average correlation ($\bar{r} = 0.94$) across journal within year. In short, the relative use of statistical techniques is similar across the two journals. One noticeable absolute difference, however, is exploratory factor analysis (see **Factor Analysis: Exploratory**). On average, 14% of studies that appeared in *JAP* reported exploratory factor analytic results compared with 24% of studies in *PP*. Given that this difference is not extremely

2 Industrial/Organizational Psychology

Table 1 Percentage of studies in Journal of Applied Psychology that used 10 statistical techniques – 1995–2003

Type of analysis	1995	1996	1997	1998	1999	2000	2001	2002	2003	Ave	SD
<i>ANOVA</i>	53	58	41	50	41	35	59	61	47	49.44	9.14
<i>CFA</i>	14	14	16	15	11	24	29	16	27	18.44	6.46
<i>CHISQ</i>	4	11	11	5	8	6	8	10	4	7.44	2.83
<i>CORR</i>	76	72	66	78	72	78	86	84	78	76.67	6.16
<i>EFA</i>	8	13	8	5	15	9	30	12	23	13.67	8.06
<i>HLM</i>	0	1	0	0	1	2	4	5	8	2.33	2.78
<i>LOGR</i>	5	3	3	4	1	5	3	4	0	3.11	1.69
<i>MA</i>	7	10	8	11	9	5	7	11	1	7.67	3.20
<i>MR</i>	46	28	32	45	40	38	57	48	43	41.89	8.68
<i>SEM</i>	7	11	11	9	17	8	17	14	13	11.89	3.66

Notes: Values rounded to the nearest whole number. Studies associated with nontraditional IO research topics such as eyewitness testimony, jury selection, and suspect lineup studies were not included. *ANOVA* = *t* Tests, analysis of variance, analysis of covariance, multivariate analysis of variance, and multivariate analysis of covariance; *CFA* = confirmatory factor analysis; *CHISQ* = chi-square; *CORR* = bivariate correlations; *EFA* = exploratory factor analysis; *HLM* = hierarchical linear modeling; *LOGR* = logistic regression; *MA* = meta-analysis; *MR* = multiple regression analysis, and *SEM* = structural equation modeling.

Table 2 Percentage of studies in Personnel Psychology that used 10 statistical techniques – 1995–2003

Type of analysis	1995	1996	1997	1998	1999	2000	2001	2002	2003	Ave	SD
<i>ANOVA</i>	56	36	59	54	36	34	46	62	40	47.00	10.95
<i>CFA</i>	0	7	19	14	32	7	15	8	27	14.33	10.30
<i>CHISQ</i>	9	11	7	8	0	5	13	8	7	7.56	3.68
<i>CORR</i>	66	71	72	86	82	66	73	77	80	74.78	6.98
<i>EFA</i>	25	29	25	29	18	24	15	15	33	23.67	6.42
<i>HLM</i>	0	4	0	0	0	0	4	0	0	0.89	1.76
<i>LOGR</i>	4	0	7	4	9	5	4	4	0	4.11	2.89
<i>MA</i>	13	7	9	0	14	7	4	0	27	9.00	8.37
<i>MR</i>	41	39	41	57	54	52	46	38	67	48.33	9.85
<i>SEM</i>	9	7	6	11	7	3	15	4	7	7.67	3.64

Notes: Values rounded to the nearest whole number. Studies associated with nontraditional IO research topics such as eyewitness testimony, jury selection, and suspect lineup studies were not included. *ANOVA* = *t* Tests, analysis of variance, analysis of covariance, multivariate analysis of variance, and multivariate analysis of covariance; *CFA* = confirmatory factor analysis; *CHISQ* = chi-square; *CORR* = bivariate correlations; *EFA* = exploratory factor analysis; *HLM* = hierarchical linear modeling; *LOGR* = logistic regression; *MA* = meta-analysis; *MR* = multiple regression analysis, and *SEM* = structural equation modeling.

pronounced and that no other techniques show similar differences, we can conclude that technique use does not vary as a function of publication outlet in the two leading journals in the IO field.

Owing to the consistency of the observed results, data were collapsed across year and journal to produce an overall ranking of statistics use (see Table 3). Rankings are based on the percentage of all studies that report each statistical technique.

Table 3 makes clear the use of statistics in IO psychology. Correlations are the statistics most frequently used, appearing in approximately 78% of empirical research in *JAP* and *PP* during

the nine years. In a second tier, ANOVA-based statistics and multiple regression appeared in 49% and 43% of studies, respectively. Confirmatory factor analysis (see **Factor Analysis: Confirmatory**) (18%), exploratory factor analysis (17%), and **structural equation modeling** (11%) comprised a third tier. Meta-analysis (9%), chi-square analysis (see **Contingency Tables**) (8%), **logistic regression** (3%), and hierarchical linear modeling (2%) appeared in the fewest coded studies.

These techniques do not exhaust the statistical ‘toolbox’ used by IO psychologists. For example, statistical techniques such as **canonical correlation**

Table 3 ‘Top ten’ statistical techniques in IO psychology

Rank	Type of analysis	Overall percent
1	Correlation	78
2	Analysis of Variance	49
3	Multiple Regression	43
4	Confirmatory Factor Analysis	18
5	Exploratory Factor Analysis	17
6	Structural Equation Modeling	11
7	Meta-Analysis	9
8	Chi-square	8
9	Logistic Regression	3
10	Hierarchical Linear Modeling	2

and **discriminant function analysis** also appear in IO-related articles. Table 3 simply provides a ‘Top 10’ of the most widely used techniques.

Information not included in the tabled data, but especially striking, is the percentage of articles in which the primary focus was some aspect of statistics or research methodology. In *JAP*, 4.8% of the articles fell into this category, whereas 15.6% of the articles in *PP* addressed a primarily psychometric, methodological, or statistical issue. Although the nature of these articles varied widely, many reported results of statistical simulations, especially simulations of the consequences of certain statistical considerations (e.g., violations of assumptions). Others detailed the development or refinement of a given technique or analytic procedure.

Types of Analyses used in IO Research

Although the distinction between industrial and organizational psychology may be perceived as somewhat nebulous, treating the two areas as unique is a useful heuristic. Industrial psychologists traditionally study issues related to worker performance, productivity, motivation, and efficiency. To understand and predict these criterion variables, researchers explore how concepts such as individual differences (e.g., cognitive ability), workplace interventions (e.g., training programs), and methods of employee selection (e.g., job interviews) impact job-related outcomes. The statistical orientation of industrial psychology is largely a function of both the need to quantify abstract psychological constructs (e.g., cognitive ability, personality traits, job performance) and the practical difficulties faced in organizational settings (e.g., lack

of random assignment, measuring change due to a specific intervention).

Organizational psychologists generally study a broader range of topics including job attitudes, worker well-being, motivation, careers, and leadership. Addressing such issues presents a number of special statistical considerations. Measurement issues and practical considerations of data collection similarly confront organizational researchers. In addition, the hierarchical nature of organizations (i.e., individuals nested within groups nested within companies) presents unique methodological challenges for organizational psychologists. To address such factors, researchers often employ a variety of methodological and statistical techniques in order to draw strong conclusions.

Bivariate Correlation

As Table 3 illustrates, simple correlational analyses appear with the most frequency in IO research. For present purposes, correlational analyses include those that involve **Pearson correlations**, phi coefficients, biserial correlations, **point-biserial correlations**, and **terachoric correlations** methods that assess relationships between two variables. Simple correlations are reported in studies related to nearly every subarea (e.g., selection, training, job performance, work attitudes, organizational climate, and motivation).

One factor responsible for the ubiquity of correlations in IO research is the field’s interest in reliability information. Whether discussing predictor tests, criterion ratings, attitude scales, or various self-report measures, IO psychologists are typically concerned about the consistency of the observed data. Types of reliability reported include test-retest, alternate forms, internal consistency (most frequently operationalized through coefficient alpha), and interrater reliability.

Analysis of Variance

Statistics such as *t* Tests, analysis of variance, **analysis of covariance** (ANCOVA), **multivariate analysis of variance** (MANOVA), and multivariate analysis of covariance (MANCOVA) appear in studies that involve comparisons of known or manipulated groups. In addition, researchers often utilize *t* Tests and ANOVA prior to conducting other analyses to ensure that different groups do not differ significantly from one another on the

primary variables of interest. ANCOVA allows one to statistically control for confounding variables, or covariates, that potentially distort results and conclusions. Because organizational reality often precludes true experimentation, ANCOVA is quite popular among IO psychologists, especially those in organizations. MANOVA and MANCOVA are useful when dealing with either multiple criteria and/or repeated measurements (*see* **Repeated Measures Analysis of Variance**). For example, evaluating separate training outcomes or evaluating one or more outcomes repeatedly would warrant one of these techniques.

Multiple Regression

Multiple regression (MR) analysis is used in three situations: (1) identifying the combination of predictors that can most accurately forecast a criterion variable, (2) testing for mediated relationships (*see* **Mediation**), and (3) testing for the presence of statistical interactions (*see* **Interaction Effects**). The first situation occurs, for example, when IO psychologists attempt to identify the optimal set of predictor variables that an organization should utilize in selecting employees. Using MR in this manner yields potential practical and financial benefits to organizations by enabling them to eliminate useless or redundant selection tests while maintaining optimal prediction.

Utilizing MR for mediated relationships is increasingly common in IO and has led to both theoretical and practical advances. For example, researchers utilize MR when attempting to identify the intervening variables and processes that explain bivariate relationships between predictor variables (e.g., cognitive ability, personality traits) and relevant criteria (e.g., job performance). Moderated MR (*see* **Moderation**) is also encountered in IO, both to uncover complex relationships that main effects fail to capture and to identify important boundary conditions that limit the generalizability of conclusions. In addition, organizations use moderated MR to ensure that the empirical relationship between a given selection measure and the criterion is constant across subgroups and protected classes. Any evidence to the contrary, revealed by a significant interaction between predictor and group, necessitates that the organization abandon the procedure.

Confirmatory & Exploratory Factor Analysis

Exploratory factor analysis is used by IO psychologists to provide construct validity evidence in many substantive interest areas. In particular, exploratory factor analysis is used in situations that involve newly created or revised measures. Often, but not always, those using exploratory factor analysis for this purpose hope to find that all of the items load on a single factor, suggesting that the measure is unidimensional.

Confirmatory factor analysis has become increasingly popular in recent years, largely due to the increasing availability of computer packages such as LISREL and EQS (*see* **Structural Equation Modeling: Software**). Unlike exploratory techniques, confirmatory approaches allow one to specify an *a priori* factor structure, indicating which items are expected to load on which factors. Confirmatory factor analysis is also useful for investigating the presence of method variance, often through **multi-trait-multimethod** data, as well as ensuring that factor structures are similar, or invariant, across different subgroups.

Structural Equation Modeling

Because **path analysis** using ordinary least squares regression does not allow for the inclusion of measurement error, structural equation modeling is used to test hypothesized measurement and structural relationships between variables. Although the use of structural equation modeling in IO remains relatively infrequent (*see* Tables 1 and 2), this approach holds great promise, especially given the increasing sophistication of IO theories and models. As IO psychologists become more familiar with structural equation modeling and the associated software, its frequency should increase.

Meta-analysis

The use of meta-analytic techniques has led to several ‘truths’ in IO psychology. From the selection literature, meta-analytic results reveal that the best predictor of job performance across all jobs is cognitive ability and the best personality-related predictor is conscientiousness. More generally, meta-analysis led to the insight that disparities in results between studies are due largely to artifacts inherent in the measurement process. This conclusion has the potential

to change the ways that IO psychologists undertake applied and academic research questions. In addition to reducing the necessity of conducting local validation studies in organizations, academic researchers may choose meta-analytic methodologies rather than individual studies.

Logistic Regression

Logistic regression is useful for predicting dichotomous outcomes, especially when fundamental assumptions underlying linear regression are violated. This technique is especially common among IO psychologists examining issues related to employee turnover, workplace health and safety, and critical performance behaviors because the relevant criteria are dichotomous. For example, a researcher might use logistic regression to investigate the variables that are predictive of whether one is involved in a driving accident or whether a work team performs a critical behavior.

Contributions of IO Psychologists to the Statistics Literature

In addition to extensively utilizing existing analytical techniques, IO psychologists also conduct research in which they examine, refine, and create statistical tools. Largely driving these endeavors is the complex nature of organizational phenomena that IO psychologists address. Often, however, these statistical advances not only enable researchers and practitioners to answer their questions but also propagate new insights and questions. In addition, other areas both within and outside of the organizational realm often benefit by applying IO psychologists' statistical advances. In the following paragraphs, we list several statistical topics to which IO researchers made especially novel and significant contributions. This listing is not exhaustive with respect to topics or results but is presented for illustrative purposes.

IO researchers have made contributions in quantitative meta-analysis, especially in terms of validity generalization. Until the late 1970s, IO psychologists believed that to identify those variables that best predicted job performance, practitioners must conduct a validation study for each job within each organization. This notion, however, was radically altered by demonstrating that predictor-criterion validity often

'generalizes' across organizations, thereby suggesting that local validation studies are not always essential. Specifically, early meta-analytic work revealed that validity estimates from different organizations and situations often differed from each other primarily as a function of statistical artifacts inherent in the measurement process (e.g., sampling error, low reliability, range restriction) and not as a result of specific contextual factors. These insights led to several lines of statistical research on how best to conceptualize and correct for these artifacts, especially when the primary studies do not contain the necessary information (e.g., reliability estimates).

Beginning in the early 1980s, IO psychologists also made strides in examining and developing various aspects of structural equation modeling. Some of these advances were related to the operationalization of continuous moderators, procedures for evaluating the influence of method variance, techniques for assessing model invariance across groups, the use of formative versus reflective manifest variables, and the impact of item parceling on model fit. Notably, some developments engendered novel research questions that, prior to these advances, IO psychologists may not have considered. For example, recent developments in latent growth modeling allowed IO psychologists to study how individuals' changes over time on a given construct impact or are impacted by their standing on another variable. Thus, to study how changes in workers' job satisfaction influence their subsequent job performance, the researcher can now measure how intra individual *changes* in satisfaction affect performance, instead of relying on a design in which Time 1 satisfaction simply is correlated with Time 2 performance.

Yet another statistical area that IO psychologists contributed to is difference scores. Organizational researchers traditionally utilized difference scores to examine issues such as the degree of 'fit' between a person and a job or a person and an organization. Throughout the 1990s, however, a series of articles highlighted several problematic aspects of difference scores and advanced the use of an alternative technique, polynomial regression, to study questions of fit and congruence.

The preceding discussion covers only a few of IO psychology's contributions to statistical methods. Given the continuing advances in computer technology as well as the ever-increasing complexity of management and organizational theory, IO psychologists

probably will continue their research work on statistical issues in the future.

The Role of Statistics in Theory Development

This entry illustrates the interest that IO psychologists have in quantitative methods. Although IO psychologists pride themselves on the relative methodological rigor and statistical sophistication of the field, our focus on these issues is not without criticism. For example, IO psychologists may be viewed as overly concerned with psychometric and statistical issues at the expense of underlying constructs and theory. To be sure, this criticism may once have possessed some merit. Recently, however, IO psychologists have made significant theoretical advancements as evidenced by recent efforts to understand, instead of simply predict, job performance and other important criteria. Without our embrace of sophisticated statistical analyses, this theoretical focus might not have emerged. Procedures such as structural equation modeling, hierarchical linear modeling, and meta-analysis have enabled researchers to assess complex theoretical formulations, and have allowed practitioners to better serve organizations and workers. Moreover, many other areas of psychology often benefit from the statistical skills of IO psychologists, especially

in terms of the availability of new or refined techniques. Thus, IO psychologists continue embracing statistics both as an instrumental tool to address theoretical research questions and as an area of study and application worthy of addressing in its own right.

References

- [1] Austin, J.T., Scherbaum, C.A. & Mahlman, R.A. (2002). History of research methods in industrial and organizational psychology: measurement, design, and analysis, in *Handbook of Research Methods in Industrial and Organizational Psychology*, S. Rogelberg, ed., Blackwell Publishers, Malden.
- [2] Landy, F. (1997). Early influences on the development of industrial and organizational psychology, *Journal of Applied Psychology* **82**, 467–477.
- [3] Sackett, P.R. & Larson Jr, J.R. (1990). Research strategies and tactics in industrial and organizational psychology, in *Handbook of Industrial and Organizational Psychology*, M.D. Dunnette & L.M. Hough, eds, Vol. 1, 2nd Edition, Consulting Psychologists Press, Palo Alto.
- [4] Stone-Romero, E.F., Weaver, A.E. & Glenar, J.L. (1995). Trends in research design and data analytic strategies in organizational research, *Journal of Management* **21**, 141–157.

RONALD S. LANDIS AND SETH A. KAPLAN

Influential Observations

LAWRENCE PETTIT

Volume 2, pp. 920–923

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Influential Observations

Introduction

An influential observation is one that has a large effect on our inferences. To measure the influence of an observation, we often compare the inferences we would make with all the data to those made excluding one observation at a time.

One of the earliest influence diagnostics, which is still widely used, is due to Cook [4]. In the context of a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is the $n \times p$ design matrix, he defined the influence of the i th observation on the estimate of $\boldsymbol{\beta}$ as

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2}, \quad (1)$$

where $\hat{\boldsymbol{\beta}}$ is the usual least squares estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares estimate omitting the i th observation and s^2 is the residual mean square. This can also be written as

$$D_i = \frac{t_i^2}{p} \frac{v_i}{(1 - v_i)}, \quad (2)$$

where

$$t_i = \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})}{s}$$

is the standardized residual and $v_i = \mathbf{x}_i(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i^T$ is the leverage, a measure of how unusual the regressor values are. Thus, an influential observation can be a gross outlier (large $|t_i|$), have a very unusual set of regressor values (large v_i), or a combination of the two.

Consider Figure 1, an example of a simple linear regression (see **Multiple Linear Regression**). The observations labelled * on their own satisfy a simple linear regression model. Observation A is an outlier but has low leverage and is not very influential because $\hat{\boldsymbol{\beta}}$ will not be affected much by it. Observation B has a high leverage but has a small residual and is not influential. Observation C is influential, has high leverage, and a large residual.

At about the time that Cook's paper was published, a number of other case deletion diagnostics were proposed, which are similarly functions of t_i and v_i .

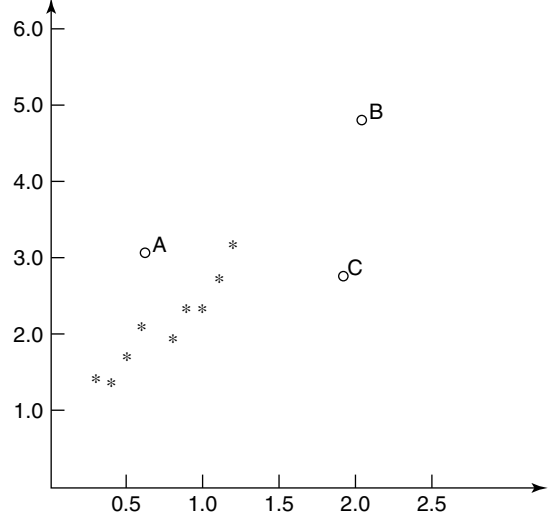


Figure 1 A simple linear regression with added outliers and influential observations

Writing $F_i = \frac{t_i^2(n - p - 1)}{n - p - t_i^2}$, we have

$$\begin{aligned} & \left(\frac{n - p}{p} F_i \right) \left(\frac{v_i}{1 - v_i} \right), \quad [17] \\ & \left(1 - \frac{t_i^2}{n - p} \right) (1 - v_i), \quad [1] \\ & \left(\frac{t_i^2}{p} \right) v_i, \quad \left(\frac{F_i}{p} \right) v_i, \quad [5] \end{aligned} \quad (3)$$

Others, for example Atkinson [2], proposed similar measures but using $s_{(i)}^2$. In general, the same observations are detected as influential by these different measures and many analysts would look at t_i , v_i as well as their preferred influence diagnostic.

Theoretical work on influence has often been based on the influence curve [9].

Suppose $X_1, X_2, \dots, X_n$ are a random sample of observations on X and a statistic of interest, T , can be written as a functional $T(F_n)$ of the empirical distribution function $\hat{F}$ of $X_1, X_2, \dots, X_n$. If X has cdf F , then the influence curve of T evaluated at $X = x$ is

$$IC_{T,F}(x) = \lim_{\varepsilon \rightarrow 0_+} \frac{T[(1 - \varepsilon)F + \varepsilon\delta_x] - T(F)}{\varepsilon} \quad (4)$$

2 Influential Observations

It gives a measure of the influence on T of adding an observation at x as $n \rightarrow \infty$.

Several finite sample versions of the influence curve have been suggested. The empirical influence curve (EIC) is obtained by substituting the sample cdf $\hat{F}$ for F in the influence curve. For linear models,

$$\begin{aligned} EIC(\mathbf{x}, y) &= n(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} (y - \mathbf{x}^T \hat{\boldsymbol{\beta}}) \\ EIC_i &= EIC(\mathbf{x}_i, y_i) = n(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i e_i, \end{aligned} \quad (5)$$

where e_i is the crude residual.

The sample influence curve (SIC) is obtained by taking $F = \hat{F}$ and $\varepsilon = -1/(n-1)$ in the definition of the influence curve. This leads to

$$SIC_i = -(n-1)(T(\hat{F}_{(i)}) - T(\hat{F})), \quad (6)$$

which in the case of the linear model gives

$$\begin{aligned} SIC_i &= (n-1)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) \\ &= \frac{(n-1)(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i e_i}{1 - v_i}. \end{aligned}$$

See [6] for more details and relationships between these measures.

A complicating factor in considering influence is that, in general, it is not ‘additive’. A pair of observations may not be individually influential but if both are deleted they may be jointly influential. With large data sets it is not clear what size sets of observations should be considered and computation may be a problem.

Bayesian case deletion diagnostics, looking at the effect on the posterior or predictive distribution, were considered by for example, Johnson and Geisser [10] and Pettit and Smith [15]. They used symmetric Kullback Leibler distance, for example,

$$I(i) = \int \log \frac{p(\boldsymbol{\theta}|\mathbf{y})}{p(\boldsymbol{\theta}|\mathbf{y}_{(i)})} [p(\boldsymbol{\theta}|\mathbf{y}) - p(\boldsymbol{\theta}|\mathbf{y}_{(i)})] d\boldsymbol{\theta}, \quad (7)$$

where $\boldsymbol{\theta}$ represents the parameters of interest. For vague prior information, the Bayesian diagnostic due to Pettit and Smith can be written as

$$I(i) = \frac{1}{2} \frac{v_i}{1 - v_i} (v_i + (2 - v_i)t_i^2). \quad (8)$$

Note that unlike most of the frequentist diagnostics it is not zero if t_i is. This is because $I(i)$ measures the effect on the whole posterior. Deleting an observation with $t_i = 0$ would not affect $\hat{\boldsymbol{\beta}}$ but may affect its variance if v_i is large.

The idea of using case deletion or a sample influence curve to measure influence has been extended to many situations, for example,

- **Principal component analysis** [7, 14]
- **Time series** [3]
- **Measures of skewness** [8]
- **Correspondence analysis** [13]
- **Cluster analysis** [11]

These may show rather different characteristics to the linear model case. For example, in principal component analysis, Pack et al. [14] show that influence is approximately additive. They also show that one influential observation, in their case caused by two measurements being swapped, can have a surprisingly large effect with the second principal component being due to this one observation.

Although influence of observations on a parameter estimate is of importance, another class of problems, model choice, has received less attention. Pettit and Young [16] and Young [18] discussed the influence of one or more observations on a Bayes factor. They defined the effect of observation d on a Bayes factor comparing models M_0 and M_1 as the difference in log Bayes factors based on all the data and omitting observation d ,

$$k_d = \log_{10} \left(\frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)} \right) - \log_{10} \left(\frac{p(\mathbf{y}_{(d)}|M_0)}{p(\mathbf{y}_{(d)}|M_1)} \right). \quad (9)$$

The diagnostic k_d can also be written as the difference in log conditional predictive ordinates (CPO) under the two models. CPO is an outlier measure. In general, an observation will have a large influence if it is an outlier under one model but not the other.

For example, when testing a mean, consider a normal sample with one observation contaminated by adding δ . Typical behavior of $|k_d|$ is to slowly increase to a maximum as δ increases and then to fall. For small δ , the contaminant is an outlier under M_0 but not M_1 . As δ increases, it becomes an outlier under both models and loses its influence.

Jolliffe and Lukudu [12] find similar behavior when looking at the effect of a contaminant on the T statistic.

The question remains as to what an analyst should do when they find an observation is influential. It should certainly be reported. Sometimes, as in the Pack et al. [14] example, it is a sign of a recording error that can be corrected. If a designed experiment results in an influential observation, it suggests that taking some more observations in that part of the design space would be a good idea. Another possibility is to use a more robust procedure that automatically down weights such observations. It may also suggest that a hypothesized model does not hold for the whole of the space of regressors.

References

- [1] Andrews, D.F. & Pregibon, D. (1978). Finding the outliers that matter, *Journal of the Royal Statistical Society. Series B* **40**, 85–93.
- [2] Atkinson, A.C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression, *Biometrika* **68**, 13–20.
- [3] Bruce, A.G. & Martin, R.D. (1989). Leave- k -out diagnostics for time series (with discussion), *Journal of the Royal Statistical Society. Series B* **57**, 363–424.
- [4] Cook, R.D. (1977). Detection of influential observations in linear regression, *Technometrics* **19**, 15–18.
- [5] Cook, R.D. & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometrics* **22**, 495–508.
- [6] Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, New York.
- [7] Critchley, F. (1985). Influence in principal components analysis, *Biometrika* **72**, 627–636.
- [8] Groeneveld, R.A. (1991). An influence function approach to describing the skewness of a distribution, *American Statistician* **45**, 97–102.
- [9] Hampel, F. (1968). *Contributions to the theory of robust estimation*, Unpublished PhD dissertation, University of California, Berkeley.
- [10] Johnson, W. & Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis, *Journal of the American Statistical Association* **78**, 137–144.
- [11] Jolliffe, I.T., Jones, B. & Morgan, B.J.T. (1995). Identifying influential observations in hierarchical cluster analysis, *Journal of Applied Statistics* **22**, 61–80.
- [12] Jolliffe, I.T. & Lukudu, S.G. (1993). The influence of a single observation on some standard statistical tests. *Journal of Applied Statistics*, **20**, 143–151.
- [13] Pack, P. & Jolliffe, I.T. (1992). Influence in correspondence analysis, *Applied Statistics* **41**, 365–380.
- [14] Pack, P., Jolliffe, I.T. & Morgan, B.J.T. (1988). Influential observations in principal component analysis: a case study, *Journal of Applied Statistics* **15**, 39–52.
- [15] Pettit, L.I. & Smith, A.F.M. (1985). Outliers and influential observations in linear models, in *Bayesian Statistics 2*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith eds, Elsevier Science Publishers B.V., North Holland, pp. 473–494.
- [16] Pettit, L.I. & Young, K.D.S. (1990). Measuring the effect of observations on Bayes factors, *Biometrika* **77**, 455–466.
- [17] Welsch, R.E. & Kuh, E. (1977). Linear regression diagnostics, Working paper No. 173, National Bureau of Economic Research, Cambridge.
- [18] Young, K.D.S. (1992). Influence of groups of observations on Bayes factors, *Communications in Statistics – Theory and Methods* **21**, 1405–1426.

LAWRENCE PETTIT

Information Matrix

JAY I. MYUNG AND DANIEL J. NAVARRO

Volume 2, pp. 923–924

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Information Matrix

Fisher information is a key concept in the theory of statistical inference [4, 7] and essentially describes the amount of information data provide about an unknown parameter. It has applications in finding the variance of an estimator, in the asymptotic behavior of **maximum likelihood estimates**, and in Bayesian inference (*see Bayesian Statistics*). To define Fisher information, let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample, and let $f(\mathbf{X}|\boldsymbol{\theta})$ denote the probability density function for some model of the data, which has parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Then the Fisher information matrix $I_n(\boldsymbol{\theta})$ of sample size n is given by the $k \times k$ symmetric matrix whose ij th element is given by the covariance between first partial derivatives of the log-likelihood,

$$I_n(\boldsymbol{\theta})_{i,j} = \text{Cov} \left[\frac{\partial \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_j} \right]. \quad (1)$$

An alternative, but equivalent, definition for the Fisher information matrix is based on the expected values of the second partial derivatives, and is given by

$$I_n(\boldsymbol{\theta})_{i,j} = -E \left[\frac{\partial^2 \ln f(\mathbf{X}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]. \quad (2)$$

Strictly, this definition corresponds to the *expected* Fisher information. If no expectation is taken, we obtain a data-dependent quantity that is called the *observed* Fisher information. As a simple example, consider a normal distribution with mean μ and variance σ^2 , where $\boldsymbol{\theta} = (\mu, \sigma^2)$. The Fisher information matrix for this situation is given by $I_n(\boldsymbol{\theta}) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{pmatrix}$.

It is worth noting two useful properties of the Fisher information matrix. Firstly, $I_n(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta})$, meaning that the expected Fisher information for a sample of n independent observations is equivalent to n times the Fisher information for a single observation. Secondly, it is dependent on the choice of parameterization, that is, how the parameters of a model are combined in the model's equation to define the probability density function. If the parameters are changed into new parameters by describing the latter as a function of the former, then the information matrix of the revised parameters can be found

analytically from the information matrix of the old parameters and the function that transforms the old parameters to the new ones [6].

The Cramer–Rao Inequality

Perhaps the most important application of the Fisher information matrix in statistics is in determining an absolute lower bound for the variance of an arbitrary unbiased estimator. Let $T(\mathbf{X})$ be any statistic and let $\psi(\boldsymbol{\theta})$ be its expectation such that $\psi(\boldsymbol{\theta}) = E[T(\mathbf{X})]$. Under some regularity conditions, it follows that for all $\boldsymbol{\theta}$,

$$\text{Var}(T(\mathbf{X})) \geq \frac{\left(\frac{d\psi(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \right)^2}{I_n(\boldsymbol{\theta})}. \quad (3)$$

This is called the *Cramer–Rao inequality* or the *information inequality*, and the value of the right-hand side of (3) is known as the famous *Cramer–Rao lower bound* [5]. In particular, if $T(\mathbf{X})$ is an unbiased estimator for $\boldsymbol{\theta}$, then the numerator becomes 1, and the lower bound is simply $1/I_n(\boldsymbol{\theta})$. Note that this explains why $I_n(\boldsymbol{\theta})$ is called the ‘information’ matrix: The larger the value of $I_n(\boldsymbol{\theta})$ is, the smaller the variance becomes, and therefore, we would be more certain about the location of the unknown parameter value. It is straightforward to generalize the Cramer–Rao inequality to the multiparameter case [6].

Asymptotic Theory

The maximum likelihood estimator has many useful properties, including reparametrization-invariance, consistency, and sufficiency. Another remarkable property of the estimator is that it achieves the Cramer–Rao minimum variance asymptotically; that is, it follows under some regularity conditions that the sampling distribution of a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$ is asymptotically unbiased and also asymptotically normal with its variance–covariance matrix obtained from the inverse Fisher information matrix of sample size 1, that is, $\hat{\boldsymbol{\theta}}_{ML} \rightarrow N(\boldsymbol{\theta}, I_1(\boldsymbol{\theta})^{-1}/n)$ as n goes to infinity.

Bayesian Statistics

Fisher information also arises in Bayesian inference (*see Bayesian Statistics*). The information matrix

2 Information Matrix

is used to define a noninformative prior that generalizes the notion of ‘uniform’. This is called *Jeffreys’ prior* [3] defined as $\pi_J(\boldsymbol{\theta}) \propto \sqrt{|I_1(\boldsymbol{\theta})|}$ where $|I_1(\boldsymbol{\theta})|$ is the determinant of the information matrix. This prior can be useful for three reasons. First, it is reparametrization-invariant so the same prior is obtained under all reparameterizations. Second, Jeffreys’ prior is a *uniform* density on the space of probability distributions in the sense that it assigns equal mass to each ‘different’ distribution [1]. In comparison, the uniform prior defined as $\pi_U(\boldsymbol{\theta}) = c$ for some constant c assigns equal mass to each different value of the parameter and is not reparametrization-invariant. Third, Jeffrey’s prior is the one that maximizes the amount of information about $\boldsymbol{\theta}$, in the Kullback–Leibler sense, that the data are expected to provide [2].

References

- [1] Balasubramanian, V. (1997). Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions, *Neural Computation* **9**, 349–368.
- [2] Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion), *Journal of the Royal Statistical Society, Series B* **41**, 113–147.
- [3] Jeffreys, H. (1961). *Theory of Probability*, 3rd Edition, Oxford University Press, London.
- [4] Lehman, E.L. & Casella, G. (1998). *Theory of Point Estimation*, 2nd Edition, Springer, New York.
- [5] Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society*, **37**, 81–91 (Republished in S. Kotz & N. Johnson, eds, *Breakthroughs in Statistics: 1889–1990*, vol. 1).
- [6] Schervish, M.J. (1995). *Theory of Statistics*, Springer, New York.
- [7] Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge.

JAY I. MYUNG AND DANIEL J. NAVARRO

Information Theory

PATRICK MAIR AND ALEXANDER VON EYE

Volume 2, pp. 924–927

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Information Theory

The first approach to formulating information theory was the communication model by **Shannon** and Weaver. Both authors worked at the Bell Laboratories. The aim of this research was the development of theoretical tools for the optimization of telephone engineering. The goal was the identification of the quickest and most efficient way to get a message from one location to another. The crucial questions were: (a) How can communication messages be converted into electronic signals most efficiently, and (b) how can signals be transmitted with a minimum of error? The results of this research culminated in the classic ‘The mathematical theory of communication’ [4].

The term *communication* can be used in a very broad sense. It includes all cases in which ideas can influence each other. Examples of such cases include words, writings, music, paintings, theater, opera, or, in brief, any human behavior. Because of the breadth of the definition, various problems of communication need to be considered. Shannon and Weaver [4] propose a three-level classification of communication problems:

1. Level A: technical problems;
2. Level B: semantic problems; and
3. Level C: effectiveness problems.

Level A is the one that is most accessible to analysis and engineering. *Technical problems* concern the accuracy of transmissions from sender (source) to receiver. Such a transmission can be an order of signs (written language), signals that change continuously (phone conversations, wireless connections), or two-dimensional patterns that change continuously (television). *Semantic problems* concern the interpretation of a message that the source sent to the receiver. Finally, *effectiveness problems* concern the success of the transmission: does a message lead to the intended response at the receiver’s end?

Levels B and C use the degree of accuracy determined by Level A. This implies that any restriction made on Level A has effects on Levels B and C. Thus, a mathematical description of technical problems is also of use at Levels B and C. These levels cover the philosophical issues of information theory. In the next section, we discuss some of the mathematical and technical issues of Level A.

Mathematical and Technical Issues of Information Theory

Shannon and Weaver [4] defined *information* as a measure of one’s freedom of choice when selecting a message. In this sense, *information* is equivalent to meaning. In different words, information refers to what could have been said instead of what has been said. As a system, communication can be represented as in Figure 1.

This communication model includes an information source and a receiver. Typically, the source encodes a message by translating it using a code in the form of bits. The word *bit* was proposed for the first time by **John W. Tukey**. It stands for *binary digit*, 0 and 1. To understand a message, the receiver must be able to decode the message. Thus, a code is a language or another set of symbols that can be used to transmit an idea through one or more channels.

An additional element in the communication model in Figure 1 is the *noise*. During a transmission, it can occur that undesirable bits of code are added to a message. Such disturbances, unintended by the source of the original information are, for instance, atmospheric disturbances, distortions of image and sound, or transmission errors.

Another reason for distortion of transmissions is channel overload. More information is fed into a channel than the channel can possibly transmit (see channel capacity, below). Because of the overload, information will be lost, and the received information will be different than the sent information.

A key question concerns the measurement of the amount of information in a message. There are two types of messages, coded as 0 and 1. The amount of information is defined as the logarithm of the number of possible choices. By convention, the logarithm with base 2 is used. If a message contains only one element of information, the amount of information transmitted is the logarithm of 2, base 2, that is, 1. This number is also called *one bit*. When a message contains four elements, there are 2^4 alternatives, and the amount of information contained in this message is $\log_2 16 = 4$.

To illustrate, consider a message with three elements. This message can contain the following eight strings of zeros and ones: 000, 001, 010, 011, 100, 101, 110, and 111. The amount of information carried by this message is $\log_2 8 = 3$, or 3 bits. In general, the amount of information contained in a message

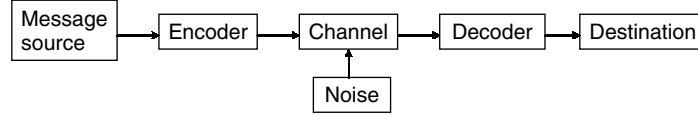


Figure 1 The communication system

with N elements is

$$H = \log_2 N, \quad (1)$$

which is known as the *Hartley formula* [2].

These first examples illustrate the theoretical constructs. An example of a practical application of the concept of information is the following: consider a person speaking. Let the elementary unit of this activity be the individual word. Thus, the speaker can select words and produce a sentence using these words. Let this sentence be the message. From an information theory perspective, it seems clear that after a particular word, it is more or less likely that another particular word will follow. For example, the probability that a noun follows after the word ‘the’ is higher than the probability that an adverb follows.

Some Background

The mathematical background of the study of sequences of words or other messages is given by Stochastic Processes and by **Markov Chains**. If a speaker builds a message word by word, the probability of the next word is considered given only by the immediately preceding word, but not by the words used before. This is the concept of a first-order Markov Chain. The mathematical and statistical terminology used in this context is that of an *Ergodic Markov Chain*, the most important case of Markov chains. In more technical terms, let P_i be the probability of state i , and $p_i(j)$ the probability of arriving at state j coming from state i . This probability is also called the *transition probability*. For a stationary process, the following constraint holds:

$$P_j = \sum_i P_i p_i(j). \quad (2)$$

In the ergodic case, it can be shown that the probability $P_j(N)$, that is, the probability of reaching state j after N signs converges to the equilibrium values

for $N \rightarrow \infty$. This statement holds for any starting distribution.

The term that takes all of the above concepts into account is *entropy*. Entropy is a function of the occurring probabilities of reaching a state in a transmission generating process, and the probability of transition from one state to another. Entropy displays the logarithm of these probabilities. Thus, the entropy can be viewed as a generalization of the logarithmic measure defined above for the simple cases. In other words, the entropy H is equivalent to information, and is, according to the Hartley formula, defined as

$$H = - \sum_{i=1}^N p_i \log p_i, \quad (3)$$

where $p_1, \dots, p_n$ are the probabilities of particular messages from a set of N independent messages.

The opposite of *information* is *redundancy*. Redundant messages add nothing or only little information to a message. This concept is important because it helps track down and minimize noise, for example, in the form of repeating a message, in a communicating system.

The aim of the original theory of information [3] was to find out how many calls can be transmitted in one phone transmission. This number is called the *channel capacity*. To determine channel capacity, it is essential to take the length of signs into account. In general, the channel capacity C is

$$C = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}, \quad (4)$$

where $N(T)$ is the number of permitted signals of length T . Based on these considerations, the following fundamental theorem of Shannon and Weaver can be derived: Using an appropriate coding scheme, a source is able to transmit messages via the transmission channel at an average transmission rate of almost C/H , where C is the channel capacity in bits per second and H is the entropy, measured in bits per sign. The exact value of C/H is never reached, regardless of which coding scheme is used.

Applications

Applications of information theory can be found, for instance, in electronics and in the social sciences. In electronics, information theory refers to engineering principles of transmission and perception. When, for instance, a person speaks into a telephone, the phone translates the sound waves into electrical impulses. The electrical impulses are then turned back into sound waves by the phone at the receiving end. In the social sciences, it is of interest how people are able or unable to communicate based on their different experiences and attitudes. An example of the use of information theory is given by Aylett [1].

The author presents a statistical model of the variation of clarity of speech. *Clarity* refers to the articulation of a word in various situations; for instance, spontaneous speech or reading words. He then investigated the degree to which a model formed for carefully articulated speech accounts for data from natural speech. Aylett states that the clarity of individual syllables is a direct consequence of a transmission process, and that a statistical model of clarity change provides insight into how such a process works. Results suggest that, if the speaker is in a noisy environment, the information content of the message should be increased in order to maximize the probability that the message is received.

Statistically, Aylett described the model of clarity variation using a density function that is composed of a mixture of Gaussians (see **Finite Mixture Distributions**)

$$Clarity = \frac{1}{n} \sum_{i=1}^n \log(p(x_i|M)), \quad (5)$$

where M is a clear speech model and n is a set of acoustic observations. The method for modeling clear speech and for comparing this model to actual speech is described in more detail in [1].

The first step of Aylett's data analysis concerned the relationship between the model and the psycholinguistic measures. Results indicate only a weak relationship between loss of intelligibility and clarity.

However, there was a stronger relationship between the speaker's average intelligibility in running speech and the average clarity of the speaker's running speech. In addition, the less intelligible a speaker's speech is, the poorer the fit of the final statistical model becomes.

In a second analytic step, the relationships between redundancy, articulation, and recognition were examined. Results suggest that clear speech is easier to recognize and to understand. Syllables that are pronounced with an emphasis are clearer than syllables with no emphasis. Aylett's model supports the assumption that out-of-context, high redundancy items in poorly articulated language are difficult to recognize. In addition, the model explains why speakers control levels of redundancy to improve the transmission process.

Aylett's study is an example of applying basic principles of information theory to modern psychological concepts. It is obvious that the main concepts defined by Shannon and Weaver [4] are not restricted to be used for modeling telephone transmission, but can be employed to answer a number of questions in the social and behavioral sciences, and statistics [5].

References

- [1] Aylett, M. (2000). Modelling clarity change in spontaneous speech, in *Information Theory and Brain*, R. Baddeley, P. Hancock & P. Földiack, eds, Cambridge University Press, Cambridge, pp. 204–220.
- [2] Hartley, R.V.L. (1928). Transmission of information, *The Bell System Technical Journal* **7**, 535–563.
- [3] Shannon, C.E. (1948). A mathematical theory of communication, *The Bell System Technical Journal* **27**, 379–423.
- [4] Shannon, C.E. & Weaver, W. (1949). *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.
- [5] von Eye, A. (1982). On the equivalence of the information theoretic transmission measure to the common χ^2 -statistic, *Biometrical Journal* **24**, 391–398.

PATRICK MAIR AND ALEXANDER VON EYE

Instrumental Variable

BRIAN S. EVERITT

Volume 2, pp. 928–928

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Instrumental Variable

The errors-in-variables model differs from the classical linear regression model in that the 'true' explanatory variables are not observed directly, but are masked by measurement error. For such models,

additional information is required to obtain consistent estimators of the parameters. A variable that is correlated with the true explanatory variable but uncorrelated with the measurement errors is one type of additional information. Variables meeting these two requirements are called *instrumental variables*.

BRIAN S. EVERITT

Intention-to-Treat

GEERT MOLENBERGHS

Volume 2, pp. 928–929

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Intention-to-Treat

A key type of intervention study is the so-called randomized **clinical trial** (RCT) [2–4, 9, 10]. In its basic form, a sample of volunteers is randomly split over two or more treatment groups. In doing so, baseline and other patient characteristics are ensured to be, on average, equal across groups, and, hence, differences in response on a clinical outcome can be ascribed solely and entirely to differences in treatment allocation. This is a powerful paradigm, since observed differences or, equivalently, associations between treatment assignment and differences in relevant outcome variables can then be given a causal interpretation. Apart from simple **randomization**, a number of variations to the randomization theme are in common use, such as **blocked randomization** and stratified randomization (*see Stratification*), to reduce the impact of chance, and, hence, to increase precision. But, whichever form is chosen, the goal is the same: to retain the interpretation of the average difference in response between the treatment groups as stemming from the treatment allocation itself, and not coming from other nuisance or confounding characteristics. The ability to reach an unbiased conclusion is a powerful asset of such a study, not shared by, for example, epidemiological or other observational studies.

However, in practice, this paradigm is jeopardized in two important ways, both of which stem from the fact that clinical trials are conducted in human subjects having a free will and, rightly so, carefully protected rights. First, some patients may not receive the treatment as planned in the study protocol, because they are sloppy with, for example, a rigorous treatment schedule, and, hence, may take less medication. Some may take more than planned at their own initiative. In rare cases, patients may even gain access to medication allocated to the other treatment arm(s). Hence, while patients remain on study, they do not follow the treatment regimen. Note that this is in line with actual practice, also outside of the clinical trial setting. Second, some patients may leave the study, some rather early after their enrollment in the trial, some at a later stage. In such cases, virtually no data or, at best, only partial data are available. This is bound to happen in studies that run over a relatively long period of time and/or when the treatment protocol is highly demanding. Again, this is in line

with the patient's rights. Having witnessed the terrible experiments conducted on humans during World War II, the Convention of Helsinki was passed. Ever since, clinical trial participation requires the patient to be given a clear and understandable statement about risks and benefits. This should be done by a qualified medical professional, and in the presence of an impartial witness. All have to sign the informed consent form. Then, still, the patient retains the right to withdraw from the study at any point in time, without the need to defend his or her decision.

Thus, after data have been collected, the researcher is faced with an incomplete sample, consisting of patients some having incomplete follow-up information, and some having followed a deviating treatment regimen.

It would then be tempting to adjust statistical analysis for discrepancies between the actual data and the way the study had been planned (*see Missing Data*). Such an approach is termed *as treated*. However, there is a key problem with such an analysis. Since dropout rates and/or deviations between the planned and the actual treatment regimen may be different between different treatment arms, the justification arising from randomization is undermined. Put differently, analyzing the data 'as treated' is likely to introduce confounding and, hence, bias.

As an answer to this, the so-called 'intention-to-treat' (ITT) principle has been introduced. It refers to an analysis that includes all randomized patients in the group to which they were randomly assigned, regardless of their adherence to the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol. While this may look strange and even offending to the novice in the clinical trial field, the principle is statistically widely accepted as providing valid tests about the null hypothesis of no treatment effect. It also refers to actual practice, where it is even more difficult to ensure patients follow the treatment as planned. The term 'intention to treat' appears to have been coined by Hill [5]. An early but clear account can be found in [11]. Careful recent accounts are given by Armitage [1] and McMahon [7].

If one is interested in the true efficacy of a medicinal product or an intervention beyond simply testing the null hypothesis, an ITT analysis is not the right tool. However, because of the bias referred to earlier, an analysis 'as treated' is not appropriate

either since it is vulnerable to bias. A large part of the recent incomplete data and so-called noncompliance literature is devoted to ways of dealing with this question [12]. This issue is nontrivial since the only definitive way to settle it would be to dispose of the unavailable data, which, by definition, is impossible. Whatever assumptions made to progress with the analysis, they will always be unverifiable, at least in part, which typically results in sensitivity to model assumptions.

The translation of the ITT principle to longitudinal clinical studies (see **Longitudinal Data Analysis**), that is, studies where patient data are collected at multiple measurement occasions throughout the study period, is a controversy in its own right. For a long time, the view has prevailed that only carrying the last measurement (also termed *last value*) actually obtained on a given patient forward throughout the remainder of the follow-up period is a sensible approach in this respect. (This is known as last observation carried forward (LOCF).) With the advent of modern likelihood-based longitudinal data analysis tools, flexible modeling approaches that avoid the need for both imputation and deletion of data have come within reach. Part of this discussion can be found in [6, 8].

References

- [1] Armitage, P. (1998). Attitudes in clinical trial, *Statistics in Medicine* **17**, 2675–2683.
- [2] Buyse, M.E., Staquet, M.J. & Sylvester, R.J. (1984). *Cancer Clinical Trials: Methods and Practice*, Oxford University Press, Oxford.
- [3] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, Springer-Verlag, New York.
- [4] Green, S., Benedetti, J. & Crowley, J. (1997). *Clinical Trials in Oncology*, Chapman & Hall, London.
- [5] Hill, A.B. (1961). *Principles of Medical Statistics*, 7th Edition, The Lancet, London.
- [6] Mazumdar, S., Liu, K.S., Houck, P.R. & Reynolds, C.F. III. (1999). Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values, *Journal of Psychiatric Research* **33**, 87–95.
- [7] McMahon, A.D. (2002). Study control, violators, inclusion criteria and defining explanatory and pragmatic trials, *Statistics in Medicine* **21**, 1365–1376.
- [8] Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C. & Carroll, R.J. (2004). Analyzing incomplete longitudinal clinical trial data, *Biostatistics* **5**, 445–464.
- [9] Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*, John Wiley, New York.
- [10] Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*, John Wiley, Chichester.
- [11] Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials, *Journal of Chronic Diseases* **20**, 637–648.
- [12] Sheiner, L.B. & Rubin, D.B. (1995). Intention-to-treat analysis and the goals of clinical trials, *Clinical Pharmacology and Therapeutics* **57**, 6–15.

(See also **Dropouts in Longitudinal Data; Dropouts in Longitudinal Studies: Methods of Analysis**)

GEERT MOLENBERGHS

Interaction Effects

LEONA S. AIKEN AND STEPHEN G. WEST

Volume 2, pp. 929–933

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Interaction Effects

In statistical analysis, we often examine the simultaneous effect of two or more variables on some outcome. Interaction effects refer to the effects of particular unique combinations of variables on an outcome that would not be expected from their average effects. Consider a statistics professor who is trying to increase performance in her graduate statistics course. She uses two study aids: a detailed study guide and a comprehensive review session. In an experiment, she tries all four combinations of providing or not providing the study guide with providing or not providing the review session. Suppose the result is as shown in Figure 1(a). When there is a study guide (left-hand pair of bars), the review session adds 30 points to performance; the same is true when there is no study guide (right-hand pair of bars). Looked at the other way, with no review session (black bars), the study guide adds 20 points to performance; with a review session (striped bars), the study guide adds the same 20 points. There is no interaction between the study guide and the review session; each has a constant effect on test performance regardless of the other.

Now, consider an alternative outcome given in Figure 1(b). When there is no study guide (right-hand pair of bars), the review session adds 30 points to performance, just as before. But, when there is a study guide (left-hand pair of bars), the review session lowers scores 10 points, perhaps due to information overload! Looked at the other way, with no review session (black bars), the study guide adds 50 points (from 40 to 90). Yet, with the review session (striped bars), the study guide adds only 10 points (from 70 to 80). The effect of each study aid depends on the other study aid; there is an interaction between the study guide and the review session.

Three Characterizations of Interactions

There are three interrelated characterizations of interactions, as (a) conditional effects, (b) nonadditive effects (*see Additive Models*), and (c) as residual effects over and above the individual effects of each variable. These are best explained with reference to the table of arithmetic means associated with Figure 1, given in Table 1. In the table, a cell mean

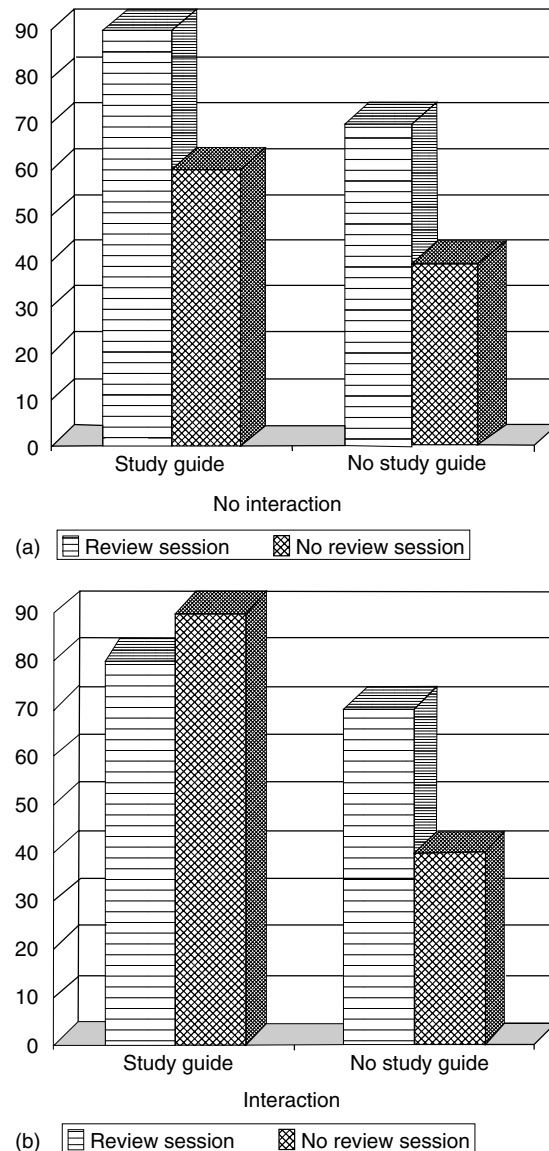


Figure 1 Effects of a study guide and a review session on statistics examination performance. In Figure 1(a), there is no interaction. In Figure 1(b), study guide and review session interact

refers to the mean at one combination of study guide (yes/no) and review session (yes/no), for example, the cell mean of 70 in Table 1(a) for the cell, 'No Study Guide/Review Session'. The column means are average effects of the study guide variable collapsed over review session; the row means are average effects of

2 Interaction Effects

Table 1 Cell means, marginal means (row and column means), and grand mean for performance on a statistics examination as a function of study guide and review session

	Table 1(a): No interaction			Table 1(b): Interaction		
	Study guide		Row mean	Study guide		Row mean
	Yes	No		Yes	No	
Review Session						
Yes	90	70	80	80	70	75
No	60	40	50	90	40	65
Column Mean	75	55	65	85	55	70

review session collapsed over study guide. The grand mean (65 in Table 1(a)) is average performance over all four cells.

Interactions as Conditional Effects

Conditional effects are the effects of one variable at a particular level of another variable. The effect of the review session when a study guide is also given is one conditional effect; effect of review session without a study guide is a second conditional effect. When there is no interaction, as in Table 1(a), the conditional effects of a variable are constant over all levels of the other variable (here the constant 30-point gain from the review session). If there is an interaction, the conditional effects of one variable differ across values of the other variable. As we have already seen in Table 1(b), the effect of the review session changes dramatically, depending on whether a study guide has been given: a 30-point gain when there is no study guide versus a 10-point loss in the presence of a study guide. One variable is said to be a **moderator** of the effect of the other variable or *to moderate* the effect of the other variable; here we say that study guide moderates the effect of review session.

Interactions as Nonadditive Effects

Nonadditive effects signify that the combination of two or more variables does not produce an outcome that is the sum of their individual effects. First, consider Table 1(a), associated with Figure 1(a) (no interaction); here, the cell means are *additive effects* of the two variables. In Table 1(a), the row means tell us that the average effect of study guide is a 30-point gain, from 50 to 80; the column means tell us that the average effect of review session is 20 points, from 55 to 75. With neither study guide nor review

session, the cell mean is 40; introducing the study guide yields a 30-point gain to a cell mean of 70; then, introducing the review session yields another 20-point gain to a cell mean of 90. Table 1(b), associated with Figure 1(b) (interaction) contains *nonadditive effects*. The row mean shows a 10 point average gain from the study guide, from 65 to 75. The column mean shows as a 30-point gain from the review session, from 55 to 85. However, the cell means do not follow the pattern of the marginal means. With neither study guide nor review session, the cell mean is 40. Introducing the study guide yields a 50-point gain to 90, and not the 30-point gain expected from the marginal mean; then, introducing the review session on top of the study guide yields a loss of 10 points, rather than the gain of 10 points expected from the marginal means. The unique combinations of effects represented by the cells do not follow the marginal means.

Interactions as Cell Residuals

The characterization of interactions as cell **residuals** [8] follows from **the analysis of variance** framework [5, 6]. By *cell residual* is meant the discrepancy between the cell mean and the grand mean that would not be expected from the additive effects of each variable. When there is an interaction between variables, the cell residuals are nonzero and are pure measures of the amount of interaction. When there is no interaction between variables, the cell residuals are all zero.

Types of Interactions by Variable (Categorical and Continuous)

Categorical by Categorical Interactions

Thus far, our discussion of interactions is in terms of variables that take on discrete values, categorical

by categorical interactions, as in the factors in the analysis of variance (ANOVA) framework. In the ANOVA framework, the conditional effects of one variable at a value of another variable, for example, the effect of review session when there is no study guide, is referred to as a *simple main effect*. See [5] and [6] for complete treatments of interactions in the ANOVA framework.

Categorical by Continuous Variable Interactions

We can also characterize interactions between categorical and continuous variables. To continue our example, suppose we measure the mathematics ability of each student on a continuous scale. We can examine whether mathematics ability interacts with having a review session in producing performance on a statistics examination. Figure 2(a) illustrates an interaction between these variables. For students who do not receive the review session, there is a strong positive relationship between mathematics ability and

performance on the statistics examination. However, the review session has a compensatory effect for weaker students. When students receive a review session, there is a much-reduced relationship between mathematics ability and performance; the weaker students ‘catch up’ with the stronger students. Put another way, the effect of mathematics ability on performance is conditional on whether or not the instructor provides a review session. An introduction to the categorical by continuous variable interaction is given in [1] with an extensive treatment in [10].

Continuous by Continuous Variable Interactions

Finally, two or more continuous variables may interact. Suppose we have a continuous measure of motivation to succeed. Motivation may interact with mathematics ability, as shown in Figure 2(b). The relationship of ability to performance is illustrated for three values of motivation along a motivation continuum. The effect of ability becomes increasingly more positive as motivation increases – with low motivation, ability does not matter. The effect of ability is conditional on the strength of motivation; put another way, motivation moderates the relationship of ability to performance.

Both continuous by continuous and continuous by categorical interactions are specified and tested in the **multiple linear regression** (MR) framework. In MR, the regression of performance on mathematics ability at one value of motivation is referred to as a *simple regression*, analogous to a simple main effect in ANOVA. In Figure 2(b), we have three simple regression lines for the effects of ability on performance, each at a different value of motivation. A complete treatment of interactions in the multiple regression framework, with prescriptions for probing and interpreting interactions involving continuous variables, is given in [1]; see also [2, 4].

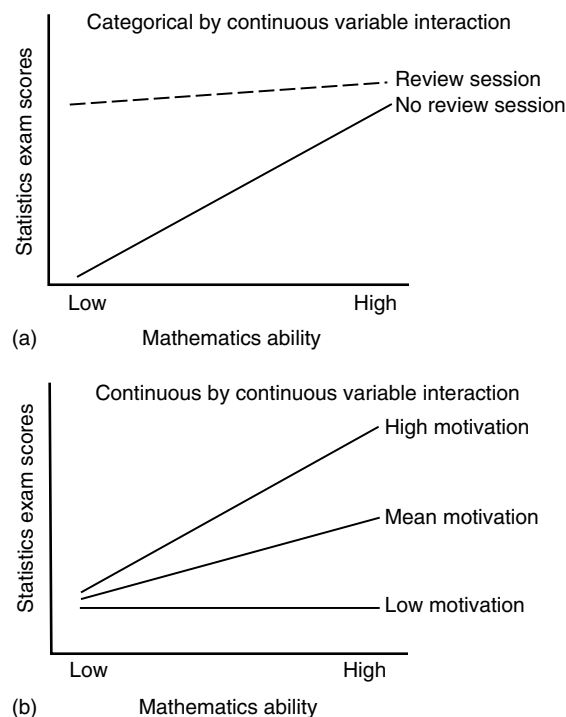


Figure 2 Interactions involving continuous variables. (a) Categorical by continuous variable interaction. (b) Continuous by continuous variable interaction

Types of Interactions by Pattern

Patterns of interactions are characterized in a variety of ways, regardless of the combination of categorical and continuous variables comprising the interaction. We consider two such categorizations: (a) crossover versus noncrossover interactions, and (b) synergistic versus buffering interactions versus compensatory interactions.

Crossover versus Noncrossover Interactions

Crossover interactions (or *disordinal interactions*) are ones in which the direction of effect of one variable reverses as a function of the variable with which it interacts. Figure 1(b) illustrates a crossover interaction. With no study guide, performance is superior with the review session; with a study guide, performance is superior without a review session. *Non-crossover interactions* (or *ordinal interactions*) are ones in which the direction of effect of one variable is constant across the values of the other variable with which it interacts. Both interactions in Figure 2 are noncrossover interactions. In Figure 2(b), performance is always higher for higher motivation across the range of ability.

Synergistic versus Buffering Interactions versus Compensatory Interactions

Synergistic interactions are ones in which two variables combine to have a joint effect that is even greater than the sum of the effects of the variables. In Figure 2(b), performance increases with ability; it increases with motivation. There is a synergy between ability and motivation – being high on both produces superb performance. *Buffering interactions* are ones in which variables are working in opposite directions, and the interaction is such that the effect of one variable weakens the effect of the other variable. Suppose we assess the number of other demands on students when they are studying for an exam. As other demands increase, performance is expected to decrease due to lack of study time. However, the study guide buffers or weakens the impact of other demands on performance by making studying much more focused and efficient. We would say that the use of the study guide buffers the negative effect of other demands. *Compensatory (or interference) interactions* are those in which both variables have an effect in the same direction, but in which one variable compensates for the other, an ‘either-or’ situation. Figure 2(a) illustrates such an interaction. Both ability and the review session have a positive effect on

performance; however, the review session weakens the positive effect of ability.

Interactions Beyond the ANOVA and Multiple Regression Frameworks

The specification and testing of interactions extends to the **generalized linear model**, including **logistic regression** [3], to hierarchical linear models [7] (see **Linear Multilevel Models**), and to structural equation modeling [9].

References

- [1] Aiken, L.S. & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*, Sage, Newbury Park.
- [2] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.
- [3] Jaccard, J. (2001). *Interaction Effects in Logistic Regression*, Sage, Thousand Oaks.
- [4] Jaccard, J. & Turrisi, R. (2003). *Interaction Effects in Multiple Regression*, 2nd Edition, Sage, Newbury Park.
- [5] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Co, Pacific Grove.
- [6] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Approach*, Lawrence Erlbaum, Mahwah.
- [7] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, Thousand Oaks.
- [8] Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*, 2nd Edition, McGraw Hill, New York.
- [9] Schumacker, R.E. & Marcoulides, G.A., eds (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*, Sage, Mahwah.
- [10] West, S.G., Aiken, L.S. & Krull, J. (1996). Experimental personality designs: analyzing categorical by continuous variable interactions, *Journal of Personality* **64**, 1–47.

LEONA S. AIKEN AND STEPHEN G. WEST

Interaction Plot

PAT LOVIE

Volume 2, pp. 933–934

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Interaction Plot

An interaction plot is a graphical tool for examining the interactions or dependencies between variables (factors) in designed experiments. Its most common role is to help with the interpretation of results from an **analysis of variance (ANOVA)**, but it can also be used as an exploratory device for determining whether an additive or interaction model is appropriate (*see Interaction Effects*).

An interaction plot for, say, a two-factor experimental design is an x–y line plot of cell (level combination) means on the response variable for each level of one factor over all the levels of a second factor. Thus, if the lines or profiles corresponding to the separate factor levels are parallel, that is, no differential effect over different combinations of the levels of the factors is revealed, then there is no interaction. If they are not parallel, then an interaction is present. Of course, since the cell means are sample means, we have to exert a degree of commonsense in what we declare as parallel and nonparallel.

Figures 1(a–c) illustrate some possible patterns for a two-factor design in which Factor A has three levels and Factor B has two levels. In Figure 1(a) we have parallel lines because the distance between the cell means for levels B1 and B2 of Factor B is essentially the same whichever level of Factor A we are looking at, that is, there is no interaction. By contrast, in Figures 1(b) and 1(c) the differences between the means vary over the levels of A, so the lines are not parallel and we conclude that there is an interaction. However, note that in Figure 1(b) the means for B1 are always greater than for B2; an interaction of this type is sometimes described as ‘ordinal’. On the other hand, in Figure 1(c), we have a crossover at A1 indicating a ‘disordinal’ interaction.

When there are three or more factors in the design, the number of graphs required becomes more unwieldy and the interpretation correspondingly more difficult. For instance, to investigate the first- and second-order interactions in a three-factor design, we would need a plot for each of the three pairs of factors, and then a sequence of plots for two of the factors at each level of the third.

Statistical packages such as SPSS and Minitab offer interaction plots as options for their ANOVA

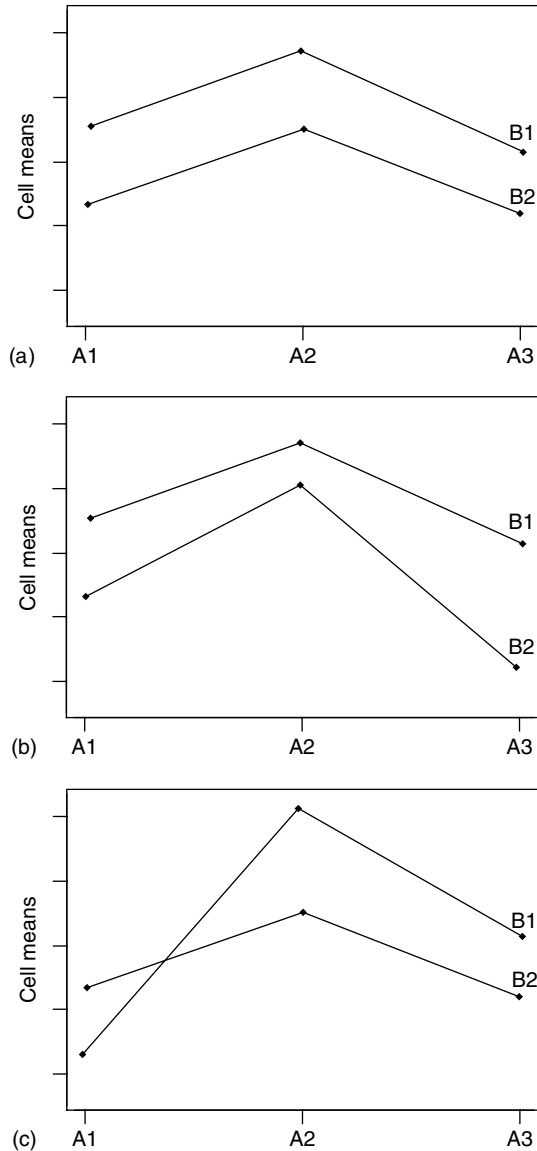


Figure 1 (a) No interaction; (b) ordinal interaction; (c) disordinal interaction

or **generalized linear model (GLM)** (*see Software for Statistical Analyses*) procedures.

PAT LOVIE

Internal Consistency

NEAL SCHMITT

Volume 2, pp. 934–936

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Internal Consistency

Internal consistency usually refers to the degree to which a set of items or measures are all indicators of a single construct or concept. Arithmetically, internal consistency is usually evaluated by assessing the degree to which the items in a measure correlate or covary highly with each other or with the total of the scores or responses in the remainder of the measure. Related notions are that of test homogeneity, again the degree to which items in a measure are indices of the same construct and unidimensionality, which also refers to the degree to which items measure a single construct.

Internal consistency, homogeneity, and unidimensionality are often used interchangeably when referring to a test and the interrelationships between the items comprising the test. Hattie [1] provided a review of 30 different methods of determining or assessing unidimensionality. Those investigators who focus on unidimensionality have usually proposed an index on the basis of **factor analysis** or **principal component analysis**. In these cases, a perfectly internally consistent measure would be that in which the first factor or component accounts for all the covariation between items. Those who focus on the reliability of items will usually focus on some version of coefficient alpha (see below) or the average correlation or covariance between items designed to measure a single construct. Still others will examine the pattern of responses across items. Obviously, the manner in which internal consistency or unidimensionality is operationalized varies with the investigator. Not often discussed in quantitative treatments of internal consistency is the obvious requirement that all items measure conceptually similar content.

Coefficient alpha is perhaps most often considered an index of internal consistency, though technically many would hold that its calculation involves an assumption of unidimensionality rather an index of it. Coefficient alpha is the ratio of the product of the squared number of items in a measure and the average covariance between items in it to the total variance in the measure. Formally,

$$\text{Alpha} = \frac{n^2(\text{average cov}_{ij})}{\sigma^2}, \quad (1)$$

where n is the number of items in the measure, cov_{ij} is the covariance between items i and j , and σ^2 is the variance of the test. In standardized terms, this index is the ratio of the product of the squared number of items in the measure and the average correlation among items in it divided by the sum of the correlations among all items in the test including the correlations of items with themselves. Formally, this is similar to the unstandardized version of alpha:

$$\text{Standardized alpha} = \frac{n^2(\text{average cor}_{ij})}{R}, \quad (2)$$

where n is the number of items in the measure, cor_{ij} is the correlation between items i and j , and R is the sum of the correlations among the items including elements above and below the diagonal as well as the diagonal 1.00s. Theoretically, this ratio is based on the notion that if items in a measure are all indices of the same construct, they should correlate perfectly. Correlations between items are usually less than 1 (likewise, covariances between items do not approach the variances of the items involved), and the difference between diagonal elements and off-diagonal elements in an item intercorrelation matrix is treated as error or lack of internal consistency reliability.

The problem with considering coefficient alpha as an index of internal consistency is that with a large number of items, the ratio referred to above can yield a relatively high value even though there is evidence that items measure more than one construct. This would be the case when there are clusters of items that are relatively highly correlated with each other but not highly correlated with items in other clusters in the test [2]. In this case, coefficient alpha might indicate high internal consistency, while an observation of the correlations among the items or a factor analysis of the items would lead a researcher to conclude that more than one construct is accounting for the between-item covariances.

So internal consistency should be indicated by relatively high intercorrelations between items with little or no variability in the intercorrelations or item covariances. Any variability (apart from variability that can be accounted for by sampling error) in item covariances or item intercorrelations is evidence that more than one construct is being measured. Also, the very obvious, but often ignored, condition is

2 Internal Consistency

that all items in the measure address conceptually similar content.

- [2] Schmitt, N. (1996). Uses and abuses of coefficient alpha, *Psychological Assessment* **8**, 81–84.

NEAL SCHMITT

References

- [1] Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9**, 139–164.

Internal Validity

DAVID C. HOWELL

Volume 2, pp. 936–937

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Internal Validity

Campbell and Stanley [1] made a two-part distinction between internal validity, which referred to the question of whether the data for a particular study justified drawing the conclusion that there was a relationship between the independent and dependent variable, and whether it was causal, and **external validity**, which referred to the degree to which the results of this experiment would generalize to other populations, settings, and variables. Cook and Campbell [2] and Shadish et al. [3] carried these distinctions further. They added statistical conclusion validity, which referred to whether the statistical conclusions were appropriate to establish that the variables covaried, and separated that from internal validity, which dealt with whether this covariance could be given a causal interpretation.

Suppose, for example, that an experimental and control group were found to differ with respect to a measure of self-esteem. If the difference between the groups was significant, but the data strongly violated the assumption of normal distribution of errors, then the statistical conclusion validity of the experiment could be compromised. If we subsequently learn that the experimental group consisted of all girls, and the control group contained all boys, we would be forced to admit that any difference that resulted could as easily be attributed to gender differences as to the experimental treatment. Such a study would be lacking in internal validity. Campbell and Stanley referred to internal validity as the *sine qua non* of interpretability of experimental design.

Campbell and Stanley [1] popularized eight different classes of extraneous variable that can jeopardize internal validity. Others have added additional classes over time. The following list is based on the nine different classes proposed by Shadish et al. [[3]]. Shadish et al. presented experimental designs that address threats from each of these sources of invalidity; those designs are given only slight reference here. Additional information can be obtained from the references and from several excellent sites on the Internet.

Threats to Internal Validity

There are at least 14 sources of threat to internal validity. They include the following:

- Ambiguous temporal precedence
 - A basic foundation of imputing a causal effect is that the cause must precede the effect (*see Hill's Criteria of Causation*). When data are collected in such a way that temporal priority cannot be established (e.g., in a cross-sectional study), our ability to indicate the direction of cause is compromised, and the internal validity of the study is compromised.
- Selection
 - This threat refers to differences between the two groups in terms of participant assignment. Unless participants are assigned at random to conditions, any resulting differences might be attributable to preexisting group differences rather than to the treatment itself.
- History
 - Differences between pre- and posttesting could be attributable to any event that occurred between the two treatments in addition to an experimental intervention. This threat can often be addressed by including a randomly assigned control group that is assessed for pre- and posttest measures, but does not receive the treatment. Presumably, that group is as likely as the experimental group to be affected by historical events, and any *additional* effect in the treatment group would represent the introduction of treatment.
- Maturation
 - Maturation refers to any changes in the participants as a result of the passage of time. If, for example, participants judged the desirability of cookies before and after a lecture on nutrition, the passage of time, and the resulting increase in hunger, might have an important effect on the outcome, independent of any effect of the lecture. Appropriate compensatory designs include a group that experiences the same maturation opportunity but not the intervention.
- Regression to the Mean
 - When groups are selected on the basis of extreme scores, differences on a subsequent measure could reflect statistical **regression to the mean**. For example, if we select the worst students in the class and provide additional tutoring, later scores may show an improvement that cannot be unequivocally attributed to our intervention. Someone scoring very

2 Internal Validity

- poorly on one test is much more likely to score better, rather than worse, on a subsequent test. Here again pre- and posttest scores from an untreated control group provide an opportunity to partial out change attributable to regression to the mean.
- Attrition or Experimental Mortality
 - Differential participant **attrition** for two or more groups can produce differences between treatments that are attributable to the attrition rather than to the treatments. Suppose that after an initial test, participants are split into two groups, one receiving special tutoring and the other assigned to watch a video. Some of those watching the video may become bored because of a short attention span and as a result, engage in behavior that results in their being expelled from the study. Any subsequent differences between treatment groups may be due to higher average attention span in the remaining members of the control group.
 - Repeated Testing
 - The effect of having taken one test on the scores that result from a subsequent test. For example, the common belief that repeated practice with the SAT exam leads to better test-taking skills would suggest that a second administration of the SAT would lead to higher scores, independent of any gain in actual knowledge.
 - Instrumentation
 - Instrumentation refers to the change in calibration of an instrument over time. While we normally think of this as applying to instruments taking physical measurement, such as a scale or meter, it applies as well to paper and pencil tests or human scorers who no longer measure the same thing that they measured when the test was developed.

- Interaction of threats
 - Each of the preceding threats may interact with any of the other preceding threats. Suppose, for example, that the control group is comprised of mostly preadolescent males, while participants in the experimental group have just entered adolescence. We might anticipate a greater maturational change between the pretest and posttesting in the experimental group than we would in the control group. This makes it impossible to clearly assign any differences we observe to the effect of our treatment.

For other threats to invalidity, *see* **Reactivity**, **Expectancy Effect by Experimenters**, **Demand Characteristics**, and **Hawthorne Effect**. For approaches to dealing with these threats, *see* **Nonequivalent Group Design**, **Regression Discontinuity Design**, and, particularly, **Quasi-experimental Designs**.

References

- [1] Campbell, D.T. & Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Skokie.
- [2] Cook, T. & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin, Boston.
- [3] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.

Further Reading

- Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research: Enlarged Edition*, Irvington Publishers, New York.

DAVID C. HOWELL

Internet Research Methods

DAVID W. STEWART

Volume 2, pp. 937–940

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Internet Research Methods

The Internet and the World Wide Web have dramatically changed the ways in which research is conducted. New tools, such as on-line information directories and search engines, have made finding information faster and more efficient than browsing library stacks that once characterized much research. Networked computers and related software provide a means for presenting stimuli and questions to respondents in many different places, often in real time. The Internet has reduced the spatial and temporal boundaries that once limited both the ready availability of extant information and the ability to gather new information. Information is posted and maintained on the Internet by government organizations, commercial businesses, universities, news organizations, research laboratories, and individuals, among others. Part of the power of the World Wide Web is that it makes such information readily accessible in a common format to any user. Text, images, audio, and video are all available on the Web.

The power of the Internet is not without its costs, however. The amount of information that is available on the Internet is daunting. A simple search can produce thousands of 'hits' that appear to be relevant. Links from one information site to another can greatly expand the scope of information searches. Distilling and integrating the volumes of information produced in a search is a significant task. Ideally, a researcher would attempt to narrow a search to only the most relevant sources, but the relevance of information is not always obvious. Even when information is narrowed in scope, there are questions about the credibility and reliability of the information. Thus, while it may appear that the Internet has taken much of the work out of finding information, the reality is not so simple. Finding information is easier and faster on the Internet, but the real work of evaluating, interpreting, and integrating such information, which is the goal of research, is as difficult as ever.

This article provides an introduction to the use of the Internet as a research tool and a brief overview of research tools and sources. The article also provides references for further exploration of information on the Internet, and for discussion of the evaluation

and use of information and data obtained from the Internet.

Two Types of Research

Secondary Research

Secondary research uses data and information collected by others and archived in some form. Such information includes government records and reports, industry studies, archived data, and specialized information services, as well as digitized versions of books and journals found in libraries. Much of the information that is accessible on the Internet is secondary information. Secondary information takes a variety of forms and offers relatively quick and inexpensive answers to many questions. It may be little more than a copy of a published report or it may involve a reanalysis of original data. For example, a number of syndicated research providers obtain government data, such as that obtained by the Census Bureau, and develop specialized reports, provide data access capabilities, or combine data from multiple sources. Several syndicated research providers obtain electronic scanner information from retailers and provide reports on the sales, prices, and other features of retail products. Other syndicated research providers collect information about product awareness, product preference, and customer satisfaction for entire industries. Similarly, some organizations offer reports of large-scale tracking studies of political trends, media usage habits, lifestyles, and purchasing habits. Table 1 provides a listing of representative Web sites that provide secondary information and data.

Secondary research generally offers a faster and less expensive means for obtaining information than collecting new data. Data and reports are already available, so they often can be obtained immediately. However, secondary research may not provide the specific information required for a researcher's purpose, and it may not be as up-to-date as would be desirable.

Much of the research that takes place on the Internet involves the use of tools, such as directories and search engines, to identify secondary sources of information. Directories are lists of resources arranged by subject that have been created for the specific purpose of organizing information content. Much like the table of contents

2 Internet Research Methods

Table 1 Useful secondary sources of information on the internet

Census Bureau (http://www.census.gov) Site of the United States Census Bureau with links to comparable sites in other countries.	Economist Intelligence Unit (http://store.eiu.com) Source of international information provided by the publisher of <i>The Economist</i> .
STAT-USA – Commerce Department (http://www.stat-usa.gov) Useful government data on business and commerce.	Commerce Department Commercial Service (http://www.usatrade.gov) Provides information related to exporting and guides to specific international markets.
FedWorld (http://www.fedworld.gov) A great place to find information from many different government sources.	Fed Stats (http://www.fedstats.gov) A gateway to data collected by virtually every federal agency. Includes a link to <i>Statistical Abstracts</i> .
Edgar (http://www.sec.gov/edaux) The place to find information about companies. This site includes 10 K and 10 Q filings of U. S. public companies.	C. A. C. I. (http://demographics.caci.com) Provides very detailed profiles of markets and geographic areas.
American Demographics (http://www.marketingtools.com) Home page of American Demographics magazine. Provides basic demographic and economic data.	Claritas (http://www.connect.claritas.com) Similar to C.A.C.I. Provides market and geographic information at various levels of aggregation down to the zip-code level.

of a book, a directory provides a list of topics contained within a data source. Perhaps the best known and largest directory on the Internet is Yahoo (<http://www.yahoo.com>). Other common directories include Hotbot (<http://www.lycos.hotbot.com>), Looksmart (<http://www.looksmart.com>), Galaxy (<http://www.Galaxy.com>) and the Librarians' Index to the Internet (<http://www.lii.org>). Directories are especially useful research tools when the researcher has identified a general area of interest and wishes to take advantage of the organizational structure provided by a directory. Also, use of a directory narrows the content, and that can make finding information more efficient. In contrast to directories, search engines provide a means for doing keyword searches across the Internet. They are most useful when the researcher needs specific information found in many sources and across many different directories. Search engines differ in the way they search the Internet and the way they organize the results of the search. Among the better known search engines are Google (<http://www.google.com>), All the Web (<http://www.alltheweb.com>), Alta Vista (<http://www.altavista.com>), and Lycos (<http://www.lycos.com>). Sometimes, a researcher will find it useful to search multiple directories and/or search engines simultaneously. Metasearch tools such as Proteus Internet Search

(<http://www.thrall.org/proteus.html>), Ixquick (<http://www.ixquick.com>), Metacrawler (<http://www.metacrawler.com>), and Dogpile (<http://www.dogpile.com>) are useful for such broad searches.

The use of directories and search engines can produce quick results. Such searches may also produce many irrelevant or out-of-date sources. For this reason, experienced researchers tend to identify and bookmark those Web sites that they find most useful and to which they return again and again.

Primary Research

The Internet also can be used for conducting primary research, that is, for obtaining data in the first instance. Primary research involves the design of research and the collection and analysis of data that address specific issues identified by the researcher. While there are many different types of primary research, most primary research can be classified as either qualitative or quantitative. **Qualitative research** is most useful for generating ideas, for generating diagnostic information, and for answering questions about why something is happening. Individual interviews and focus groups are two of the more common qualitative research techniques, and

Table 2 Representative tools for conducting research on the internet**SurveyMonkey** (<http://www.surveymonkey.com>)

A tool for designing and conducting on-line survey research. Free for surveys involving up to 10 items and 100 respondents.

Zoomerang (<http://info.zoomerang.com>)

Another on-line survey research tool. Provides templates for more than a 100 common types of surveys and preconstructed panels of potential respondents with known characteristics.

SurveySite (<http://www.surveysite.com>)

Web site for a commercial provider of on-line focus groups and other on-line research tools.

E-Prime (<http://www.pstnet.com>)

A software package published by Psychology Software Tools, Inc. that enables the design and administration of behavioral experiments on-line.

both are now routinely conducted via the Internet. The interactive nature of the Internet with threaded discussions, chatrooms, and bulletin boards provides an ideal venue for the dialog that characterizes qualitative research. Quantitative tools, which include surveys, are most useful for obtaining numeric summaries that can be used to characterize larger populations. On-line surveys have become ubiquitous on the Web and provide a means for obtaining current information from survey respondents.

The Internet provides a powerful means for conducting primary behavioral research. The Internet can be used to present stimulus materials to respondents ranging from text to audio to streaming video. The respondents can be asked to respond immediately. Stimulus materials are easily customized with a computer. Hence, it is easy to construct alternative treatment conditions and scenarios for the conduct of on-line experiments. Specialized software has been developed to facilitate the conduct of both qualitative and quantitative research on the Internet. Table 2 provides a brief description of representative types of such software.

Evaluating Research on the Internet

Not all information on the Internet is equally reliable or valid. Information must be evaluated carefully and weighted according to its recency and credibility. The same questions that arise in the evaluation of secondary sources also arise in the context of primary research. The only difference is that these questions must be addressed retrospectively for secondary research, while they must be addressed prospectively for primary research. When evaluating information, six questions must be answered: (a) What was (is) the

purpose of the study? (b) Who collected (or will collect) the information? (c) What information was (will be) collected? (d) When was (will) the information (be) collected? (e) How was (will) the information (be) obtained? (f) How consistent is the information with other sources? In answering these basic questions, other more specific questions will arise. These more specific questions include the source(s) of the data, measures used, the time of data collection, and the appropriateness of analyses and conclusions.

Summary

The Internet provides powerful tools for obtaining data. However, it is always necessary to carefully evaluate the information. There are numerous comprehensive treatments of the use of the Internet as a research tool [1–3]. It is important to recognize that the Internet has brought with it some unique challenges, including concerns about its intrusiveness, invasion of individuals' privacy, and its appropriate use with children [4]. The Internet is a tool for efficiently obtaining information but its power assures that it will yield a great deal of irrelevant and unreliable information. The researcher's task is to determine how to best use the information.

References

- [1] Ackermann, E. & Hartman, K. (2003). *Searching and Researching on the Internet and the World Wide Web*, Third Edition, Franklin, Beedle & Associates, Wilsonville.
- [2] Dillman, D.A. (2000). *Mail and Internet Surveys*, 2nd Edition, John Wiley & Sons, New York.

4 Internet Research Methods

- [3] Hewson, C., Peter, Y., Carl, V. & Dianna, L., (2003). *Internet Research Methods: A Practical Guide for the Social and Behavioural Sciences*, Sage, Thousand Oaks.
- [4] Kraut, R., Judith O., Mahzarin, B., Amy, B., Jeffrey, C. & Couper, M. (2002). *Psychological Research Online: Opportunities and Challenges*, Monograph of the APA Board of Scientific Affairs Advisory Group on Conducting Research on the Internet, American Psychological Association, Washington, (available online at <http://www.apa.org/science/bsaweb-agcri.html>).

DAVID W. STEWART

Interquartile Range

DAVID CLARK-CARTER

Volume 2, pp. 941–941

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Interquartile Range

The interquartile range (IQR) is a measure of spread defined as the range of the middle 50% of scores in an ordered set of data, that is, the difference between the upper **quartile** (Q_3) and the lower quartile (Q_1). For example, if Q_3 were 25 and Q_1 were 5, then the IQR would be 20.

As with other measures that are based on quartiles, such as the median, the IQR has the advantage that it is little affected by the presence of extreme scores. In a **box plot**, it is represented by the length of the box and is sometimes referred to as the H-range [1] or the midspread.

As an example, assume data from a normal distribution on the number of arithmetic problems solved by 12 children. These data are 11 10 12 12 12 14 14 15 17 17 18 19, and have a mean of 14.17, and a standard deviation of 3.07. $Q_1 = 12$ and $Q_3 = 17$, so the interquartile range is 5.0.

If the data had been slightly more spread in the tails, we might have 6 8 12 12 12 14 14 15 17

19 21 23 Here the mean is still 14.17, the standard deviation has increased by nearly 2/3 to 5.0, and the interquartile range has increased only slightly to 6.0.

Finally, if we have one extreme score, giving us data values of 6 8 12 12 12 14 14 15 17 19 21 33, The mean will increase to 15.25 (the median would have remained constant), the standard deviation would increase to 7.0, while the interquartile range would remain at 6.0.

Dividing the IQR by 2 produces the semi-interquartile range (sometimes known as the quartile deviation). The semi-interquartile range is an appropriate measure of spread to quote when reporting medians.

Reference

- [1] Clark-Carter, D. (2004). *Quantitative Psychological Research: A Student's Handbook*, Psychology Press, Hove.

DAVID CLARK-CARTER

Interrupted Time Series Design

JOHN FERRON AND GIANNA RENDINA-GOBIOFF

Volume 2, pp. 941–945

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Interrupted Time Series Design

The interrupted time-series design [2] provides a method for researchers to examine the effect of an intervention on a single case, where the case may be a group or an individual. The basic interrupted time-series design is diagrammed in Figure 1, which shows there are multiple observations both prior to and after the intervention. A graph, showing an example of the kind of data obtained from an interrupted time-series design is also provided in Figure 1. The points represent observations, the solid lines are trend lines, and the dotted line is a projection from baseline. The difference between the dotted line and the post intervention solid line provides an estimate of treatment effect.

Discussion of the interrupted time-series design occurs both in the literature dealing with **quasi-experimental designs** [6] and the literature dealing with **single-case designs** [5]. In the single-case design literature, the basic interrupted time-series design is often referred to as an AB design, where the time prior to the intervention makes up the baseline phase (A) and the time after the intervention makes up the treatment phase (B). Although the form of the design remains the same, the applications used to characterize the design differ considerably between the quasi-experimental and single-case literatures.

Applications

Applications of interrupted time-series designs used to illustrate quasi-experiments often focus on archival

data that contain an aggregate value for a group across a large number of time points – for instance, research examining the frequency of behavior before and after a policy implementation. Specifically, one might be interested in the frequency of fatal motorcycle accidents before and after a mandatory helmet law was implemented. Another example would be to examine the frequency of behaviors before and after a major event. Specifically, one might be interested in teacher morale before and after the introduction of a new principal at their school.

Applications of interrupted time-series designs in the single-case literature tend to focus on the behavior of a single participant in a context where the researcher has some level of control over the measurement process, the observational context, and decisions about when to intervene – for example, research examining the effects of the introduction of some kind of treatment. In particular, one might be interested in measuring the aggressive behavior of a child before and after the introduction of therapy. Similarly, one might be interested in measuring the drug behavior of someone before and after treatment in a drug facility.

Types of Treatment Effects

Treatment effects can vary greatly from application to application in both their magnitude and their form. Figure 1 shows a treatment effect that remains constant over time, which is characterized by an immediate (or abrupt) upward shift in level but no change in slope. Figure 2 illustrates a variety of different treatment effects. Figures 2(a, c, e, and f), all show immediate effects, or discontinuities in the time series at the time of intervention.

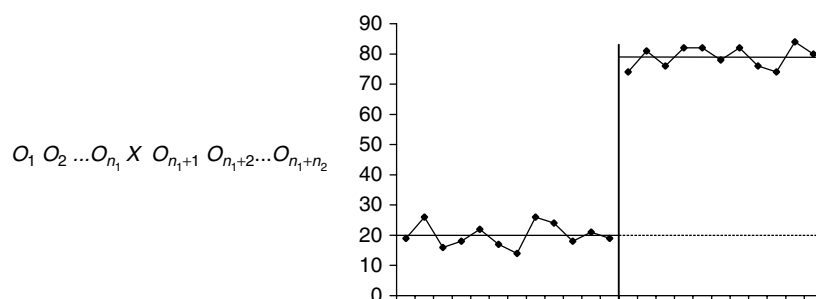


Figure 1 Diagram of an interrupted time-series design where O s represent observations and the X represents the intervention (left panel), and graphical display of interrupted time-series data (right panel)

2 Interrupted Time Series Design

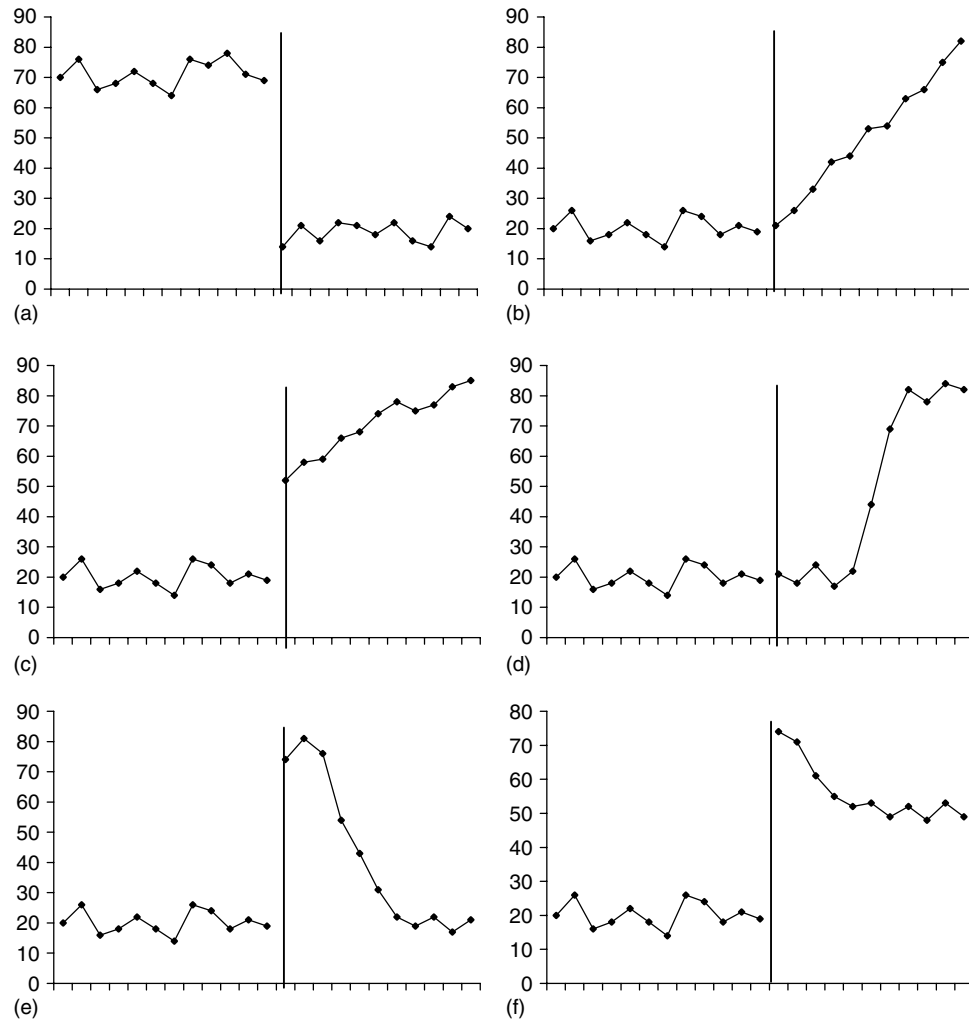


Figure 2 Illustration of different types of effects that may be found using an interrupted time-series design: immediate change in level (a), change in slope (b), a change in both level and slope (c), a delayed change in level (d), a transitory effect (e), and an effect that diminishes over time (f)

Figure 2(a), like Figure 1, shows a constant effect, while Figures 2(c, e, and f) show effects that change over time. In Figure 2(c) the effect becomes larger over time, whereas in Figures 2(e and f) the effect diminishes over time. In Figure 2(e), the effect may be referred to as transitory because there appears to be no effect by the end of the time series. For Figures 2(b and d), there is no immediate shift and thus it takes longer before a treatment effect is observed. In Figure 2(b) the treatment effect gradually increases over time, while in Figure 2(d) the treatment effect is delayed.

Design Issues

Number of Observations

The number of observations within a time series varies greatly across applications. While some studies have over 100 observations, others contain fewer than 10. Generally speaking, greater numbers of observations lead to stronger statements about treatment effects, and allow for more flexibility in analyzing the data. The number of observations needed, however, depends greatly on the application. A researcher

facing substantial variation around the trend lines will need a large number of observations to make a precise inference about the size of the treatment effect. In contrast, a researcher who encounters no variation in the baseline data may be able to infer a positive treatment effect with relatively few observations.

Baseline Variability

In the single-case literature, where the number of observations tends to be smaller, researchers often work to actively reduce the variability around their trend lines. For example, the researcher may notice a baseline variation that is associated with the time of day the observation is made, the particular activity being observed, and the people present during the observation. Under these circumstances, the researcher could reduce variation by moving toward greater commonality in the experimental conditions. There may also be inconsistencies in how the observations are rated, stemming from the use of different observers on different days or an observer who has ratings that tend to drift over time. In these circumstances, it may be possible to provide training to reduce the variability. In situations in which the researcher has control over the study, it is often possible to reduce variability, making it possible to make effect inferences from relatively short interrupted time-series studies. In situations in which the researchers do not have this much control, they may still work to identify and measure variables associated with the variability. This additional information can then be taken into account during the analysis.

Threats to Validity

When effect inferences are made, researchers often consider alternative explanations for why there may have been a change in the time series. Put another way, they consider threats to the validity of their inferences (*see* **Validity Theory and Applications; Internal Validity; External Validity**). It is possible, for example, that some other event occurred around the same time as the intervention and this other event is responsible for the change. It is also possible that a child happened to reach a new developmental stage, and this natural maturation led to the observed

changes in behavior; or that the instrumentation changed, leading to the observed change. Researchers may choose to add design elements to the basic interrupted time-series design to reduce the plausibility of alternative explanations. Doing so leads to more complex interrupted time-series designs.

More Complex Interrupted Time-series Designs

One option would be to remove the treatment from the participants, which is referred to as a **reversal design** (or withdrawal design). The simplest reversal design, which is diagrammed in Figure 3, has multiple observations during an initial baseline phase (A), multiple observations in the treatment phase (B), and then multiple observations in a second baseline phase (A). If the series reverses back to baseline levels when the treatment is removed, it becomes more obvious that the treatment led to the changes. Put another way, history and maturation effects become less plausible as explanations for the observed change. It is also possible to extend the ABA design to include more phases, creating more complex designs like the ABAB or the ABABAB design.

Another common extension of the interrupted time-series design is **the multiple baseline design**. In the multiple baseline design, a baseline phase (A) and treatment phase (B) are established for multiple participants, multiple behaviors, or multiple settings. The initiation of the treatment phases is staggered across time, creating baselines of different lengths for the different participants, behaviors, or settings. A diagram for the two-baseline version of the design and a corresponding graphical display are presented in Figure 3. As the number of baselines increases, it becomes less likely that history or maturational effects would stagger themselves across time in a manner that coincided with the staggered interventions.

A third variation, Figure 3(c), is obtained by including a comparison series. The comparison series could come from a control group that does not receive the treatment, or it could include data from a nonequivalent dependent variable. It is also possible to combine several of these additional elements. For example, one could combine elements from the reversal and multiple baseline design.

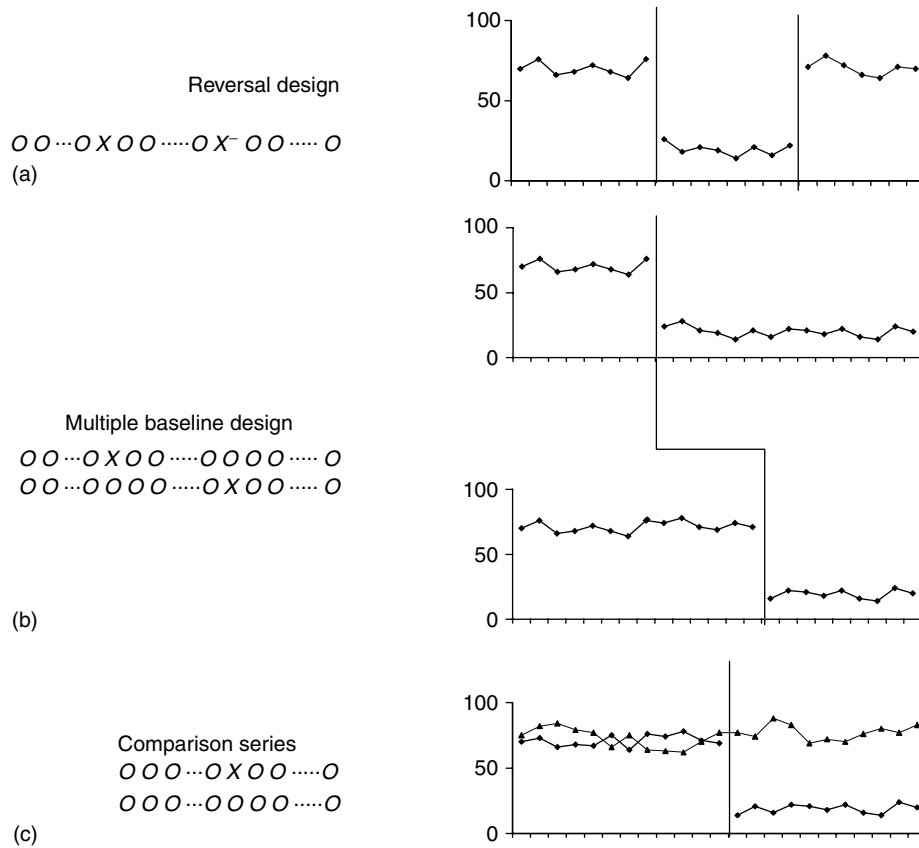


Figure 3 Design diagram and graphical display of interrupted time-series designs incorporating a withdrawal of the treatment (a), multiple baselines (b), and a comparison series (c). The O s represent observations, X represents the implementation of the intervention, and the X^- represents removal of the intervention

Analysis of Interrupted Time-series Data

Researchers typically start by visually analyzing a graph of the data. Additional analyses depend on the goals of the researcher, the variability in the data, and the amount of data available. Those interested in making inferences about the size and form of the treatment effect may consider selecting from a variety of statistical models for time-series data. Different models allow estimation of different types of treatment effects. For example, a model could contain one parameter for the level before treatment and one parameter for the change in level that occurs with the intervention. Such a model makes sense for a treatment effect that leads to an immediate shift in level, like the effect shown in Figure 1. As the effects become more complicated, additional

parameters need to be included to fully describe the treatment effect [4].

Statistical models for time-series data also differ in how the errors are modeled, where errors are the deviations of the observations from the trend lines that best fit the time-series data. The simplest model would result from assuming the errors were independent, which is what is assumed if one uses a standard regression model. There is reason, however, to suspect that the errors will often not be independent in time-series data. We often anticipate that unaccounted for factors may influence multiple consecutive observations. For example, a child's behavior may be affected by being sick, and being sick could be an unaccounted for factor that influenced observations across multiple days in the study. Under these circumstances, we expect that

errors that are closer in time will be more similar than errors further apart in time. Consequently, researchers often turn to statistical models that allow for dependencies, or autocorrelation, in the error structure [1].

Dependent error models differ in complexity. A researcher with relatively few observations will often have to assume a relatively simple model for the errors, while researchers with larger numbers of observations may use the data to help select from a wide range of possible models. Uncertainties that arise about the appropriateness of assuming a statistical model for the time-series data can lead to consideration of alternative approaches for analyzing the data.

Randomization tests [3] provide an alternative for researchers who wish to test the no-treatment-effect hypothesis without assuming a model for the data. The logic of randomization tests requires researchers to incorporate randomization in their designs. For example, they could randomly select the intervention point from among possible intervention points. A variety of randomization schemes and tests have been developed, and generally more statistical power is obtained for more complex designs.

References

- [1] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis, Forecasting, and Control*, Holden-Day, San Francisco.
- [2] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-experimental Designs for Research*, Rand McNally, Chicago.
- [3] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [4] Huitema, B.E. & McKean, J.W. (2000). Design specification issues in time-series intervention models, *Educational and Psychological Measurement* **60**, 38–58.
- [5] Kratochwill, T.R. (1978). Foundations of time series research, in *Single Subject Research: Strategies for Evaluating Change*, T.R. Kratochwill, ed., Academic Press, New York.
- [6] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

JOHN FERRON AND
 GIANNA RENDINA-GOBIOFF

Intervention Analysis

P.J. BROCKWELL

Volume 2, pp. 946–948

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Intervention Analysis

When a time series is recorded over an extended time interval $\{1, 2, \dots, n\}$, it frequently happens that, at some known time T in the interval, an event occurs that causes long-term or short-term changes in level of the series. For example, a change in the tax laws may have a long-term effect on stock returns, or the introduction of seat-belt legislation may have a long-term effect on the rate of occurrence of serious traffic injuries, while a severe storm in a particular city is likely to increase the number of insurance claims for a short period following the storm.

To account for the possible patterns of change in level of a time series $\{Y_t\}$, Box and Tiao [4] introduced the following model, in which it is assumed that the time T at which the change (or “intervention”) occurs is known:

$$Y_t = \sum_{j=0}^{\infty} \tau_j x_{t-j} + N_t, \quad (1)$$

where $\{N_t\}$ is the undisturbed time series (that can very often be modeled as an ARMA or ARIMA process), $\{\tau_j\}$ is a sequence of weights, and the “input” sequence $\{x_t\}$ is a deterministic sequence, usually representing a unit impulse at time T or possibly a sequence of impulses at times $T, T+1, T+2, \dots$, that is,

$$x_t = \begin{cases} 1, & \text{if } t = T, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

or

$$x_t = \begin{cases} 1, & \text{if } t \geq T, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

respectively. (Other deterministic input functions $\{x_t\}$ can also be used, representing for example impulses at times T_1 and T_2 .) By selecting an appropriate input sequence and a parametric family of weights, $\tau_j(\omega)$, $j = 0, 1, 2, \dots$, where ω is a finite-dimensional parameter vector, the problem of fitting the model (1) becomes that of estimating ω and finding an appropriate ARMA or ARIMA model for the series $\{N_t\}$ (see **Time Series Analysis**).

For example, if it is believed that the effect on the number of insurance claims following a major storm at time $T-1$ will be the addition of new claims at times $T, T+1, T+2, \dots$, with

the numbers declining geometrically as time goes by, then the intervention term $\sum_{j=0}^{\infty} \tau_j x_{t-j}$ with x_t defined by (2) and $\tau_j = c\theta^j$ for some $c > 0$ and $\theta \in (0, 1)$ is appropriate. The contribution of the intervention term at time $T+j$ is then equal to $c\theta^j$, $j = 0, 1, 2, \dots$ and the problem of fitting the intervention model (1) is that of estimating the parameters c and θ , while simultaneously finding a suitable model for $\{N_t\}$. The weights in this case are members of a two-parameter family indexed by c and θ .

By allowing the weights to belong to a larger parametric family than the two-parameter family of the preceding paragraph, it is possible to achieve an extremely wide range of possible intervention effects. Box and Tiao proposed the use of weights τ_j that are the coefficients of z^j in the power series expansion of the ratio of polynomials $z^b w(z)/v(z)$, where b is a nonnegative integer (*the delay parameter*), and the zeroes of the polynomial $v(z)$ in the complex plane all have absolute values greater than 1. The intervention term can then be expressed as

$$\sum_{j=0}^{\infty} \tau_j x_{t-j} = \sum_{j=0}^{\infty} \tau_j B^j x_t = \frac{B^b w(B)}{v(B)} x_t, \quad (4)$$

where B is the backward shift operator and $B^b w(B)/v(B)$ is referred to as the *transfer function* of the filter with weights τ_j , $j = 0, 1, 2, \dots$. The geometrically decreasing weights of the preceding paragraph correspond to the special case in which the ratio of polynomials is $c/(1 - \theta z)$. The intervention model (1) with intervention term (4) is very closely related to the transfer function model of Box and Jenkins [3]. In the transfer function model, the deterministic input sequence $\{x_t\}$ is replaced by an observed *random* input sequence. In intervention modeling, the input sequence is *chosen* to provide a parametric family of interventions of the type expected. Having selected the function $\{x_t\}$ and the parametric family for the weights $\{\tau_j\}$, the problem of fitting the model (1) reduces to a nonlinear regression problem in which the errors $\{N_t\}$ constitute an ARMA or ARIMA process whose parameters must also be estimated. Estimation is by minimization of the sums of squares of the estimated white noise sequence driving $\{N_t\}$. For details, see, for example, [3], [4], or [5].

The goals of intervention analysis are to estimate the effect of the intervention as indicated by the term $\sum_{j=0}^{\infty} \tau_j x_{t-j}$ and to use the resulting model (1)

2 Intervention Analysis

for forecasting. For example, in [9], intervention analysis was used to investigate the effect of the American Dental Association's endorsement of Crest toothpaste on Crest's market share. Other applications of intervention analysis can be found in [1], [2], and [3]. A more general approach can also be found in [6], [7], and [8].

Example Figure 1 shows the number of monthly deaths and serious injuries on UK roads for 10 years beginning in January 1975. (Taking one month as our unit of time we shall write these as $D_t, t = 1, \dots, 120$. The data are from [7].) Seat-belt legislation was introduced in the UK in February, 1983 ($t = 98$), in the hope of reducing the mean of the series from that time onwards. In order to assess the effect of the legislation, we can estimate the coefficient θ in the model,

$$D_t = \sum_{j=0}^{\infty} \theta x_{t-j} + W_t, \quad t = 1, \dots, 120, \quad (5)$$

where x_t is defined by (3) with $T = 98$ and $D_t = W_t$ is the model for the data prior to the intervention. The intervention term in (5) can be expressed more simply as θf_t where $f_t = 0$ for $t \leq 97$ and $f_t = 1$ for $t \geq 98$. Figure 1 clearly indicates (as expected) the presence of seasonality with period 12 in $\{W_t\}$. If, therefore, we apply the differencing operator $(1 - B^{12})$ to each side of (5), we might hope to obtain a model for the differenced data, $Y_t = D_t - D_{t-12}, t = 13, \dots, 120$, in which the noise $\{N_t = W_t - W_{t-12}\}$ is representable as an ARMA process. Carrying out

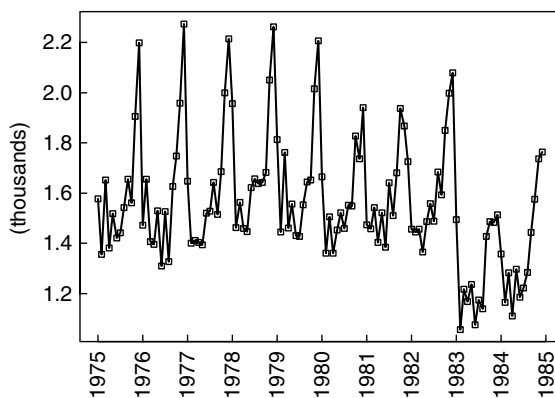


Figure 1 Monthly deaths and serious injuries on UK roads, January 1975–December 1984

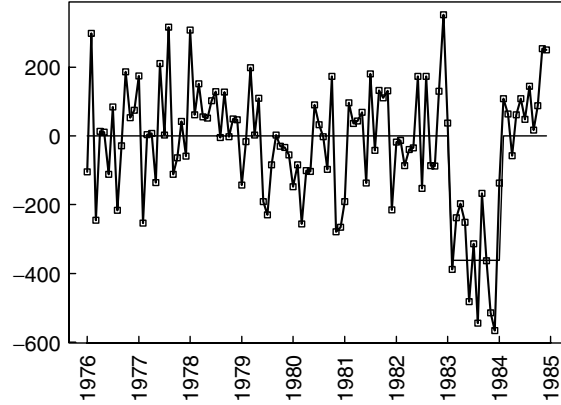


Figure 2 The differenced deaths and serious injuries on UK roads, showing the fitted intervention term

the differencing of the data, this does indeed appear to be the case. Our model for $\{Y_t\}$ is thus

$$Y_t = \theta \sum_{j=0}^{\infty} x'_{t-j} + N_t, \quad t = 13, \dots, 120, \quad (6)$$

where $\{x'_t\}$ is the differenced sequence $\{(1 - B^{12})x_t\}$, that is,

$$x'_t = \begin{cases} 1, & \text{if } 98 \leq t < 110, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The model (6) is now of the form (1) with $\{N_t\}$ representable as an ARMA process. For details of the least squares fitting of the model (6) to the differenced data, see [5], where it is found that the estimated value of θ is -362.5 (highly significantly different from zero) and a moving average model of order 12 is selected for $\{N_t\}$. Figure 2 shows the *differenced* data $\{Y_t\}$ together with the fitted intervention term in the model (6). The corresponding intervention term in the model (5) for the *original* data is a permanent level-shift of -362.5 .

References

- [1] Atkins, S.M. (1979). Case study on the use of intervention analysis applied to traffic accidents, *Journal of the Operational Research Society* **30**, 651–659.
- [2] Bhattacharya, M.N. & Layton, A.P. (1979). Effectiveness of seat belt legislation on the Queensland road toll – an Australian case study in intervention analysis, *Journal of the American Statistical Association* **74**, 596–603.

-
- [3] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis; Forecasting and Control*, 3rd Edition, Prentice Hall, Englewood Cliffs.
- [4] Box, G.P.E. & Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems, *Journal of the American Statistical Association* **70**, 70–79.
- [5] Brockwell, P.J. & Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, 2nd Edition, Springer-Verlag, New York.
- [6] Harvey, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
- [7] Pole, A., West, M. & Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*, Chapman & Hall, New York.
- [8] West, Mike. & Harrison, Jeff. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd Edition, Springer-Verlag, New York.
- [9] Wichern, D. & Jones, R.H. (1978). Assessing the impact of market disturbances using intervention analysis, *Management Science* **24**, 320–337.

P.J. BROCKWELL

Intraclass Correlation

ANDY P. FIELD

Volume 2, pp. 948–954

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Intraclass Correlation

Commonly used correlations such as the **Pearson product moment correlation** measure the bivariate relation between variables of different measurement classes. These are known as *interclass* correlations. By ‘different measurement classes’, we really just mean variables measuring different things. For example, we might look at the relation between attractiveness and career success; clearly one of these variables represents a class of measures of how good looking a person is, whereas the other represents the class of measurements of something quite different: how much someone achieves in their career. However, there are often cases in which it is interesting to look at relations between variables *within* classes of measurement. In its simplest form, we might compare only two variables. For example, we might be interested in whether anxiety runs in families, and we could look at this by measuring anxiety within pairs of twins (see [1]). In this case, the objects being measured are twins, and both twins are measured on some index of anxiety. As such, there is a pair of variables, which both measure anxiety and are, therefore, from the same class. In such cases, an intraclass correlation (ICC) is used and is commonly extended beyond just two variables to look at the consistency between judges. For example, in gymnastics, ice-skating, diving, and other Olympic sports, the contestant’s performance is often assessed by a panel of judges. There might be 10 judges, all of whom rate performances out of 10; therefore, the resulting measures are from the same class (they measure the same thing). The objects being rated are the competitors. This again is a perfect scenario for an intraclass correlation.

Models of Intraclass Correlations

There are a variety of different intraclass correlations (see [4] and [5]) and the first step in calculating one is to determine a model for your sample data. All of the various forms of the intraclass correlation are based on estimates of mean variability from a one-way **repeated measures Analysis of Variance (ANOVA)**.

All situations in which an intraclass correlation is desirable will involve multiple measures on different entities (be they twins, Olympic competitors, pictures,

sea slugs etc.). The objects measured constitute a random factor (see **Fixed and Random Effects**) in the design (they are assumed to be random exemplars of the population of objects). The measures taken can be included as factors in the design if they have a meaningful order, or can be excluded if they are unordered as we shall now see.

One-way Random Effects Model

In the simplest case, we might have only two measures (refer to our twin study on anxiety). When the order of these variables is irrelevant (for example, with our twin study it is arbitrary whether we treat the data from the first twin as being anxiety measure 1 or anxiety measure 2), the only systematic source of variation is the random variable representing the different objects. In this case, the only systematic source of variation is the random variable representing the different objects. As such, we can use a one-way ANOVA of the form:

$$x_{ij} = \mu + r_i + e_{ij}, \quad (1)$$

in which r_i is the effect of object i (known as the *row effects*), j is the measure being considered, and e_{ij} is an error term (the residual effects). The row and residual effects are random, independent, and normally distributed. Because the effect of the measure is ignored, the resulting intraclass correlation is based on the overall effect of the objects being measured (the mean between-object variability MS_{Rows}) and the mean within-object variability (MS_{W}). Both of these will be formally defined later.

Two-way Random Effects Model

When the order of measures is important, then the effect of the measures becomes important also. The most common case of this is when measures come from different judges or raters. Hodgins and Makarchuk [3], for example, show two such uses; in their study they took multiple measures of the same class of behavior (gambling) and also measures from different sources. They measured gambling both in terms of days spent gambling and money spent gambling. Clearly these measures generate different data so it is important to which measure a datum belongs (it is not arbitrary to which measure a datum

2 Intraclass Correlation

is assigned). This is one scenario in which a two-way model is used. However, they also took measures of gambling both from the gambler and a collateral (e.g., spouse). Again, it is important that we attribute data to the correct source. So, this is a second illustration of where a two-way model is useful. In such situations, the intraclass correlation can be used to check the consistency or agreement between measures or raters.

In this situation a two-way model can be used as follows:

$$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}, \quad (2)$$

where c_j is the effect of the measure (i.e., the effect of different raters, or different measures), and rc_{ij} is the interaction between the measures taken and the objects being measured. The effect of the measure (c_j) can be treated as either a fixed effect or a random effect. How it is treated does not affect the calculation of the intraclass correlation, but it does affect the interpretation (as we shall see). It is also possible to exclude the interaction term and use the model:

$$x_{ij} = \mu + r_i + c_j + e_{ij}. \quad (3)$$

We shall now turn our attention to calculating the sources of variance needed to calculate the intraclass correlation.

Sources of Variance: An Example

Field [2] uses an example relating to student concerns about the consistency of marking between lecturers. It is common that lecturers obtain reputations for being ‘hard’ or ‘light’ markers, which can lead students to believe that their marks are not based solely on the

intrinsic merit of the work, but can be influenced by who marked the work. To test this, we could calculate an intraclass correlation. First, we could submit the same eight essays to four different lecturers and record the mark they gave each essay. Table 1 shows the data, and you should note that it looks the same as a one-way repeated measures ANOVA in which the four lecturers represent four levels of an ‘independent variable’, and the outcome or dependent variable is the mark given (in fact, these data are used as an example of a one-way repeated measures ANOVA).

Three different sources of variance are needed to calculate an intraclass correlation. These sources of variance are the same as those calculated in one-way repeated measures ANOVA (see [2] for the identical set of calculations!).

The Between-object Variance (MS_{Rows})

The first source of variance is the variance between the objects being rated (in this case the between-essay variance). Essays will naturally vary in their quality for all sorts of reasons (the natural ability of the author, the time spent writing the essay, etc.). This variance is calculated by looking at the average mark for each essay and seeing how much it deviates from the average mark for all essays. These deviations are squared because some will be positive and others negative, and so would cancel out when summed. The squared errors for each essay are weighted by the number of values that contribute to the mean (in this case, the number of different markers, k). So, in general terms we write this as:

$$SS_{\text{Rows}} = \sum_{i=1}^n k_i (\bar{X}_{\text{Row } i} - \bar{X}_{\text{all rows}})^2. \quad (4)$$

Table 1 Marks on eight essays by four lecturers

Essay	Dr Field	Dr Smith	Dr Scrote	Dr Death	Mean	S^2	$S^2(k-1)$
1	62	58	63	64	61.75	6.92	20.75
2	63	60	68	65	64.00	11.33	34.00
3	65	61	72	65	65.75	20.92	62.75
4	68	64	58	61	62.75	18.25	54.75
5	69	65	54	59	61.75	43.58	130.75
6	71	67	65	50	63.25	84.25	252.75
7	78	66	67	50	65.25	132.92	398.75
8	75	73	75	45	67.00	216.00	648.00
Mean:	68.88	64.25	65.25	57.38	63.94	Total:	1602.50

Or, for our example we could write it as:

$$SS_{\text{Essays}} = \sum_{i=1}^n k_i (\bar{X}_{\text{Essay } i} - \bar{X}_{\text{all essays}})^2. \quad (5)$$

This would give us:

$$\begin{aligned} SS_{\text{Rows}} &= 4(61.75 - 63.94)^2 + 4(64.00 - 63.94)^2 \\ &\quad + 4(65.75 - 63.94)^2 + 4(62.75 - 63.94)^2 \\ &\quad + 4(61.75 - 63.94)^2 \\ &\quad + 4(63.25 - 63.94)^2 + 4(65.25 - 63.94)^2 \\ &\quad + 4(67.00 - 63.94)^2 \\ &= 19.18 + 0.014 + 13.10 + 5.66 \\ &\quad + 19.18 + 1.90 + 6.86 + 37.45 \\ &= 103.34. \end{aligned} \quad (6)$$

This sum of squares is based on the total variability and so its size depends on how many objects (essays in this case) have been rated. Therefore, we convert this total to an average known as the *mean squared error (MS)* by dividing by the number of essays (or in general terms the number of rows) minus 1. This value is known as the *degrees of freedom*.

$$MS_{\text{Rows}} = \frac{SS_{\text{Rows}}}{df_{\text{Rows}}} = \frac{103.34}{n-1} = \frac{103.34}{7} = 14.76. \quad (7)$$

The mean squared error for the rows in Table 1 is our estimate of the natural variability between the objects being rated.

The Within-judge Variability (MS_W)

The second variability in which we are interested is the variability within measures/judges. To calculate this, we look at the deviation of each judge from the average of all judges on a particular essay. We use an equation with the same structure as before, but for each essay separately:

$$SS_{\text{Essay}} = \sum_{k=1}^p (\bar{X}_{\text{Column } k} - \bar{X}_{\text{all columns}})^2. \quad (8)$$

For essay 1, for example, this would be:

$$\begin{aligned} SS_{\text{Essay}} &= (62 - 61.75)^2 + (58 - 61.75)^2 \\ &\quad + (63 - 61.75)^2 + (64 - 61.75)^2 = 20.75. \end{aligned} \quad (9)$$

The degrees of freedom for this calculation is again one less than the number of scores used in the calculation. In other words, it is the number of judges, k , minus 1.

We calculate this for each of the essays in turn and then add these values up to get the total variability within judges. An alternative way to do this is to use the variance within each essay. The equation mentioned above is equivalent to the variance for each essay multiplied by the number of values on which that variance is based (in this case the number of Judges, k) minus 1. As such we get:

$$\begin{aligned} SS_W &= s_{\text{essay1}}^2(k_1 - 1) + s_{\text{essay2}}^2(k_2 - 1) \\ &\quad + s_{\text{essay3}}^2(k_3 - 1) + \dots + s_{\text{essayn}}^2(k_n - 1). \end{aligned} \quad (10)$$

Table 1 shows the values for each essay in the last column. When we sum these values we get 1602.50. As before, this value is a total and so depends on the number essays (and the number of judges). Therefore, we convert it to an average by dividing by the degrees of freedom. For each essay, we calculated a sum of squares that we saw was based on $k - 1$ degrees of freedom. Therefore, the degrees of freedom for the total within-judge variability are the sum of the degrees of freedom for each essay $df_W = n(k - 1)$, where n is the number of essays and k is the number of judges. In this case, it will be $8(4 - 1) = 24$.

The resulting mean squared error is, therefore:

$$MS_W = \frac{SS_W}{df_W} = \frac{1602.50}{n(k-1)} = \frac{1602.50}{24} = 66.77. \quad (11)$$

The Between-judge Variability (MS_{Columns})

The within-judge or within-measure variability is made up of two components. The first is the variability created by *differences* between judges. The second is the unexplained variability (error for want of a better word). The variability between judges is again calculated using a variant of the same equation that we have used all along, only this time we are interested in the deviation of each judge's mean from the mean of all judges:

$$SS_{\text{Columns}} = \sum_{k=1}^p n_i (\bar{X}_{\text{Column } i} - \bar{X}_{\text{all columns}})^2 \quad (12)$$

4 Intraclass Correlation

or

$$SS_{\text{Judges}} = \sum_{k=1}^p n_i (\bar{X}_{\text{Judge } i} - \bar{X}_{\text{all Judges}})^2, \quad (13)$$

where n is the number of things that each judge rated. For our data we would get:

$$\begin{aligned} SS_{\text{Columns}} &= 8(68.88 - 63.94)^2 + 8(64.25 - 63.94)^2 \\ &\quad + 8(65.25 - 63.94)^2 + 8(57.38 - 63.94)^2 \\ &= 554. \end{aligned} \quad (14)$$

The degrees of freedom for this effect are the number of judges, k , minus 1. As before, the sum of squares is converted to a mean squared error by dividing by the degrees of freedom:

$$MS_{\text{Columns}} = \frac{SS_{\text{Columns}}}{df_{\text{Columns}}} = \frac{554}{k-1} = \frac{554}{3} = 184.67. \quad (15)$$

The Error Variability (MS_E)

The final variability is the variability that cannot be explained by known factors such as variability between essays or judges/measures. This can be easily calculated using subtraction because we know that the within-judges variability is made up of the between-judges variability and this error:

$$\begin{aligned} SS_W &= SS_{\text{Columns}} + SS_E \\ SS_E &= SS_W - SS_{\text{Columns}}. \end{aligned} \quad (16)$$

The same is true of the degrees of freedom:

$$\begin{aligned} df_W &= df_{\text{Columns}} + df_E \\ df_E &= df_W - df_{\text{Columns}}. \end{aligned} \quad (17)$$

So, for these data we obtain:

$$\begin{aligned} SS_E &= SS_W - SS_{\text{Columns}} \\ &= 1602.50 - 554 \\ &= 1048.50 \end{aligned} \quad (18)$$

and

$$\begin{aligned} df_E &= df_W - df_{\text{Columns}} \\ &= 24 - 3 \\ &= 21. \end{aligned} \quad (19)$$

The average error variance is obtained in the usual way:

$$MS_E = \frac{SS_E}{df_E} = \frac{1048.50}{21} = 49.93. \quad (20)$$

Calculating Intraclass Correlations

Having computed the necessary variance components, we shall now look at how the ICC is calculated. Before we do so, however, there are two important decisions to be made.

Single Measures or Average Measures?

So far we have talked about situations in which the measures we have used produce single values. However, it is possible that we might have measures that produce an average score. For example, we might get judges to rate paintings in a competition on the basis of style, content, originality, and technical skill. For each judge, their ratings are averaged. The end result is still the ratings from a set of judges, but these ratings are an average of many ratings. Intraclass correlations can be computed for such data, but the computation is somewhat different.

Consistency or Agreement?

The next decision involves whether we want a measure of overall consistency between measures/judges. The best way to explain this distinction is to return to our example of lecturers and essay marking. It is possible that particular lecturers are harsh (or lenient) in their ratings. A consistency definition views these differences as an irrelevant source of variance. As such the between-judge variability described above (MS_{Columns}) is ignored in the calculation (see Table 2). In ignoring this source of variance, we are getting a measure of whether judges agree about the relative merits of the essays without worrying about whether the judges anchor their marks around the same point. So, if all the judges agree that essay 1 is the best and essay 5 is the worst (or their rank order of essays is roughly the same), then agreement will be high: it does not matter that Dr. Field's marks are all 10% higher than Dr. Death's. This is a consistency definition of agreement.

Table 2 Intraclass correlation (ICC) equations and calculations

Model	Interpretation	Equation	ICC for example data
<i>ICC for Single Scores</i>			
One-way	Absolute agreement	$\frac{MS_R - MS_W}{MS_R + (k - 1)MS_W}$	$\frac{14.76 - 66.77}{14.76 + (4 - 1)66.77} = -0.24$
Two-way	Consistency	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E}$	$\frac{14.76 - 49.93}{14.76 + (4 - 1)49.93} = -0.21$
	Absolute agreement	$\frac{MS_R - MS_E}{MS_R + (k - 1)MS_E + \frac{k}{n}(MS_C - MS_E)}$	$\frac{14.76 - 49.93}{14.76 + (4 - 1)49.93 + \frac{4}{8}(184.67 - 49.93)} = -0.15$
<i>ICC for Average Scores</i>			
One-way	Absolute agreement	$\frac{MS_R - MS_W}{MS_R}$	$\frac{14.76 - 66.77}{14.76} = -3.52$
Two-way	Consistency	$\frac{MS_R - MS_E}{MS_R}$	$\frac{14.76 - 49.93}{14.76} = -2.38$
	Absolute agreement	$\frac{MS_R - MS_E}{MS_R + (MS_C - MS_E/n)}$	$\frac{14.76 - 49.93}{14.76 + (184.67 - 49.93/8)} = -1.11$

The alternative is to treat relative differences between judges as an important source of disagreement. That is, the between-judge variability described above (MS_{Columns}) is treated as an important source of variation and is included in the calculation (see Table 2). In this scenario, disagreements between the relative magnitude of judge's ratings matters (so, the fact that Dr Death's marks differ from Dr Field's will matter even if their rank order of marks is in agreement). This is an absolute agreement definition. By definition, the one-way model ignores the effect of the measures and so can have only this kind of interpretation.

Equations for ICCs

Table 2 shows the equations for calculating ICC on the basis of whether a one-way or two-way model is assumed and whether a consistency or absolute agreement definition is preferred. For illustrative purposes, the ICC is calculated in each case for the example used in this article. This should enable the reader to identify how to calculate the various sources of variance. In this table, MS_{Columns} is abbreviated to MS_C , and MS_{Rows} is abbreviated to MS_R .

Significance Testing

The calculated intraclass correlation can be tested against a value under the null hypothesis using a

standard F test (see **Analysis of Variance**). McGraw and Wong [4] describe these tests for the various intraclass correlations we have discussed; Table 3 summarizes their work. In this table, ICC is the observed intraclass correlation whereas ρ_0 is the value of the intraclass correlation under the null hypothesis. That is, it is the value against which we wish to compare the observed intraclass correlation. So, replace this value with 0 to test the hypothesis that the observed ICC is greater than zero, but replace it with other values such as 0.1, 0.3, or 0.5 to test that the observed ICC is greater than known values of small, medium, and large-effect sizes respectively.

Fixed versus Random Effects

I mentioned earlier that the effect of the measure/judges can be conceptualized as a fixed or random effect. Although it makes no difference to the calculation, it does affect the interpretation. Essentially, this variable should be regarded as random when the judges or measures represent a sample of a larger population of measures or judges that could have been used. In other words, the particular judges or measures chosen are not important and do not change the research question that is being addressed. However, the effect of measures should be treated as fixed when changing one of

6 Intraclass Correlation

Table 3 Significance test for intraclass correlations (Adapted from McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients, *Psychological Methods* **1**(1), 30–46.)

Model	Interpretation	<i>F</i> -ratio	<i>Df</i> 1	<i>Df</i> 2
<i>ICC for Single Scores</i>				
One-way	Absolute agreement	$\frac{MS_R}{MS_W} \times \frac{1 - \rho_0}{1 + (k - 1)\rho_0}$	$n - 1$	$n(k - 1)$
Two-way	Consistency	$\frac{MS_R}{MS_E} \times \frac{1 - \rho_0}{1 + (k - 1)\rho_0}$	$n - 1$	$(n - 1)(k - 1)$
	Absolute agreement	$\frac{MS_R}{aMS_C + bMS_E}$		
	In which;		$n - 1$	$\frac{(aMS_C + bMS_E)^2}{\frac{(aMS_C)^2}{k - 1} + \frac{(bMS_E)^2}{(n - 1)(k - 1)}}$
	$a = \frac{k\rho_0}{n(1 - \rho_0)}$			
	$b = 1 + \frac{k\rho_0(n - 1)}{n(1 - \rho_0)}$			
<i>ICC for Average Scores</i>				
One-way	Absolute agreement	$\frac{1 - \rho_0}{1 - ICC}$	$n - 1$	$n(k - 1)$
Two-way	Consistency	$\frac{1 - \rho_0}{1 - ICC}$	$n - 1$	$(n - 1)(k - 1)$
	Absolute agreement	$\frac{MS_R}{cMS_C + dMS_E}$		
	In which;		$n - 1$	$\frac{(cMS_C + dMS_E)^2}{\frac{(cMS_C)^2}{k - 1} + \frac{(dMS_E)^2}{(n - 1)(k - 1)}}$
	$c = \frac{\rho_0}{n(1 - \rho_0)}$			
	$b = 1 + \frac{\rho_0(n - 1)}{n(1 - \rho_0)}$			

the judges or measures would significantly affect the research question (see **Fixed and Random Effects**). For example, in the gambling study mentioned earlier it would make a difference if the ratings of the gambler were replaced: the fact the gamblers gave ratings was intrinsic to the research question being addressed (i.e., do gamblers give accurate information about their gambling?). However, in our example of lecturers' marks, it should not make any difference if we substitute one lecturer with a different one: we can still answer the same research question (i.e., do lecturers, in general, give inconsistent marks?). In terms of interpretation, when the effect of the measures is a random factor then the

results can be generalized beyond the sample; however, when they are a fixed effect, any conclusions apply only to the sample on which the ICC is based (see [4]).

References

- [1] Eley, T.C. & Stevenson, J. (1999). Using genetic analyses to clarify the distinction between depressive and anxious symptoms in children, *Journal of Abnormal Child Psychology* **27**(2), 105–114.
- [2] Field, A.P. (2005). *Discovering Statistics Using SPSS*, 2nd Edition, Sage, London.

- [3] Hodgins, D.C. & Makarchuk, K. (2003). Trusting problem gamblers: reliability and validity of self-reported gambling behavior, *Psychology of Addictive Behaviors* **17**(3), 244–248.
- [4] McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients, *Psychological Methods* **1**(1), 30–46.
- [5] Shrout, P.E.F. & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing reliability, *Psychological Bulletin* **86**, 420–428.

ANDY P. FIELD

Intrinsic Linearity

L. JANE GOLDSMITH AND VANCE W. BERGER

Volume 2, pp. 954–955

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Intrinsic Linearity

When a response variable is related to one or more predictor variables in a nonlinear model (see **Nonlinear Models**), parameter estimation must be accomplished through a complicated, iterative process. In some cases, a transformation can be used to change the model to a linear one, thereby permitting analysis using the highly developed, efficient methods of linear models (see **Least Squares Estimation; Multiple Linear Regression**). In the classic linear model $Y = X\beta + \varepsilon$, parameter estimates are found through the familiar closed-form solution

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (1)$$

Some examples and the appropriate linearizing transformations are given below.

- (1) $Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$ is a nonlinear model, because it is not linear in the parameters. However, one can apply the natural logarithm (ln) transformation to both sides of the equation to obtain $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, which is a linear model with $\ln Y$ as the new response variable. One could say, then, that the model was intrinsically linear, but it appeared nonlinear because of an unfortunate choice of the response variable Y . Using $\ln(Y)$ instead makes all the difference.
- (2) $Y = \theta_1 X / (\theta_2 + X)$, the so-called Michaelis–Menten model, is also a nonlinear model. One can apply the reciprocal transformation to both sides of the equation to obtain

$$\frac{1}{Y} = \frac{\theta_2 + X}{\theta_1 X} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1} \frac{1}{X}. \quad (2)$$

For clarity, this can be seen to be the form of a linear model $Z = \beta_0 + \beta_1 V$, with $Z = 1/Y$,

$$\beta_0 = \frac{1}{\theta_1}, \quad \beta_1 = \frac{\theta_2}{\theta_1}, \quad \text{and } V = \frac{1}{X}. \quad (3)$$

- (3) $Y = \alpha X^\beta e^{-\nu X}$. Take natural logs of both sides of the equation to obtain

$$\ln Y = \ln \alpha + \beta \ln X - \nu X. \quad (4)$$

This is of the form $Z = \beta_0 + \beta_1 V_1 + \beta_2 V_2$, with $Z = \ln Y$, $\beta_0 = \ln \alpha$, $\beta_1 = \beta$, $V_1 = \ln X$, $\beta_2 = -\nu$, and $V_2 = X$.

The Error Term

The above examples demonstrate how transformations can change nonlinear models into models of linear form. An important *caveat*, however, is that for the correct application of linear estimation theory, the random error term ε must appear as an additive term in the resulting (transformed) linear model and exhibit what may be called *normal sphericity*; that is, the errors for different observations in the linear model must be independent and identically distributed as normal variates from a Gaussian distribution of mean 0 and variance σ^2 . More succinctly, $\varepsilon \sim N(0, \sigma^2)$.

Thus, for example, in the original nonlinear model (1) above, the error should appear in the exponent: $Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$, as opposed to being added on to the exponentiated expression so that after the ln transformation, the error appears properly in the linear model $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Furthermore, of course, the error term must have the normal sphericity properties. Similarly, for model (3), ε should appear in the exponent $Y = \alpha X^\beta e^{-\nu X + \varepsilon}$ of the original model, suitably distributed, for the transformed linear model to be a candidate for application of linear model least squares theory for analysis. Note, however, that needing the error term to satisfy a certain condition does not constitute a valid reason for assuming this to be the case. The error term actually does satisfy this required condition, or it does not. If it does not, then the model that assumes it does should not be used. The more realistic model should be used, even if this means that it will remain nonlinear even after the transformation.

For example, Model (2), above, the Michaelis–Menten model, usually does not exhibit the correct error structure after transformation and is usually analyzed as a nonlinear model to avoid bias in the analysis, even though the reciprocal transformation yields a tempting linear form. Standard regression residual analysis, including scatterplots and quartile plots, can be used to substantiate appropriate error term behavior in transformed linear models. There are some excellent texts dealing with nonlinear regression models, including material on transforming them to linear models [1–4].

References

- [1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

2 **Intrinsic Linearity**

- [2] Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition, Wiley, New York.
- [3] Ratkowsky, D.A. (1990). *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York.
- [4] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*, Wiley, New York.

L. JANE GOLDSMITH AND VANCE W. BERGER

INUS Conditions

LEON HORSTEN AND ERIK WEBER

Volume 2, pp. 955–958

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

INUS Conditions

The term 'INUS condition' was introduced in 1965 by philosopher John Mackie, in the context of his attempt to provide a logical analysis of the relation of causality in terms of necessary and sufficient conditions [5, 6]. His theory of causation can be seen as a refinement of the theories of D. Hume and of J.S. Mill.

We will start with a rough and provisional statement of Mackie's theory of causality, which we will then look at in some detail. In the course of this, we will have occasion to slightly modify and qualify the rough provisional explication with which we have started until we reach a more detailed and more accurate account. To conclude, the notion of an INUS condition will be related to the methodology of the behavioral sciences.

A Rough Sketch of Mackie's Theory of Causality

C is a cause of E if and only if

1. C and E are both actual
2. C occurs before E
3. C is an INUS condition of E

Here, the notion of an INUS condition is spelled out as follows:

C is an INUS condition of E if and only if C is an **I**nsufficient but **N**onredundant part of a Condition, which is itself **U**nnecessary but exclusively **S**ufficient for E in the circumstances.

The notion of INUS condition is thus the central component of Mackie's theory of causation. We will now explain Mackie's theory of causation in some more detail.

We start with an illustration of what it means for one event to cause another. If we say that a particular flash of lightning was a cause of a forest fire, then both the lightning flash and the forest fire must have occurred, and the occurrence of the lightning flash must have temporally preceded the occurrence of the forest fire. The particular lightning flash is a component of a complex condition, which was, in the circumstances, sufficient for the forest fire to start. This complex condition might, for instance, also contain the dryness of the grass and the fact that when

lightning struck, there was no fire truck driving by. So the lightning flash was not by itself sufficient to start the fire, but conjoined with these other components, it was. Now this complex condition (lightning flash, dry grass, no fire truck) need not be a necessary condition for the forest fire to start. For instance, if there was no lightning flash, but someone had thrown a lighted match in the grass, that might also have set the forest on fire.

This gives us a way to depict the logical structure of the relation of causality. The *full cause* C of an effect E is a disjunction of complex conditions $C_1, \dots, C_n$, where each C_i is a conjunction of positive and negative conditions $(\neg)p_1^i, \dots, (\neg)p_k^i$. So the logical form of a typical full cause C of an effect E could be:

$$(A \wedge B \wedge \neg D) \vee (G \wedge B \wedge \neg D)$$

We can apply this to the example above by setting

1. A = lightning flash
2. B = grass is dry
3. D = there is a fire truck passing
4. G = a lighted match is thrown

The disjuncts of the full cause must contain no nonredundant components. The individual conjuncts of the disjuncts out of which the full cause C of E consists are called the *INUS conditions* of E. So in our example, $\{A, B, \neg D, G\}$ is the collection of INUS conditions of E. The INUS conditions of E can be seen as *possible causes* of E in the *ordinary sense of the word*, for ordinarily, when asked what caused E, we do not give what Mackie calls the *full cause* of E. The actual causes in the ordinary sense of the word are the components of the disjuncts of the full cause that actually obtain. Suppose the disjunct $(A \wedge B \wedge \neg D)$ has occurred. What are called the *actual causes* of E are then the components A, B, $\neg D$ of the disjunct of C that has obtained.

This is the approximate core of Mackie's theory of causality.

Qualifications and Complications

First, Mackie stresses that the account given up to now needs to be put into the perspective of a background context which is assumed to obtain, and which Mackie calls the *causal field* F. To come back

to our example, without oxygen, the forest would not have burned. But it would be strange to call the presence of oxygen a cause of the forest fire, since forests in fact *always* contain oxygen. So it may be taken to be a part of the causal field that the part of the atmosphere where the forest is located contains enough oxygen to allow the burning of the forest. If we take this into account, then the logical form of our example of causality will be:

$$F \rightarrow [((A \wedge B \wedge \neg D) \vee (G \wedge B \wedge \neg D)) \leftrightarrow E] \quad (1)$$

Second, it needs to be stressed that $\rightarrow$ and $\leftrightarrow$ in (1) are not merely material implications, but must be assumed to have *counterfactual strength*. When $\rightarrow$ and $\leftrightarrow$ are read as material implications, (1) only makes a claim about the actual world. But the implicative relations expressed by the formula above are intended to hold also in counterfactual situations. So formula (1) implies, for instance, that in a counterfactual situation in which not the first but the second disjunct $G \wedge B \wedge \neg D$ obtains, E must also obtain.

Let us pause here for a moment to notice a consequence of this. The fact that (1) is assumed to hold in *all* counterfactual circumstances implies that irreducibly *indeterministic causation* cannot be modeled by Mackie's theory, for an indeterministic cause is not required to produce the effect in all counterfactual circumstances (see [7, Chapter 7]). Mackie's theory can, therefore, be no more than a theory of deterministic causation. Also, Mackie's account presupposes that we have a theoretical grip on counterfactual conditionals. However, it has proved to be difficult to construct a satisfactory theory of the logic of counterfactual conditionals. A description and defense of one of the leading theories of counterfactuals is given in [3, 4].

Third, there is a *fourth* condition, which, in addition to the three conditions mentioned in the provisional description of Mackie's account, has to be satisfied for C to be a cause of E. Mackie requires that the disjunct of the full cause of E to which C belongs is the *only* disjunct which is true. The reason for this requirement is that Mackie wants to exclude the possibility of *causal overdetermination*. To come back to our example, imagine the situation in which *both* lightning struck the grass (A) *and* a lighted match was thrown in the grass (G) (and the grass was dry and no fire truck was driving by). In this imagined situation, the fourth condition that we have just sketched is violated. Mackie would say

that *neither A nor G* would count as a cause of E. Many have found this an unwelcome consequence of Mackie's theory. They would say that in the imagined situation, *both* the lighted match *and* the lightning flash are causes of the forest fire. In other words, many authors find causal overdetermination a genuine possibility. (For this reason, we have not included this requirement in the provisional account.) See, for instance, [8].

Fourth, in the provisional statement of Mackie's explication of the causal relation between C and E, we required C to temporarily precede E. Mackie objects to this requirement as we have formulated it since it rules out by fiat the possibility that a cause is simultaneous with its effect, and a fortiori that an effect precedes its cause (*backwards causation*). Mackie wants to leave simultaneous and backwards causation open as a possibility. For this reason, he tries to ensure the asymmetry of the relation of causality by a weaker requirement. For C to cause E, Mackie requires that C is *causally prior* to E. And C is causally prior to E if C is fixed earlier than E is fixed, where a condition C is fixed at a moment t if it is *determined* at t that C will occur. This leaves open the possibility that a condition C at t_2 causes a condition E at t_1 even if $t_1 < t_2$. For, it is possible that it is fixed (or determined) that C will occur *before* it is fixed that E will occur. In this way, the requirement of temporal precedence must be replaced by Mackie's more complicated requirement of causal priority. Note, however, that even in the amended account of Mackie's theory, it is *some* relation of temporal precedence that ensures the asymmetry of causality. A good reference on problems related to backward causation is [2].

Mackie's Theory and Experimental Methods

In most situations, the full cause of an event E that has happened is not known. What we usually have is a list of conditions that may be causally relevant, and the information that one (or some) of them, call this condition A, has occurred immediately before event E. From this information, we can, using Mill's *method of difference*, try to derive that A must be an actual cause of E. Suppose that we have a list $A_1, \dots, A_n$ of factors that are thought to be causally relevant. Then we can investigate which

factors are present whenever the event E occurs and which factors are absent whenever E occurs. This information can then be gathered into disjunction of conjunctions of conditions, which may give us the full cause of E. Even if it does not, this information may tell us that A is an INUS condition of E. And, even if that is not the case, the information obtained from the method of difference may strengthen our belief that A is an INUS condition of E.

Mackie's theory can also be related to statistical testing procedures (see **Classical Statistical Inference: Practice versus Presentation**). Suppose we do the following experiment (see also [1, p. 210 ff]). A random sample of laboratory rats is randomly divided into an experimental group and a control group. The experimental group is put on a diet that contains 5% saccharin. The control group receives the same diet, minus the saccharin. When a rat dies, we check whether it has developed bladder cancer. After two years, the remaining rats are killed and examined. We observe that the fraction of animals with bladder cancer is higher in the experimental group. The difference is statistically significant, so we conclude that saccharin consumption is positively causally relevant for bladder cancer in populations of laboratory rats. In other words, we conclude that, in a population where all rats would consume saccharin (this hypothetical population is represented in the experiment by the experimental group), there would be more bladder cancers than in a population without saccharin consumption.

If we claim that smoking is positively causally relevant for lung cancer in the Belgian population, we make a similar counterfactual claim. In Belgium, some people smoke, others do not. Positive causal relevance means that, if every Belgian would smoke (while their other characteristics remained the same), there would be more instances of lung cancer than in the hypothetical population where no Belgian smokes (while their other characteristics are preserved). INUS conditions provide a framework for interpreting such claims about causal relevance in populations. Since not all rats that consume saccharin develop bladder cancer, it is plausible to assume that there is at least one set of properties X such that all rats which have X and consume saccharin develop bladder cancer. In fact, there may be several such properties, which we denote as $X_1, \dots, X_m$. This means that experiments like the one described above can be

seen as methods for establishing that something (e.g., saccharin consumption) is an INUS condition without having to spell out what the other components of the conjunct are. We show that there is at least one X such that $X \wedge C$ is a sufficient cause of E, but we do not know (and do not need to know) what this X is.

We have been cautious in our formulations in the previous paragraph ('it is plausible' and 'can be seen'). The reason for this is that interpreting causal hypotheses about populations in the indicated way assumes that the individuals that constitute the populations are deterministic systems. It is equally plausible to assume that there is also indeterminism at the individual level. But as was mentioned above, Mackie's theory of causation is a conceptual framework for developing deterministic interpretations of experimental results. This means that it is incomplete: we need a complementary framework for developing genuinely indeterministic interpretations.

Finally, a brief remark on the notion of a causal field. The causal hypotheses we considered were hypotheses about populations. If we interpret the claim along the lines of Mackie, the characteristics of the population (e.g., the properties that all rats have in common) constitute the causal field. So also the notion of a causal field has its counterpart in experimental practice.

References

- [1] Giere, R. (1997). *Understanding Scientific Reasoning*, Harcourt Brace College Publishers, Fort Worth.
- [2] Hausman, D. *Causal Asymmetries*, Cambridge University Press, Cambridge.
- [3] Lewis, D. (1973). *Counterfactuals*, Basil Blackwell.
- [4] Lewis, D. Causation, *Journal of Philosophy* **70**, 556–567. p.
- [5] Mackie, J. (1965). Causes and conditions, *American Philosophical Quarterly* **2/4**, 245–255 and 261–264.
- [6] Mackie, J. (1974). *The Cement of the Universe. A Study of Causation*, Clarendon Press, Oxford.
- [7] Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton.
- [8] Scriven, M. (1966). Defects of the necessary condition analysis of causation, in *Philosophical Analysis and History*, W. Dray, ed., Harper and Row, New York, pp. 258–262.

LEON HORSTEN AND ERIK WEBER

Item Analysis

RICHARD M. LUECHT

Volume 2, pp. 958–967

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Analysis

Item analysis is a fundamental psychometric activity that impacts both test construction and scoring.¹ Within the test construction process, it supplements the qualitative evaluation of item content and presentation format that item writers and test editors carry out, providing a focused, quantitative analysis of every test item to determine whether each is at an appropriate level of difficulty and exhibits adequate information to distinguish among examinees at different proficiency levels. As part of the test scoring process, item analysis can statistically verify the correctness of answer keys and scoring rules, aid in identifying flawed or ambiguous items, and help to determine if any items are potentially biased or otherwise problematic for certain subgroups in the population.

This entry has two sections. The first section covers the primary components used in scoring test items and highlights some similarities and differences between different item types. The second section illustrates some of the common statistics and presentation formats used in item analysis.

Scoring Expressions and Scoring Evaluators

Item analysis for paper-and-pencil multiple-choice (MC) tests is covered in most textbooks about testing and measurement theory (e.g., see references [1, 2, 6, 7] and [13]). However, traditional MC item analysis methods have not been readily extended to the growing variety of item types used on modern tests. The intent in this section is to present a structured framework, based on modern database concepts, for understanding scoring mechanisms applicable to a wide variety of item types. The framework provides the basis for discussing item analysis methods in the final section.

Classes of Item Types

In developing a framework for item analysis, it is almost imperative to consider some of the similarities among the different item types. Most item types actually fall into one of two broad classes: (a) selected-response items and (b) constructed-response items.

These item type classes do share some common data components that can be exploited.

Selected-response (SR) items include binary response items such as true-false or alternate-choice items, multiple-choice (MC) items with one-best answer, multiple-choice items with two or more best answers, and matching items that require the test taker to match the elements of two or more lists (see [8] and [13]). **Computer based testing** (CBT) has further expanded the array of selected-response item types, adding functional descriptions like *hot-spot* items, *drag-and-drop* items, *build-list-and-reorder* items, and *create-a-tree* items (e.g., [12]). Despite some of their apparent surface differences, virtually all SR item types have two common features: (a) a fixed list of alternative answer choices – often called the *answer options* or *distractors* and (b) a response-capturing mechanism that determines how the test taker must select his or her response(s) from the list of choices (e.g., filling in a ‘bubble’ on an answer sheet, using a computer mouse to click on a ‘radio button’ option, using a computer mouse to drag a line connector to indicate a relationship between two network objects).

Constructed-response (CR) items provide the examinee with prompts or sets of task directives. The examinees must respond using some type of response-capturing mechanism that can range from fill-in-the-blank fields on an answer sheet or in a test booklet to a large collection of computer software *controls* such as input boxes, text editors, spreadsheets, and interactive applications that record what the examinee does or that yield a complex performance-based work-sample. The prompts or task directives and the response-capturing mechanism may constrain the nature of the responses. However, unlike SR items, CR items do not incorporate fixed lists of alternatives from which the examinee chooses his or her answer. Therefore, examinees can produce an enormous number of possible responses.

A Framework for Scoring and Analysis of SR and CR Items

Both SR and CR items can be scored and analyzed under a broad item analysis framework by developing a system that includes three components: (a) scoring expressions; (b) scoring evaluators; and (c) a scoring expression database. These components readily generalize to most item types.

2 Item Analysis

A *scoring expression* is a type of scoring rule that converts the test taker's response(s) into a point value – usually zero or one. A *scoring evaluator* is a computer software procedure that performs a specific comparative or evaluative task. *Scoring evaluators* can range from simple pattern matching algorithms to complex analytical algorithms capable of evaluating essays and constructed responses.² A *scoring database* is where the answers and other score-related data are stored until needed by the scoring evaluator.

A scoring expression used by a scoring evaluator takes the form

$$y_i = f(r_i, a_i) \quad (1)$$

where r_i is the response set provided by the examinee and a_i is an answer key or set of answer keys. Selected-response items can directly use this type of scoring expression. For example, a conventional MC item has three to five distractors, a response-capturing mechanism that allows the examinee to select only one answer from the list of available distractors. Consider a multiple-choice item with five options, 'A' to 'E'. The examinee selects option 'B', which happens to be the correct answer. Using (1) to represent a simple pattern matching algorithm, the scoring rule would resolve to $y = 1$ if $r = a$ or $y = 0$ otherwise. In this example, $r = 'B'$ (the examinee selected 'B') and $a = 'B'$ (the answer key), so the item score resolves to $y = 1$ (correct). If r is any other value, $y = 0$ (incorrect). More complicated scoring rules may require more sophisticated scoring evaluators that are capable of resolving complex logical comparisons involving conjunctive rules such as $a = 'A \text{ AND NOT}(B \text{ OR } C)'$, or response ordering sequences such as $a = 'Object 7 \text{ before Object } 4'$. As Luecht [12] noted, when the examinee's response involves multiple selections from the list of options, the number of potential scoring expressions can grow extremely large. That is, given m choices with k response selections allowed for a given item, there are

$$C_k^m = \frac{m!}{k!(m-k)!} \quad (2)$$

possible combinations of responses, r_i , of which, only a subset will comprise the correct answers.

There are numerous types of scoring mechanisms and scoring rubrics employed with CR items. The scoring evaluators can employ simple matching

algorithms that compare the test taker's response to an idealized answer set of responses. Analytical scoring evaluators – sometimes called *automated scoring* when applied to essays or performance assessments (e.g., see references [3, 4] and [5]) – typically break-down and reformat the test taker's responses into a set of data components that can be automatically scored by a computer algorithm. The scoring algorithms are applied to the reformatted responses and may range from weighted functions of the scorable data components, based upon Boolean logic, to self-learning neural nets. Holistic scoring rubrics, typically used by trained human raters evaluating essays or written communications, outline specific aspects of performance at ordered levels of proficiency and may state specific rules for assigning points.

A scoring function for CR items is

$$y_i = f[g(r_i), h(a_i, r_{i-1})] \quad (3)$$

where $g(r_i)$ denotes a transformation of the original examinee's responses into a scorable format (e.g., extracting the entry in a spreadsheet cell or converting a parsed segment of text into a given grammatical or semantic structure), and where $h(a_i, r_{i-1})$ is an embedded answer processing function that may involve consideration of prior responses. For example, suppose that we wish to allow the examinee's computations or decisions to be used resolving subsequent answers.³ In that case, $h(a_i, r_{i-1})$ might involve some auxiliary computations using the prior responses, r_{i-1} , provided by the examinee, along with the computational formulas and other values embodied in a_i . When no prior responses are considered, the scoring expression (3) simplifies to

$$y_i = f[g(r_i), a_i]. \quad (4)$$

When no transformation of the examinee's responses is performed, the scoring expression simplifies to (1).

As noted earlier, the scoring expressions must be stored in a *scoring expression database*. Each item may have one or more scoring expression records in the database. The scoring records identify the item associated with each scoring expression, the scoring answers, arguments, rules, or other data needed to evaluate responses, and finally, reference to a scoring evaluator (explained below).

ItemID	ExpressionID	Answers	Evaluator	Value
ITM001	EXPR000001	ANS = "B"	SPM	1
ITM002	EXPR000005	ANS = "A" AND ANS = "E"	CPM	1
ITM003	EXPR000011	ANS = "945.3"; TOLER = "0.01"	SMC	1

Figure 1 A scoring expression database for three items

In special cases, programmed scripts can be incorporated as part of the data. In general, it is preferable to store as much scoring information as possible in a structured database to facilitate changes.

Figure 1 presents a simple database table for three items: ITM0001, ITM002, and ITM003. ITM001 is a one-best answer MC item where 'B' is the correct answer. ITM002 is a multiple-selection SR item with two correct answers, 'A' and 'E'. Finally, ITM003 is a CR item that requires the examinee to enter a value that is compared to a target answer, 945.3 within a prescribed numerical tolerance, ± 0.01 . Three scoring evaluators are referenced in Figure 1: (a) a simple pattern-match (SPM) evaluator for performing exact text matches such as $r_i = a_i$; (b) a complex pattern-match (CPM) evaluator that compares a vector of responses, $\mathbf{r}_i$, to a vector of answers, $\mathbf{a}_i$, using Boolean logic; and (c) a simple math compare (SMC) evaluator that numerically compares r_i to a_i , within the prescribed tolerance (TOLER). More advanced scoring evaluator might compare the ordering of selected objects or perform very complex analyses of the responses using customized scripts to process

the data. A thorough discussion of scoring evaluators is beyond the scope of this entry.

A special class of scoring expressions is *auxiliary-scoring expressions* (ASE). Their primary use is to facilitate the item analysis. For SR items, auxiliary-scoring expressions represent all possible response options other than the primary answer key(s). The item analysis procedures evaluate whether the examinee's response 'hits' on one of these auxiliary-scoring expressions (i.e., the ASE Boolean logic is *true*). For example, the four *incorrect* response options associated with a five-choice, one-best answer MC item would each constitute an auxiliary-scoring expression. Anytime an examinee selects or enters data that matches the conditions of the scoring expression, there is a 'hit' on that scoring expression. An analysis of the 'hits' on the auxiliary-scoring expressions for each item is identical to conducting a distractor analysis. These auxiliary expressions can suggest alternative correct answers or might hint at the actual correct answer for a miskeyed MC item. The auxiliary-scoring expressions can be predetermined for SR items since the list of alternative responses is known in advance. For CR items, they may need

ItemID	ExpressionID	Status	Answers	Evaluator	Value
ITM001	EXPR000000	PAK	ANS = "B"	SPM	1
ITM001	EXPR000001	ASE	ANS = "A"	SPM	0
ITM001	EXPR000002	ASE	ANS = "C"	SPM	0
ITM001	EXPR000003	ASE	ANS = "D"	SPM	0
ITM001	EXPR000004	ASE	ANS = "E"	SPM	0
ITM002	EXPR000005	PAK	ANS = "A" AND ANS = "E"	CPM	1
ITM002	EXPR000006	ASE	ANS = "A" AND NOT (ANS = "E")	CPM	0
ITM002	EXPR000007	ASE	ANS = "E" AND NOT (ANS = "A")	CPM	0
ITM002	EXPR000008	ASE	ANS = "B"	CPM	0
ITM002	EXPR000009	ASE	ANS = "C"	CPM	0
ITM002	EXPR000010	ASE	ANS = "D"	CPM	0
ITM003	EXPR000011	PAK	ANS = "945.3"; TOLER = "0.01"	SMC	1

Figure 2 A scoring expression database for three items with primary answer keys (PAK) and auxiliary-scoring expressions (ASE)

to be generated from the examinees' responses and reviewed by subject matter experts.

Figure 2 presents an elaborated scoring expression database table that incorporates the auxiliary-scoring expressions for the three items shown in Figure 1. A new database field has been added called 'Status', where PAK refers to the primary answer key and ASE denotes a auxiliary-scoring expression. The point values of the ASE records are set to zero. If any of those auxiliary-scoring expressions were subsequently judged to provide valid information, the database field could easily be set to a nonzero point value and the batch of response records rescored. This approach can incorporate fractional or even negative scoring weights.

Using this framework, item analysis becomes a set of prescribed statistical procedures for operating on stored scoring expression records. This simple database structure easily accommodates a wide variety of scoring expressions and scoring evaluators, making it relatively straightforward to generalize a common set of item analysis methods to a wide variety of item types. For example, some of the same techniques used for multiple-choice items can easily be adapted for use with complex, computer-based performance exercises.

Statistical Analysis Methods for Evaluating Scoring Expressions

There are many types of statistical analyses we might employ in an item analysis. This section highlights some of the more common ones, including tabular and graphical methods of presenting the item analysis results.

As noted earlier, traditional item analysis methods have primarily been devised for MC items. These include descriptive statistics such as frequency distributions, means, variances, tabular, and graphical presentations of score-group performance, and various discrimination indices that denote the degree to which an item or scoring expression differentiates between examinees at various levels of apparent proficiency. This section highlights four broad classes of item analysis statistics: (a) unconditional descriptive statistics; (b) conditional descriptive statistics; (c) discrimination indices; and (d) IRT misfit statistics and nonconvergence.

Unconditional Descriptive Statistics

Perhaps the most obvious descriptive statistics to use in an item are the count or percentage of total examinees that hit on a particular scoring expression, the mean score, and the variance. The count or percentage of hits on a scoring expression can highlight problems such as scoring expressions that no one ever chooses. The mean provides an indication of difficulty and the variance indicates whether the full range of possible scores is being adequately used.

Means and variances can also be computed conditionally, as discussed in the next section. The point values can also be aggregated over multiple scoring expressions that belong to the same unit (e.g., an SR item with multiple correct answers as unique scoring expressions). Provided that an appropriate database field exists for grouping together multiple scoring expressions (e.g., an item identifier), it is relatively straightforward to add this aggregation functionality to the item analysis system.

Conditional Descriptive Statistics

The capability of scoring expressions to distinguish between individuals at different levels of proficiency is referred to as *discrimination*. Scoring expressions that discriminate well between the examinees at different proficiency levels are usually considered to be better than expressions that fail to differentiate between the examinees. Discrimination can be evaluated by computing and reporting descriptive statistics conditionally for two or more score groups, $k = 1, \dots, G$, usually based on total test performance or some other criterion score. Conditional descriptive statistics include frequencies, means, and standard deviations that are computed and reported in tabular or graphical form for each score group.

The criterion test score is essential for creating the score groups used in computing the conditional statistics. The raw total test score is usually used as the criterion score for paper-and-pencil tests. Even then, the total test score, $T_j = \sum y_{ij}$, must be systematically adjusted by subtracting y_{ij} for each examinee, $j = 1, \dots, N$. The raw total test score also fails as the optimal choice as the criterion test score when adaptive testing algorithms are used, or when the test forms seen by individual examinees may differ in average difficulty. For example, a very able

examinee and a less-able examinee can have the same raw score on an adaptive test simply because the more able examinee was administered more difficult items.

A useful alternative to the total test score is an IRT (*see Item Response Theory (IRT) Models for Dichotomous Data*) ability score⁴ computed from operational test items, using a set of stable item parameter estimates from an item bank. Assuming that all of the IRT item parameters in the item bank have been calibrated to a common metric, the estimated criterion score, T_j , $j = 1, \dots, N$, will be on a common scale, regardless of which test form a particular examinee took. T_j can then be computed once for each examinee and then used in a simultaneous item analysis of all scoring expressions.

Given the criterion scores obtained for all examinees in the item analysis sample, two or more score groups can be formed, $k = 1, \dots, G$. Item analyses typically form the score groups to have approximately equal numbers of examinees in each group – that is, $F_1 = F_2 = \dots = F_G$ – versus using equally spaced intervals on the score scale.

One of the most common conditional item analysis methods is a high–low table of conditional frequencies or relative frequency percentages. This type of table provides a convenient way to simultaneously evaluate the correct answers and incorrect options (distractors) for MC items. The same technique can easily be extended to other types of SR and CR item. First, two contrasting score groups are formed, based on the criterion test scores. Kelley [11] recommended that the two score groups be formed by taking the upper and lower 27% of the examinees as the high and low groups – optionally, other percentages, up to a 50:50 split of the sample, have been effectively used.

Figure 3 shows a distractor analysis for a single MC item with five response options. The correct

answer is ‘B’ (i.e., the scoring evaluator returns $y_i = 1$ if $r_i = \text{‘B’}$). The remaining four response options are auxiliary-scoring expressions. The distractor analysis indicates that 91% of the examinees in the high score group selected the correct response. In contrast, only 26% of the low-score group examinees selected the correct answer. All of the other response options were attractive to the examinees in the lower score group.

A multi-line plot can also be used to simultaneously evaluate a number of scoring expressions when more than two score groups are used. For example, suppose that we compute criterion scores and categorize the examinees into five equal-sized groups or quintiles (i.e., 0–20% of the examinees are placed in Group 1, 21–40% in Group 2, etc.). The multi-line plot shows the proportion of examinees within each group who selected each of the scoring expressions. Figure 4 shows the simultaneous percentages of examinees within five equal-sized score groups, labeled 1 to 5 along the horizontal axis. There is one line for each of the six scoring expressions for ITM002, shown earlier in Figure 2. Scoring expression EXPR000005 is the primary answer key. The remaining expressions are auxiliary scoring expressions. The multi-line plot shows that the majority of the candidates in score groups 4 and 5 properly selected both options ‘A’ and ‘E’. The remaining lines are fairly flat. This type of plot provides two types of useful information for evaluating the scoring expressions. First, if the line for a primary answer key is decreasing across the low-to-high score groups, there may be a serious problem with that particular key. Second, if any auxiliary-scoring expression demonstrates an increasing trend across the low-to-high score groups, that expression might be a valid secondary answer key (or the correct key, if line for the primary answer

Response options	A	B*	C	D	E	Omits total
High score group (N_H)	6	179	7	5	0	0 197
Low score group (N_L)	54	52	37	47	5	2 197
Total group (N)	105	448	75	83	27	3 741
High score group %	3%	91%	4%	3%	0%	0% 100%
Low score group %	27%	26%	19%	24%	3%	1% 100%
Total group %	14%	60%	10%	11%	4%	0% 100%

* Primary answer key

Figure 3 High-low group distractor analysis for a five-option MC item

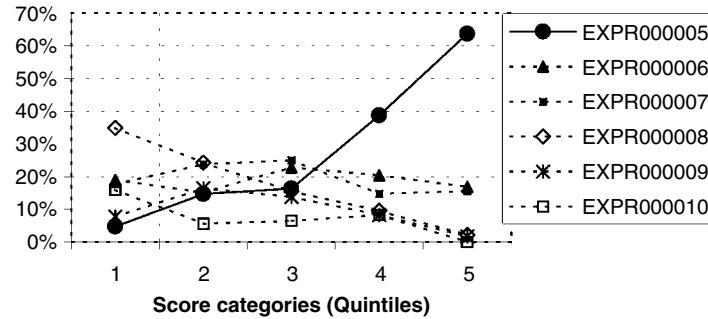


Figure 4 A multi-line for six scoring expressions with five score groups

ItemID	ExpressionID	Status	Answer	Count	Mean	SD	Min.	Max.
ITM001	EXPR000000	PAK	ANS = "B"	448	0.65	0.23	-0.15	2.58
ITM001	EXPR000001	ASE	ANS = "A"	105	-0.26	0.35	-2.16	0.30
ITM001	EXPR000002	ASE	ANS = "C"	75	-0.13	0.41	-1.88	0.78
ITM001	EXPR000003	ASE	ANS = "D"	83	-0.38	0.59	-2.43	0.69
ITM001	EXPR000004	ASE	ANS = "E"	27	-0.22	0.60	-1.93	0.51

Figure 5 A summary of conditional descriptive statistics for the criterion test for five scoring expressions associated with an MC item

key is simultaneously decreasing across the score groups).

A third way to indicate how well the scoring expressions discriminate is to compute the means and standard deviations of the criterion test values conditional for examinees who ‘hit’ on each of the scoring expressions. Examinees whose responses match the primary answer keys should have higher mean scores than those whose responses match the auxiliary answer keys. This method is particularly useful with computer-adaptive tests if IRT scoring is used to compute the criterion scores.

Figure 5 presents a table of conditional descriptive statistics for the criterion scores associated with the five expressions for ITM001 shown earlier in Figure 2. The criterion scores reported are IRT proficiency scores computed for each examinee. Therefore, all of the proficiency scores are on the same scale. As we might hope, the mean score on the criterion test is higher for the primary answer key than for any of the other options. Similar to the distractor analysis, the implication is that higher scoring examinees tend to select the correct response more frequently than lower scoring examinees.

Figure 6 presents the same information as Figure 5, but in graphical form. The symbols distinguish between the primary answer key (PAK) and the auxiliary-scoring expressions (ASI). The error bars denote the standard deviations of the proficiency scores about the mean proficiency score for each expression. If the PAK were incorrect, the mean might be lower than some or all of the ASE means. A possible valid auxiliary-scoring expression would appear with a higher mean than the other ASE values. If the PAK versus ASE means are similar in magnitude, or if the standard error bars overlap a great deal, we might look to change one or more scoring expressions to extract more discriminating information.

Discrimination Indices

There are three discrimination indexes typically used in an item analysis: the U-L index, the point-biserial correlation, and the biserial correlation. These tend to augment some of the descriptive conditional statistical methods.

The U-L index of discrimination, sometimes called *D*, (see references [7] and [11]) is the simple difference between the mean scores for the examinees

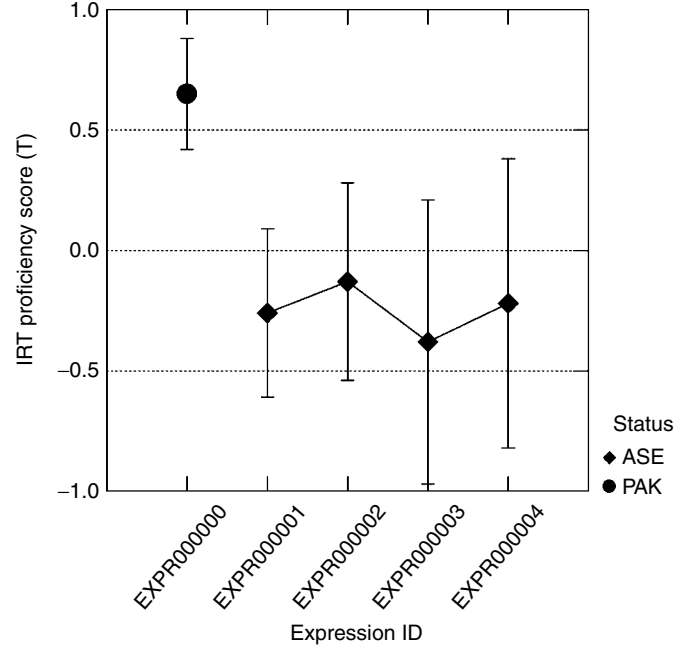


Figure 6 Plot of conditional means and standard deviations for the criterion test for five scoring expressions associated with an MC item

in the upper (U) 27% and the lower 27% of the distribution of total test scores. That is,

$$D = \frac{\bar{Y}_{Ui} - \bar{Y}_{Li}}{m_i} \quad (5)$$

where m_i denotes the total possible points for the item. When applied to dichotomous items, D is the simple difference between the P values (item means). The upper bound of D is +1.0; the lower bound is -1.0. $D \geq 0.4$ is generally considered to be satisfactorily. Values in the range $0.2 \leq D \leq 0.39$ indicate possible problems, but nothing serious. Values of $D \leq 0.19$ are considered to be unacceptable and indicate a scoring expression that should be eliminated or that requires serious revision.

Point-biserial and biserial correlations can be used to denote the relative strength of the association between getting a 'hit' on a particular scoring expression and the criterion test score. When the expression-criterion score is positive and reasonably high, it indicates that higher performance on the total test is associated with correspondingly higher performance on the item or scoring expression. Low values

indicate nondiscriminating items and negative values usually signal a potential serious flaw in the scoring expression. When scoring expression represents a dichotomously score, we use a point-biserial correlation.

A **point-biserial correlation** is the **Pearson product-moment correlation** between a dichotomous and a continuous variable,

$$r_{pbis(i)} = \frac{\bar{T}_{i+} - \bar{T}}{s_T} \sqrt{\frac{p_i}{(1 - p_i)}} \quad (6)$$

where $\bar{T}_{i+}$ is the mean criterion test score for examinees who correctly answered item i , $\bar{T}$ is the mean for the entire sample, s_T is the sample standard deviation, and p_i is the item P value.

If we assume that there is an underlying normal distribution of ability that is artificially dichotomized for the primary answer key score, as well as for hits on the auxiliary-scoring expressions, we can use a biserial correlation. The biserial correlation is computed as

$$r_{bis(i)} = \frac{\bar{T}_{i+} - \bar{T}}{s_T} \sqrt{\frac{p_i}{h}} \quad (7)$$

where $\bar{T}_{i+}$, $\bar{T}$, s_T , and p_i are defined the same as for the point-biserial correlation. The variable h represents the ordinate of the standard normal curve at a z -value corresponding to p_i .

The point-biserial is easier to compute than the biserial correlation, but that is no longer a major consideration for computerized item analysis. However, the biserial correlation is about one-fifth greater in magnitude in the mid-range of P values. The relationship between r_{pbis} and r_{bis} is

$$r_{bis(i)} = r_{pbis(i)} \frac{\sqrt{p_i q_i}}{h}. \quad (8)$$

Some test developers prefer the biserial as an overall discrimination indicator – perhaps because it is systematically higher than the point biserial. However, the point-biserial correlation tends to be more useful in item analysis and test construction.

As part of an item analysis, point-biserial correlations can be presented in much the same way as graphics or tables computed using conditional statistics. Primary answer keys with positive values of the point biserial – usually at least 0.3 or higher – are considered to be acceptable. Primary answer keys within the range $0.1 \leq r_{pbis} \leq 0.29$ typically warrant serious scrutiny, but may still be associated with a useful item. The utility of keeping primary answer keys where $r_{pbis} \leq 0.1$ is highly questionable.

If each hit on a secondary scoring expression is treated as a dichotomous score (i.e., $y_{ij} = 1$, regardless of the point values for the scoring expression) point-biserial correlations can also be computed for the auxiliary-scoring expressions. Plausible secondary answer keys can sometimes be detected by finding ASEs with moderate-to-high r_{pbis} values. If the r_{pbis} value associated with an ASE is higher than the PAK r_{pbis} , the analysis may indicate that the primary answer key is a miskey and simultaneously suggest what the appropriate primary key should be.

IRT Misfit Analysis and Convergence

Item response theory (IRT) can also be used to evaluate the quality of each scoring key (e.g., [9] and [10]). There are two broad classes of analyses that deserve mention here. The first is IRT misfit analysis. IRT misfit analyses can be employed to evaluate how well the empirical response data match a probability function curve produced by a particular IRT

model. Primary answer key scoring expressions that fail to fit a given IRT model may signal serious scaling problems if the item is retained for operational scoring. IRT model misfit can further be evaluated to uncover dependencies among various scoring expressions, to detect differential functioning of the items for different population subgroups, or even to suggest possible multidimensionality problems – especially when a test is assumed to measure one primary trait or a composite of primary traits.

A second type of IRT analysis involves a review of convergence problems with respect to particular item types. Some of these problems may be due to nondiscriminating scoring expression, violations of IRT model assumptions that lead to statistical identifiability problems, or messy data that can be traced to faulty instructions to the examinees or an overly complex user interface used with a new CBT item type.

A thorough discussion of IRT or these issues is far beyond the scope of this entry. Nonetheless, as testing organizations move more toward using IRT – especially with CBT and some of the new item types – it is essential to find ways to discover problems with the scoring expressions, before the items are used operationally. As noted in the previous two sections, item analysis using conditional descriptive statistics or scoring expression-criterion test correlations, where the latter is based on an IRT proficiency score, is also an approach that combines IRT and classical test item analysis methods.

Notes

1. The focus in this entry is on large-scale testing applications. In classroom settings, item analysis can be employed to help teachers evaluate their assessments and to use the results to improve instruction. However, this assumes that the teachers have proper training in interpreting item analysis results and are provided with convenient and easy-to-use software tools [10].
2. Human raters can be considered to be *scoring evaluators* and incorporated into this framework.
3. Introducing obvious dependencies among the scored responses is typically frowned upon as an item writing practice. Nonetheless, it can be argued that the scoring and item analysis system should be capable of handling those types of responses.
4. The three most common estimators used for IRT scoring are a maximum likelihood estimator, a Bayes

mean estimator, or a Bayes mode estimator. Hambleton and Swaminathan [9] provide details about IRT estimation and scoring.

References

- [1] Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*, Brooks/Cole, Monterey.
- [2] Anastasi, A. (1986). *Psychological Testing*, 6th Edition, Macmillan, New York.
- [3] Baxter, G.P. Shavelson, R.J. Goldman, S.R. & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessments, *Journal of Educational Measurement* **29**, 1–17.
- [4] Braun, H.I. Bennett, R.E. Frye, D. & Soloway, E. (1990). Scoring constructed responses using expert systems, *Journal of Educational Measurement* **27**, 93–180.
- [5] Clauser, B.E. Margolis, M.J. Clyman, S.G. & Ross, L.P. (1997). Development of automated scoring algorithms for complex performance assessment, *Journal of Educational Measurement* **36**, 29–45.
- [6] Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*, Holt, Rinehart and Winston, New York.
- [7] Ebel, R.L. (1965). *Measuring Educational Achievement*, Prentice-Hall, Englewood Cliffs.
- [8] Haladyna, T.M. Downing, S.M. & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines, *Applied Measurement in Education* **15**, 209–334.
- [9] Hambleton, R.K. & Swaminathan, H.R. (1985). *Item Response Theory: Principles and Applications*, Kluwer Academic Publishers, Boston.
- [10] Hsu, T. & Yu, L. (1989). Using computers to analyze item response data, *Educational Measurement: Issues and Practice* **8**, 21–28.
- [11] Kelley, T.L. (1939). Selection of upper and lower groups for the validation of test items, *Journal of Educational Psychology* **30**, 17–24.
- [12] Luecht, R.M. (2001). Capturing, codifying, and scoring complex data for innovative, computer-based items, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, Seattle.
- [13] Popham, W.J. (2002). *Classroom Assessment: What Teachers Need to Know*, 3rd Edition, Allyn & Bacon, Boston.

RICHARD M. LUECHT

Item Bias Detection: Classical Approaches

GIDEON J. MELLEMBERGH

Volume 2, pp. 967–970

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Bias Detection: Classical Approaches

In the 1960s, concern was raised on the fairness of educational and psychological tests. The test performance of American minority group members stayed behind the performance of Whites, and tests were blamed for being biased against minorities [2]. Methods were developed to detect biased items, but these methods did not clearly distinguish between impact and item bias. Impact means that between-group item performance differences are explained by between-group ability distribution differences, whereas item bias means that between-group item performance differences exist for examinees of equal ability from different groups [10].

Classical test theory (CTT) applies to the observed test score [11]. The observed test score is used as an estimate of an examinee's true test score. Scheuneman [16] proposed to study item bias conditional on the test score: She considered an item to be biased between different (e.g., ethnic, sex, cultural) groups when item performance differs between examinees from different groups who have the same test score. Her proposal founded CTT-based item bias detection methods that do not confuse item bias and impact.

The classical concept of item bias had a rather limited scope. It applied to achievement and ability tests that consist of dichotomously scored items. The concept was broadened into several directions. The term 'item bias' was replaced by the more neutral term **differential item functioning** (DIF), and DIF was extended from achievement and ability tests to attitude and personality questionnaires, and from dichotomously scored item responses to item response scales that have more than two answer categories.

The situation is considered of a test that predominantly measures one latent variable, such as a cognitive ability, attitude, or personality trait. The test consists of n items; the observed test score, X , is the unweighted or weighted sum of the n item scores. It is studied whether an item is differentially functioning with respect to a violator variable V [14]. Usually, V is an observed group membership variable, where the focal group ($V = f$) is the particular group of interest (e.g., a minority group) and the reference group ($V = r$) is the group with whom the focal

group is compared (e.g., the majority group). An item is differentially functioning between focal and reference group when the item responses of equal ability (trait level) examinees (respondents) differ between the two groups.

Dichotomous Item Responses

The items of the test are dichotomously scored: $I = 1$ for a correct (affirmative) answer and $I = 0$ for an incorrect (negative) answer. $P_{ir}(x)$ and $P_{if}(x)$ are the probabilities of giving the correct (affirmative) answer to the i th item, given the test score, in the reference and focal group, respectively. $P_{ir}(x)$ and $P_{if}(x)$ are the regression functions of the item score on the test score, that is $P_{ir}(x) = E(I_i | X = x, V = r)$ and $P_{if}(x) = E(I_i | X = x, V = f)$. Under CTT, item i is considered to be free of DIF if

$$P_{ir}(x) = P_{if}(x) = P_i(x) \quad \text{for all values } x \text{ of } X. \quad (1)$$

The **Logistic Regression, Logit, and Mantel-Haenszel Methods** for DIF detection use similar specifications of (1). The regression function of the Logistic Regression Method [18] is

$$P_i(x) = \frac{e^{\beta_{i0} + \beta_{i1}x}}{1 + e^{\beta_{i0} + \beta_{i1}x}}, \quad (2)$$

where β_{i0} is the difficulty (attractiveness) parameter and β_{i1} the discrimination parameter of the i th item. Under Model (2), the CTT definition is satisfied if the parameters of the reference and focal group are equal, that is, $\beta_{i0r} = \beta_{i0f}$ and $\beta_{i1r} = \beta_{i1f}$. If $\beta_{i1r} \neq \beta_{i1f}$, DIF is said to be nonuniform; and if $\beta_{i0r} \neq \beta_{i0f}$ and $\beta_{i1r} = \beta_{i1f}$, DIF is said to be uniform because the regression functions of reference and focal group only differ in location (difficulty or attractiveness) and not in discrimination. Likelihood ratio tests are used for testing the null hypotheses of equal item parameters across groups.

The Logistic Regression Method treats the test score as a quantitative covariate. In contrast, the Logit Method [13] treats the test score as a nominal variable. The studied item i is deleted from the test, and for each of the examinees (respondents) the rest score, that is, the score on the remaining $n - 1$ items is computed. The sample size per rest score may be too small, and mostly the rest scores are classified

2 Item Bias Detection: Classical Approaches

into a smaller number (three to seven) categories. The Logit Method adapts (2) to the situation of a nominal test score for matching reference and focal group members. Technically, this adaptation is done by using dummy coding (*see Dummy Variables*) of the test score categories and substituting dummy variables for the test score into (2). Likelihood ratio tests for the equality of item parameters across groups are used for the study of nonuniform and uniform DIF.

The Mantel–Haenszel Method [9] uses the test score itself for matching reference and focal group members. The Mantel–Haenszel Method adapts (2) to the situation of a nominal test score by using dummy coding of the test scores and substituting dummy variables for the test score into (2). Moreover, the Mantel–Haenszel Method assumes that DIF is uniform, that is, it sets β_{i1r} equal to β_{i1f} [1]. The Mantel–Haenszel statistic is used for testing the null hypothesis of uniform DIF ($\beta_{i0r} = \beta_{i0f}$). This statistic has some advantages above the likelihood ratio test for uniform DIF. First, the Mantel–Haenszel test is less affected by test scores that have small sample sizes [1]. Second, the test has the largest **power** for testing the null hypothesis $\beta_{i0r} = \beta_{i0f}$ under the assumption $\beta_{i1r} = \beta_{i1f}$. Therefore, the Mantel–Haenszel test is the best test for DIF when it can be assumed that DIF is uniform [5]. However, the test is less appropriate when DIF is nonuniform. The Mantel–Haenszel Method was adapted to a procedure that can detect nonuniform DIF by [12].

The Standardization and SIBTEST methods use the test score for matching reference and focal group members, but they do not use parametric regression functions. The Standardization Method [6] defines the standardization index as

$$SI_i = \sum_x \frac{N_{xf}}{N_f} \{P_{if}(x) - P_{ir}(x)\}, \quad (3)$$

where $\sum_x$ means that the summation is over all observed test scores, N_{xf} is the number of focal group members who have test score x and N_f is the total number of focal group members. Estimates of the index and its variance can be used to test the null hypothesis that the index is 0. The value of SI can be near 0 when positive and negative differences cancel out in (3). Therefore, SI is appropriate when all differences have the same sign across the test scores, which was called unidirectional DIF [17]. Unidirectional DIF means that the item is always

biased against the same group at each of the test scores. However, the SI is less appropriate when $\{P_{if}(x) - P_{ir}(x)\}$ is positive at some test scores and negative at other scores. It is remarked that a uniform DIF item is always a unidirectional DIF item, but a unidirectional DIF item does not need to be a uniform DIF item [8].

The SIBTEST Method for the detection of a DIF item splits the test into a subtest of $n - 1$ items that is used for matching and a single studied item. A modified version of SI is used that compares the probabilities of giving the correct (affirmative) answer of reference and focal group members, given the examinees' (respondents') estimated true scores, where the true scores are estimated separately for the reference and focal group using the CTT regression of true score on observed score.

Ordinal-polytomous Item Responses

The item response variable has C ($C > 2$) ordered categories, for example, a five-point Likert scale ($C = 5$). $P_{icr}(x)$ and $P_{icf}(x)$ are the probabilities of responding in the c th category of the i th item, given the test score, in the reference and focal group, respectively. Under CTT, item i is considered to be free of DIF if these probabilities are equal across all values of the test score, which means that $C - 1$ probabilities must be equal because the C probabilities sum to 1:

$$P_{icr}(x) = P_{icf}(x) = P_{ic}(x), \quad c = 1, 2, \dots, C - 1, \\ \text{for all values } x \text{ of } X. \quad (4)$$

The previously discussed methods for DIF detection of dichotomous items were extended to polytomous items; a framework is given by [15].

The Logit Method was described as a special case of the Loglinear Method for polytomous items (*see Log-linear Models*) [13]. The Logistic Regression Method was extended to the Multiple-Group Logistic Regression Method, which subsumes the Proportional Odds Method as a special case [20]. The Mantel–Haenszel Method was extended to the Generalized Mantel–Haenszel and Mantel Methods [21]. These extended methods treat the test score in the same way as their corresponding methods for dichotomous items: The Loglinear, Generalized Mantel–Haenszel, and Mantel Methods

use the test score as a nominal matching variable, whereas the Multiple-Group Logistic Regression and Proportional Odds Methods use the test score as a quantitative covariate. Moreover, the extended methods differ in the way the categories of the response variable are handled: The Loglinear and Generalized Mantel–Haenszel Methods treat them as nominal categories, the Multiple-Group Logistic Regression and Proportional Odds Methods treat them as ordered categories, and the Mantel Method assigns quantitative values to the ordered categories.

The Standardization and SIBTEST Methods were extended to the Standardized Mean Differences [7, 22] and POLYSIBTEST Methods [3], respectively. Both methods assign quantitative values to the item response categories. The Standardized Mean Differences Method treats the test score as a nominal matching variable, whereas the POLYSIBTEST Method uses the estimated true score per group for matching reference and focal group members.

Comments

The test score is used to correct for ability (trait) differences between reference and focal group. Therefore, the test must predominantly measure one latent construct. Psychometric methods can be used to study the dimensionality of a test, but DIF yields a conceptual problem. A differentially functioning item is by definition multidimensional, for example, a mathematics item may be differentially functioning between native and nonnative speakers because nonnative speakers have difficulty to understand the content of the item. A pragmatic approach to the multidimensionality problem is a stepwise application of DIF detection methods to purify the test [10, 19].

The statistics of the different DIF detection methods are sensitive to different types of DIF, and test different types of null hypotheses. The Logit and Logistic Regression Methods and their extensions are sensitive to both uniform and nonuniform DIF, the Mantel–Haenszel Method and its extensions to uniform DIF, and the Standardization and SIBTEST Methods and their extensions to unidirectional DIF.

From a theoretical point of view, each of the CTT-based methods is flawed. The test score is handled as a nominal matching variable or as a quantitative covariate, but it can better be considered as an ordinal variable [4]. From the CTT perspective, the per group

estimated true score must be preferred for correction of ability (trait) differences, but most of the methods use the test score. Usually, the response categories of a polytomous item are ordinal, but only the Multiple-Group Regression and Proportional Odds Methods treat them correctly as ordered. Theoretically, IRT-based methods (*see Item Bias Detection: Modern Approaches*) should be preferred. However, the application of IRT-based methods requires that a number of conditions are fulfilled. Therefore, CTT-based methods are frequently preferred in practice.

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.
- [2] Angoff, W.H. (1993). Perspectives on differential item functioning methodology, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Erlbaum, Mahwah, pp. 3–23.
- [3] Chang, H.-H., Mazzeo, J. & Roussos, L. (1996). Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure, *Journal of Educational Measurement* **33**, 333–353.
- [4] Cliff, N. & Keats, J.A. (2003). *Ordinal Measurement in the Behavioral Sciences*, Erlbaum, Mahwah.
- [5] Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Erlbaum, Mahwah, pp. 35–66.
- [6] Dorans, N.J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic attitude test, *Journal of Educational Measurement* **23**, 355–368.
- [7] Dorans, N.J. & Schmitt, A.P. (1993). Constructed response and differential item performance: a pragmatic approach, in *Construction Versus Choice in Cognitive Measurement*, R.E. Bennett & W.C. Ward, eds, Erlbaum, Hillsdale, pp. 135–166.
- [8] Hessen, D.J. (2003). Differential item functioning: types of DIF and observed score based detection methods, Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.
- [9] Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure, in *Test Validity*, H. Wainer & H. Braun, eds, Erlbaum, Hillsdale, pp. 129–145.
- [10] Lord, F.M. (1977). A study of item bias, using item characteristic curve theory, in *Basic Problems in Cross-cultural Psychology*, Y.H. Poortinga, ed., Swets & Zeitlinger, Amsterdam, pp. 19–29.
- [11] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

4 Item Bias Detection: Classical Approaches

- [12] Mazor, K.M., Clauser, R.E. & Hambleton, R.K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure, *Educational and Psychological Measurement* **54**, 284–291.
- [13] Mellenbergh, G.J. (1982). Contingency table models for assessing item bias, *Journal of Educational Statistics* **7**, 105–118.
- [14] Oort, F.J. (1993). Theory of violators: assessing unidimensionality of psychological measures, in *Psychometric Methodology, Proceedings of the 7th European Meeting of the Psychometric Society in Trier*, R. Steyer, K.F. Wender & K.F. Widaman, eds, Fischer, Stuttgart, pp. 377–381.
- [15] Potenza, M. & Dorans, N.J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation, *Applied Psychological Measurement* **19**, 23–37.
- [16] Scheuneman, J.A. (1979). A method of assessing bias in test items, *Journal of Educational Measurement* **16**, 143–152.
- [17] Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF, *Psychometrika* **58**, 159–194.
- [18] Swaminathan, H. & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement* **27**, 361–370.
- [19] Van der Flier, H., Mellenbergh, G.J., Adèr, H.J. & Wijn, M. (1984). An iterative item bias detection method, *Journal of Educational Measurement* **21**, 131–145.
- [20] Wellkenhuysen-Gijbels, J. (2003). The detection of differential item functioning in Likert score items, Unpublished doctoral dissertation, Catholic University of Leuven, Belgium.
- [21] Zwick, R., Donoghue, J.R. & Grima, A. (1993). Assessment of differential item functioning for performance tasks, *Journal of Educational Measurement* **30**, 233–251.
- [22] Zwick, R. & Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items, *Journal of Educational and Behavioral Statistics* **21**, 187–201.

(See also **Classical Test Models**)

GIDEON J. MELLENBERGH

Item Bias Detection: Modern Approaches

GIDEON J. MELLEBERGH

Volume 2, pp. 970–974

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Bias Detection: Modern Approaches

In the 1960s, concern was raised on the fairness of educational and psychological tests. It was found that the test performance of American minority group members stayed behind the performance of Whites, and tests were claimed to be biased against minorities [1]. Methods were developed to detect test items that put minority group members at a disadvantage. These methods did not clearly distinguish between impact and item bias. Impact means that between-group item performance differences are explained by between-group ability distribution differences, whereas item bias means that between-group item performance differences exist for examinees of equal ability from different groups. Lord [9] described an **item response theory (IRT)** based item bias detection method, which does not confuse item bias and impact.

The concept of item bias had a rather limited scope. It applied to achievement and ability tests, which consist of dichotomously scored items, and to ethnic, sex, and cultural groups. The concept was broadened into several directions. The term ‘item bias’ was replaced by the more neutral term ‘**differential item functioning**’ (DIF); DIF was extended from achievement and ability tests to attitude and personality questionnaires, from dichotomously scored item responses to other response scales, and from a group membership variable to other types of variables that may affect item responses; and DIF was subsumed under the general concept of ‘measurement invariance’ (MI) [13, 14].

DIF Definition

Mellenbergh [11] gave a definition of an unbiased item, which subsumes item bias under the more general MI definition. This definition is adapted to the situation of a unidimensional test (questionnaire), which is intended to measure one latent variable Θ . The observed item response variable is indicated by I , and the unintended variable which may affect the item responses is called the *violation* V [18]. The definition of a non-DIF item is:

Item i for the measurement of the **latent variable** Θ is not differentially functioning with respect to the potential violator V , if and only if

$$f(I_i|\Theta = \theta, V = v) = f(I_i|\Theta = \theta) \quad (1)$$

for all values θ of Θ and v of V , where $f(I_i|\Theta = \theta, V = v)$ is the conditional distribution of the responses to the i th item given the latent variable and the violator and $f(I_i|\Theta = \theta)$ is the conditional distribution of the responses to the i th item given only the latent variable; otherwise, item i is differentially functioning with respect to V .

Equation (1) means that, given the intended latent variable, the item responses and the potential violator are independently distributed. This definition is very general. It is assumed that I is intended to measure one latent variable Θ , but it does not make any other assumption on I , Θ , and V , for example, V can be an observed or a latent variable [8].

Modern (IRT) DIF detection methods are based on this definition. The general strategy is to check whether (1) applies to observed item responses. In practice, specifications must be made. Some well-known DIF situations are discussed. For each of these specifications, the latent variable is a continuous variable, that is, a latent trait; and the violator, a nominal (group membership) variable; a group of special interest (e.g., Blacks) is called the *focal group* ($V = f$) and the group with whom the focal group is compared (e.g., Whites) is called the *reference group* ($V = r$). For this situation, (1) converts into:

$$\begin{aligned} f(I_i|\Theta = \theta, V = r) &= f(I_i|\Theta = \theta, V = f) \\ &= f(I_i|\Theta = \theta). \end{aligned} \quad (2)$$

Dichotomous Item Responses

The item response is dichotomously scored: $I = 1$ for a correct (affirmative) answer and $I = 0$ for an incorrect (negative) answer. It is assumed that the item responses are, conditionally on the latent trait, Bernoulli distributed with parameters $P_{ir}(\theta)$ and $P_{if}(\theta)$ of giving the correct (affirmative) answer to the i th item in the reference and focal group, respectively. The $P_{ir}(\theta)$ and $P_{if}(\theta)$ are the regression functions of the observed item response on the latent trait, that is, $P_{ir}(\theta) = E(I_i|\Theta = \theta, V = r)$ and $P_{if}(\theta) = E(I_i|\Theta = \theta, V = f)$, and are called *item response functions* (IRFs).

2 Item Bias Detection: Modern Approaches

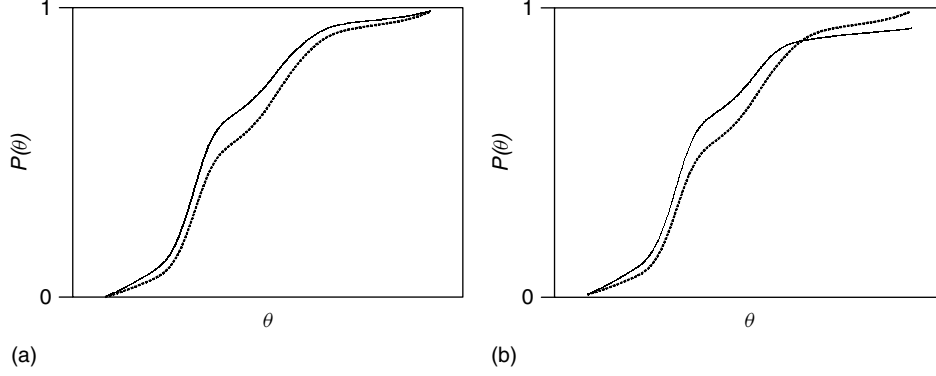


Figure 1 Examples of (a) unidirectional and (b) bidirectional DIF

A Bernoulli distribution (*see Catalogue of Probability Density Functions*) is completely determined by its parameters. Therefore, (2) is equivalent to:

$$P_{ir}(\theta) = P_{if}(\theta) = P_i(\theta). \quad (3)$$

Equation (3) states that the IRFs of the reference and focal group are equal. The general strategy of DIF detection is to check whether (3) holds across the values of the latent trait.

If (3) does not hold, item i is differentially functioning between reference and focal group. Equation (3) can be violated in different ways, and different types of DIF can be distinguished [4, 5], for example unidirectional and bidirectional DIF [22]. Unidirectional DIF means that the IRFs of the reference and focal group do not intersect (Figure 1(a)), whereas bidirectional DIF means that the two IRFs intersect at least once (Figure 1(b)).

Unidirectional DIF implies that any bias of the item is always against the same group, whereas bidirectional DIF implies that the item is biased against one group for some values of the latent trait and biased against the other group for other values of the latent trait.

Nonparametric Item Response Theory models do not use parameters to characterize the IRF. Mokken's [15, 16] nonparametric double monotonicity model (DMM) can be used to study between-group DIF. The DMM assumes that the IRF is a nonparametric monotonically nondecreasing function of the latent trait, and that the IRFs of the n items of a test (questionnaire) do not intersect; an example is shown in Figure 1(a). Methods to detect DIF under the DMM are described by [23].

Parametric IRT models specify IRFs that are characterized by parameters; models that specify a normal ogive or logistic IRF are discussed by [3] and [10]. These IRFs are characterized by one-, two-, or three-item parameters; DIF is studied by testing null hypotheses on these parameters. An example of a parametric IRF is Birnbaum's [2] two-parameter logistic model (2PLM):

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad (4)$$

where a_i is the item discrimination parameter and b_i is the item difficulty (attractiveness) parameter.

Under the 2PLM, the definition of a non-DIF item (3) is satisfied if the item parameters of the reference and focal group are equal, that is, $a_{ir} = a_{if}$ and $b_{ir} = b_{if}$. Lord [10] developed a chi-square test for the null hypothesis of equal item parameters across groups, and a likelihood ratio test for this null hypothesis is given by [24]. An overview of statistics for testing the null hypothesis of equal item parameters under IRT models for free-answer and multiple-choice items is given by [26].

The rejection of the null hypothesis of equal item parameters means that the item is differentially functioning between reference and focal group, but it does not show the amount of DIF. Three different types of DIF-size measures were described in the literature. The first type quantifies the amount of DIF at the item level [19, 20]. The signed area measure is the difference between the two IRFs across the latent trait:

$$SA_i = \int_{-\infty}^{\infty} \{P_{if}(\theta) - P_{ir}(\theta)\} d\theta. \quad (5)$$

Under some IRT models (e.g., the 2PLM), DIF can be bidirectional. The SA measure can be near 0 for intersecting IRFs when positive and negative differences cancel each other out; the unsigned area measure is the distance (area) between the two IRFs:

$$UA_i = \int_{-\infty}^{\infty} |P_{if}(\theta) - P_{ir}(\theta)| d\theta. \quad (6)$$

The SA and UA measures assess DIF size at the item level, but they do not show the effect of DIF at the examinees' (respondents') level. The second type of DIF-size measures [28] combines DIF at the item level with the latent trait distribution of the focal group. Some of the items of a test (questionnaire) can be biased against the focal group, whereas other items can be biased against the reference group, which means that bias can compensate within a test (questionnaire). Therefore, a DIF-size measure that is compensatory within a test (questionnaire) was developed by [21].

Ordinal-polytomous Item Responses

The item response variable has C ($C > 2$) ordered categories, for example, a five-point Likert scale ($C = 5$). It is assumed that the item responses are, conditionally on the latent trait, multinomially distributed with probabilities $P_{icr}(\theta)$ and $P_{icf}(\theta)$ of answering in the c th category ($c = 1, 2, \dots, C$) in the reference and focal group, respectively.

A C -category multinomial distribution is completely determined by $C - 1$ probabilities because the C probabilities sum to 1. Therefore, the definition of a non-DIF item (2) is equivalent to

$$P_{icr}(\theta) = P_{icf}(\theta) = P_{ic}(\theta), c = 1, 2, \dots, C - 1. \quad (7)$$

Parametric IRT models were defined for these probabilities under the constraint that the ordered nature of the C categories is preserved [27]. DIF detection proceeds by testing the null hypothesis of equal item parameters across the reference and focal group.

Continuous Item Responses

The item response is on a continuous scale, for example, the distance to a mark on a line. It is

assumed that the item responses are, conditionally on the latent trait, normally distributed. This assumption may be violated in practical applications because response scales are usually bounded (e.g., a 4-inch line segment).

A normal distribution is completely determined by its mean and variance. Therefore, (2) is equivalent to:

$$\begin{aligned} E(I_i | \Theta = \theta, V = r) &= E(I_i | \Theta = \theta, V = f) \\ &= E(I_i | \Theta = \theta), \end{aligned} \quad (8a)$$

$$\begin{aligned} Var(I_i | \Theta = \theta, V = r) &= Var(I_i | \Theta = \theta, V = f) \\ &= Var(I_i | \Theta = \theta), \end{aligned} \quad (8b)$$

where Var denotes variance. A unidimensional latent trait model for continuous item responses is the model for congeneric measurements [6], which is Spearman's one-factor model applied to item responses [12]. The one-factor model specifies three item parameters: the intercept (difficulty or attractiveness), loading (discrimination), and residual variance. The definition of a non-DIF item (8) is satisfied if the parameters of the reference and focal group are equal; null hypotheses on the equality of item parameters can be tested using multiple group **structural equation modeling (SEM)** methods [7].

The Underlying Variable Approach

The previously mentioned DIF detection methods are based on IRT models that directly apply to the observed item responses. However, within the SEM context, an indirect approach to dichotomous and ordinal-polytomous variables is known [7]. It is assumed that a continuous variable underlies a discrete variable and that this continuous variable is divided into two or more categories. Spearman's one-factor model is applied to the underlying response variables, and the equality of the factor model item parameters across the reference and focal group is tested [17].

Test Purification

The null hypothesis of equal item parameters is tested for an item that is studied for DIF. However, the item parameters of the other $n - 1$ test (questionnaire) items are constrained to be equal across the two

groups. A conceptual problem of DIF detection is that the constraints may be incorrect for one or more of the $n - 1$ other items. This problem is handled by using a stepwise or anchor test method. The stepwise procedure purifies the test (questionnaire) by removing DIF items in steps [10]. The anchor test method allocates a part of the n items to an anchor test, which is free of DIF [25]. The bottleneck of the anchor test method is the choice of the anchor items because these items must be free of DIF. Classical (test theory-based) DIF detection methods can be used to select non-DIF anchor items (Mellenbergh, this volume).

Acknowledgment

The author thanks Ineke van Osch for her assistance in preparing this entry.

References

- [1] Angoff, W.H. (1993). Perspectives on differential item functioning methodology, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 3–23.
- [2] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 397–479.
- [3] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory*, Kluwer-Nijhoff, Boston.
- [4] Hanson, B.A. (1998). Uniform DIF and DIF defined by difference in item response function, *Journal of Educational and Behavioral Statistics* **23**, 244–253.
- [5] Hessen, D.J. (2003). *Differential Item Functioning: Types of DIF and Observed Score Based Detection Methods*, Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.
- [6] Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests, *Psychometrika* **36**, 109–133.
- [7] Kaplan, D. (2000). *Structural Equation Modeling*, Sage Publications, Thousand Oaks.
- [8] Kelderman, H. (1989). Item bias detection using loglinear IRT, *Psychometrika* **54**, 681–697.
- [9] Lord, F.M. (1977). A study of item bias, using item characteristic curve theory, in *Basic Problems in Cross-Cultural Psychology*, Y.H. Poortinga, ed., Swets & Zeitlinger Publishers, Amsterdam, pp. 19–29.
- [10] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.
- [11] Mellenbergh, G.J. (1989). Item bias and item response theory, *International Journal of Educational Research* **13**, 127–143.
- [12] Mellenbergh, G.J. (1994). A unidimensional latent trait model for continuous item responses, *Multivariate Behavioral Research* **29**, 223–236.
- [13] Millsap, R.E. & Everson, H.T. (1993). Methodology review: statistical approaches for assessing measurement bias, *Applied Psychological Measurement* **17**, 297–334.
- [14] Millsap, R.E. & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias, *Applied Psychological Measurement* **16**, 389–402.
- [15] Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis with Applications in Political Research*, Walter de Gruyter/Mouton, New York.
- [16] Mokken, R.J. (1997). Nonparametric models for dichotomous responses, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer, New York, pp. 351–367.
- [17] Muthén, B. & Lehman, J. (1985). Multiple group IRT modeling: applications to item bias analysis, *Journal of Educational Statistics* **10**, 133–142.
- [18] Oort, F.J. (1993). Theory of violators: assessing unidimensionality of psychological measures, in *Psychometric Methodology, Proceedings of the 7th European Meeting of the Psychometric Society in Trier*, R. Steyer, K.F. Wender & K.F. Widaman, eds, Fischer, Stuttgart, pp. 377–381.
- [19] Raju, N.S. (1988). The area between two item characteristic curves, *Psychometrika* **53**, 495–502.
- [20] Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement* **14**, 197–207.
- [21] Raju, N.S., Van der Linden, W.J. & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests, *Applied Psychological Measurement* **19**, 333–368.
- [22] Shealy, R.T. & Stout, W.F. (1993). An item response theory model for test bias and differential test functioning, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 197–239.
- [23] Sijtsma, K. & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*, Sage Publications, Thousand Oaks.
- [24] Thissen, D., Steinberg, L. & Gerrard, M. (1986). Beyond group mean differences: the concept of item bias, *Psychological Bulletin* **99**, 118–128.
- [25] Thissen, D., Steinberg, L. & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines, in *Test Validity*, H. Wainer & H. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 147–169.
- [26] Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 67–113.
- [27] Van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

- [28] Wainer, H. (1993). Model-based standardized measurement of an item's differential impact, in *Differential Item Functioning*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 123–135.

(See also **Item Bias Detection: Classical Approaches**)

GIDEON J. MELLENBERGH

Item Exposure

FREDERIC ROBIN

Volume 2, pp. 974–978

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Exposure

Major professional organizations [1, 3] clearly stress in their guidelines that, to ensure the validity of test scores, it is critical that all aspects of test security be taken into account as tests are developed and delivered. According to these guidelines, three principal aspects of test security need to be addressed. First, test materials (e.g., test booklets and their associated keys stored in paper or electronic form) must be kept secure throughout test development and administrations, until they are released to the public. Second, cheating must be prevented during test administration. And third, the possibility for examinees to gain specific knowledge about items to be used in future administrations, through past administrations (and possibly pretesting), must be minimized. The focus of this entry is on this third aspect of test security.

Consider item exposure (or simply exposure), which reflects the extent to which an item has been used in past administrations, as an indicator of how much risk there is that it has been compromised for future administrations. That risk depends on future examinees' ability, individually through past testing experience or collectively through organized or informal networks, to recollect and gather detailed information on items used or reused in past administrations. It also depends on the extent to which future exposure is predictable, allowing examinees to focus their preparation on items identified as likely to appear in the test and even memorize answers keys, and as a result, obtaining scores significantly different from the ones they would have obtained under the standardized conditions test users expect. Worse, coaching organizations or dedicated internet web sites may actually engage in concerted efforts to gain and disseminate highly accurate information to their clients and distort testing outcomes for small or large subpopulations.

Of course, the most effective way to avoid this problem is not to reuse items at all. However, increasingly, testing clients are demanding greater scheduling flexibility and shorter testing time, which are generally attainable only through higher administration frequency and more efficient tests designs. Unfortunately, for many testing programs, this strategy is not viable. Given that the amount of work and the costs involved in developing new items are largely

incompressible, increasing testing frequency (resulting in fewer examinees per administration to share the costs) quickly results in resource requirements far exceeding what an organization or its clients can afford. And, although shorter test length could in principle decrease item needs, in practice this has not happened because to maintain measurement accuracy tests have to rely more heavily on highly informative items that have to be 'cherry picked.'

When item reuse is necessary, maintaining test security requires: (a) the establishment of acceptable item exposure limits, (b) the implementation of effective test design and development procedures to maintain exposure below the acceptable limits, (c) the monitoring of item-measurement properties and examinee scores over administrations, and (d) the monitoring of examinee information networks (e.g., internet websites) for evidence of security breach.

Although no firm guideline exists for determining acceptable exposure limits, values used by existing programs may be considered [5] and item memorization or 'coachability' studies may be conducted. Once in operation, these limits may be further established as monitoring information becomes available.

This entry provides a brief description of measures useful for monitoring item exposure and evaluating test security. Then it discusses alternative traditional linear (TL), multistage (MST) or **computerized adaptive testing (CAT)** test design choices as they relate to exposure. And finally, it outlines the methodologies that can be employed to maintain acceptable item exposure rates.

Evaluation and Monitoring of Item Exposure

The most obvious way to measure item exposure is to keep track of the total number of times an item has been administered. However, this measure alone may not be a sufficient indicator of the security risk an item presents for future test administrations. Additional measures, including exposure rate, conditional exposure rate and test overlap, are also useful for deciding if an item can be safely included in new test forms, if it should be set aside or 'docked' for a period of time before it is made available for testing again, or if it is time to retire it. On the basis of these measures, operational item-screening rules

2 Item Exposure

and algorithms can be devised to effectively manage the operational item bank and assemble new test forms [8].

Let's call X the total number of times an item has been administered. If X_M represents the maximum acceptable number of exposures, then item j should be retired whenever $X_j > X_M$.

Furthermore, an item that tends to be seen by larger proportions of examinees is more likely to become known in greater detail even before it reaches X_M than another one seen by only few of the examinees taking the test. Considering n_j the number of examinees who have seen item j and N the total number of examinees who have been tested during a given time period, the item's exposure rate for that time period is simply: $x_j = n_j/N$. If x_M represents the maximum acceptable exposure rate over the last time period, then item j should be docked whenever $x_j > x_M$.

To illustrate how these measures and criteria may be applied, let us imagine three items and follow their exposure over six administration cycles, during each of which tests are drawn from a new item pool following a typical CAT design. Figure 1 displays the three-item exposure rates and the maximum acceptable limit x_M . It shows that during the first administration cycle, item 3 was highly exposed at a 0.6 rate, coming close to the maximum acceptable exposure, say $X_M = 7000$, as a result of 10 000 administrations. Because of that, the item was retired. Item 2 was exposed at a much lower rate: first below $x_M = 0.3$, then slightly above x_M and consequently docked for the next cycle, then still slightly above x_M reaching X_M to be finally retired. Item 1, exposed at an even lower rate, was used throughout.

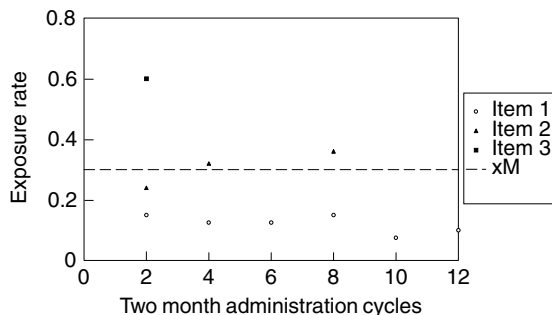


Figure 1 Exposure rates and maximum acceptable limit for three items

Conditional exposure and test-overlap measures may also be needed in order to effectively manage test security [5]. Conditional measures in particular may be necessary for programs using adaptive testing (CAT or MST). Because adaptive testing seeks to match item difficulty with examinee performance (oftentimes through maximizing item information), for some items, exposure can become very high among examinees of similar ability while their overall exposure is below the acceptable limit or even quite low. Moreover, the security risk associated with high conditional exposure amongst, for example, the most able examinees may not be limited to that group as information sharing is likely to go across ability groups. Although there is the added complexity of having to evaluate exposure across multiple groups, conditional exposures can be evaluated essentially the same way overall measures are.

Concerning test overlap, if the proportion of items shared between any two test forms is large, then examinees who retake the test may be unduly advantaged as they may have already experienced a large proportion of the test items. However, overlap can be tedious and difficult to keep track of. Fortunately, Chen and Ankenmann [2] have shown that there is a very strong relationship between the easy-to-measure exposure rates and test overlap and they have provided an easy way to estimate overall or conditional test overlap based on overall or conditional exposure rates. From a practical point of view, it might be more meaningful to set item test overlap rather than item exposure limits. Therefore, overlap may be used instead of, or as a complement to, exposure for evaluating, implementing, and monitoring test security.

Test Design and Item Exposure

Options for test design are numerous. Choosing an appropriate test design involves evaluating alternatives and making trade-offs between many competing goals, one of which is maintaining test security through exposure control. The purpose of this section is to briefly review various aspects of test design as they may facilitate or impede exposure control.

From this point of view, the most important aspects of test design include its administration schedule, length, and degree of adaptation to examinees. The administration schedule may be defined by

its frequency and by the number of alternative forms used during each administration occasion or cycle. On the one end, occasional administration of a single test form (at say a couple of dates per year) may not require any item reuse. On the other end, continuous administration with as many forms as examinees assembled from the same item pool is only possible with item reuse. However, it is worth noticing that higher administration frequency, in addition to providing scheduling flexibility, may also allow for more effective exposure management as items use may be spread across many more forms and their future rendered less predictable. This is discussed further in the next section.

The situation with test length is less intuitive. *A priori* a smaller test length would ease item demand and reduce the need for item reuse. But, experience has shown that significant length reduction tends to greatly increase the demand on the most informative items (in order to maintain test-reliability requirements) and practically disqualify large numbers of medium- to low-information items from being usable. Unchecked, this can lead to both the overexposure and depletion of the most sought after items, jeopardizing not only the security but also the psychometric quality of the tests produced.

Test adaptation is another important aspect of test design. By tailoring tests to examinees, there is some potential for more secure and more efficient testing. Indeed, with the traditional linear design (nonadaptive), test forms have to include items that cover a wide range of abilities, including items that after the fact will be of little value for the estimation of the examinee's score. With CAT, the increasingly precise ability estimations that can be made as testing progresses can be used to make more judicious item selections. By avoiding items too easy or too difficult, items for one examinee but useful to other examinees, CAT and to a certain extent MST may in fact provide greater testing flexibility and reduce testing time without the need for more item development.

Methods for Controlling Item Exposure

For any given design, test development can be viewed as a three-step process, with exposure control measure taking part in each one of these steps. In the first step, newly developed items are added to

the operational bank and existing items are screened through the application of the docking and retirement rules (as discussed earlier). With the TL test design, in the second step, a specified number of parallel forms are assembled. And finally, the assembled forms are randomly administered to examinees until the next test development cycle. In that case, either or both more stringent bank-screening rules and larger number of forms will lead to reduced item exposures. Also, because forms are assembled in advance of administration, there is no interaction between examinee performance and test assembly; thus, conditional exposure is not an issue.

With adaptive testing, in the second step, instead of test forms, one or more item pools are assembled. And, finally from each randomly assigned pool tests are assembled and administered (generally on-demand but not necessarily) to examinees. As with linear testing, exposure can be controlled primarily through bank management (first step) and through the use of multiple parallel pools. However, because pools generally need to be large, the number of parallel pools that can be used within the same administration cycle is often very limited. An additional challenge with adaptive testing may also arise from its emphasis on shortening test length and maximizing measurement efficiency. Many of the adaptive test assembly algorithms that have been developed tend to rely heavily on the most informative items available resulting in only a small proportion of any pool being used (exposed) at a high rate overall and at very high rates conditionally.

In assembling tests from an item pool, two main approaches have been developed in order to strike a balance between the need for exposure control and the desire for measurement efficiency. The first approach is to partition (stratify) the pool to create subsets of items with similar characteristics that can be selected interchangeably (e.g., through a purely random process). In that way, exposure is spread evenly among all the items available within each subgroup and exposure rates are determined by the subgroup size and the number of times, within each test assembly, items are selected from the subgroup. However, effective implementation may require a relatively complex partitioning of the pool (e.g., based on item discrimination and difficulty as well as item content). Yi and Chang [9] provide further discussion and an example of implementation of this approach.

The second approach is based on a probabilistic control of the item selection during test assembly. With that approach, the selection algorithm first determines the best items available to maximize measurement efficiency. Then, starting with the best item, a random number is drawn and compared with the item probability parameter resulting in either its rejection (without replacement) and the trial selection of the next best item or its selection; the selection being done with the first accepted item. Before operational use, this approach requires the repeated simulations of large number of tests in order to find, by successive adjustment, the item exposure control parameters that will result in exposures below the desired limits. Stocking and Lewis [5] and van der Linden [7] provide further discussions and examples of implementation of this approach.

Conclusion

For most testing programs, ensuring appropriate exposure control without compromising measurement requires choices of appropriate test design and exposure control mechanisms as well as significant test development efforts renewed over each test administration cycle. For further information, the two recent and complementary books, 'Computerized Adaptive Testing: Theory and Practice' [6] and 'Computer-based Testing: Building the Foundation for Future Assessment' [4], are primary sources to be consulted as they provide broad and practical as well as in-depth treatment of test development and exposure control issues.

References

- [1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.
- [2] Chen, S., Ankenmann, R.D. & Spray, J.A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing, *Journal of Educational Measurement* **40**, 129–145.
- [3] International Test Commission. (draft version 0.5, March 2004). International guidelines on computer-based and internet delivered testing [On-line], Available: http://www.intestcom.org/itc_projects.htm
- [4] Mills, C.N., Potenza, M.T., Fremer, J.J. & Ward, W.C., eds (2002). *Computer-based Testing: Building the Foundation for Future Assessment*, Lawrence Erlbaum Associates, Mahwah.
- [5] Stocking, M.L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing, *Journal of Educational and Behavioral Statistics* **23**, 57–75.
- [6] van der Linden, W.J. & Glas, C.A.W., eds (2000). *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers, Boston.
- [7] van der Linden, W.J. & Veldkamp, B.P. (2004). Constraining item-exposure in computerized adaptive testing with shadow tests, *Journal of Educational and Behavioral Statistics* **29**, 273–291.
- [8] Way, W.D., Steffen, M. & Anderson, G.S. (2002). Developing, maintaining, and renewing the item inventory to support CBT, in *Computer-based Testing: Building the Foundation for Future Assessment*, C.N. Mills, M.T. Potenza, J.J. Fremer & W.C. Ward, eds, Lawrence Erlbaum Associates, Mahwah, pp. 143–164.
- [9] Yi, Q. & Chang, H.H. (2003). A-stratified CAT design with content blocking, *British Journal of Mathematical and Statistical Psychology* **56**, 359–378.

FREDERIC ROBIN

Item Response Theory (IRT): Cognitive Models

AMY B. HENDRICKSON AND ROBERT J. MISLEVY

Volume 2, pp. 978–982

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Response Theory (IRT): Cognitive Models

That **item response theory (IRT)** was called latent trait theory into the 1980s reflects its origins in trait psychology. By modeling response probabilities at the level of items rather than test scores, IRT provided practical solutions to thorny problems in classical test theory, including item banking, adaptive testing, and linking tests of different lengths or difficulties. But operationally, the person parameter in IRT models (usually denoted θ) was essentially the same as that of true score in classical test theory, namely, a propensity to perform well in some domain of tasks. The meaning of that propensity lay outside IRT, and was inferred from the content of the tasks.

The ‘cognitive revolution’ of the 1960s and 1970s, exemplified by Newell and Simon’s *Human Information Processing* [20], called attention to the nature of knowledge, and the ways in which people acquire and use it. How do people represent the information in a situation? What operations and strategies do they use to solve problems? What aspects of problems make them difficult, or call for various kinds of knowledge or processes? Researchers such as Carroll [4] and Sternberg [27] studied test items as psychological tasks. Others, including Whitely [29]¹ and Klein et al. [15] designed aptitude and achievement test items around features motivated by theories of knowledge and performance in a given domain.

Several psychometric models have been introduced to handle claims cast in information-processing terms, explicitly modeling performance in terms of theory-based predictions of performance. Cognitively based IRT models incorporate features of items that influence persons’ responses, and relate these features, through θ , to response probabilities. Examples of IRT models that incorporate cognitively motivated structures concerning task performance are presented below (see [13] for a more detailed discussion of building and estimating such models).

Measurement Models

In a basic IRT model for dichotomous items, θ represents person proficiency in the domain, and the observable variables ($X_1, \dots, X_n$) are responses to

n test items. Possibly vector-valued item parameters ($\beta_1, \dots, \beta_n$) determine the conditional probabilities of the X s given θ . Under the usual IRT assumption of **conditional independence**,

$$P(x_1, \dots, x_n | \theta, \beta_1, \dots, \beta_n) = \prod_{j=1}^n P(x_j | \theta, \beta_j). \quad (1)$$

Under the **Rasch model** [22] for dichotomous items, for example,

$$P(x_j | \theta, \beta_j) = \begin{cases} \Psi(\theta - b_j) & \text{if } x_j = 1 \\ 1 - \Psi(\theta - b_j) & \text{if } x_j = 0 \end{cases}, \quad (2)$$

where $\Psi(\cdot) \equiv \exp(\cdot) / [1 + \exp(\cdot)]$ is the cumulative logistic probability distribution, θ is a one-dimensional measure of proficiency, b_j is the difficulty of item j , and x_j is 1 if right and 0 if wrong. This formulation does not address the nature of θ , the rationale for the difficulty of item j , or the connection between the two.

Linear Logistic Test Model (LLTM)

In the Linear Logistic Test Model (LLTM; [6, 25]), cognitively based features of items and persons’ probabilities of response are related through a Q -matrix [28], where q_{jk} indicates the degree to which feature k applies to item j . In simple cases, q_{jk} is 1 if feature k is present in item j and 0 if not. The LLTM extends the Rasch model by positing a linear structure for the b_j s:

$$b_j = \sum_k q_{jk} \eta_k = q'_j \eta, \quad (3)$$

where η_k is a contribution to item difficulty entailed by feature k . Features can refer to a requirement for applying a particular skill, using a particular element of information, carrying out a procedure, or some surface feature of an item.

Fischer [6] used the LLTM to model the difficulty of multistep calculus items, as a function of how many times each of seven differentiation formulas had to be applied. Hornke and Habon [9] generated progressive matrix items in accordance with the rules of the patterns across stimuli, and modeled item difficulty in these terms. Sheehan and Mislevy [26] fit an LLTM-like model with random item effects, using features based on Mosenthal and Kirsch’s [17] cognitive analysis of the difficulty of document

literacy tasks. The features concerned the structure of a document and the complexity of the task to be performed. In each of these applications, any given value of θ can be described in terms of expected performance in situations described by their theoretically relevant features.

Multidimensional Cognitive IRT Models

Providing theoretically derived multidimensional characterizations of persons' knowledge and skills is called cognitive diagnosis [21]. Most cognitively based **multidimensional item response theory (IRT) models** posit either compensatory or conjunctive ('noncompensatory') combinations of proficiencies.

Compensatory Models. In compensatory models, proficiencies combine linearly, so a lack in one proficiency can be made up with an excess in another proficiency, that is, $a'_j\theta = a_{j1}\theta_1 + \dots + a_{jD}\theta_D$, where θ is a D -dimensional vector. The a_{jd} s indicate the extent to which proficiency d is required to succeed on item j . The A -matrix is analogous in this way to the Q -matrix decomposition of examinee proficiencies. A is estimated in some models [1], but in applications more strongly grounded in cognitive theory they are treated as known, their values depending on the knowledge and skill requirements that have been designed into each item. For example, Adams et al.'s [2] Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) posits known a_{dk} s and q_{jk} s. The probability of a correct response for a dichotomous item is

$$\Pr(X_j = 1 | \theta, \eta, a'_j, q'_j) = \Psi(a'_j\theta + q'_j\eta). \quad (4)$$

When the MRCMLM is applied to polytomous responses, each response category has its own a and q vectors to indicate which aspects of proficiency are evidenced in that response and which features of the category context contribute to its occurrence. Different aspects of proficiency may be involved in different responses to a given item, and different item features can be associated with different combinations of proficiencies. Adams et al. provide an example with a mathematical problem-solving test using the structure of learning outcome (SOLO) taxonomy; constructed responses are scored according to an increasing scale of cognitive complexity, and the θ s indicate persons' propensities to give explanations

that reflect these increasing levels of understanding. De Boeck and his students (e.g., [10, 12, 23]) have carried out an active program of research using such models to investigate hypotheses about the psychological processes underlying item performance.

Conjunctive Models. In conjunctive models, an observable response is the (stochastic) outcome of sub processes that follow unidimensional IRT models. Embretson [5] describes a multicomponent latent trait model (MLTM) and a general component latent trait model (GLTM) in which distinct subtasks are required to solve a problem, and the probability of a success in each depends on a component of θ associated with that subtask. An 'executive process' parameter e represents the probability that correct subtask solutions are combined appropriately to produce an overall success, and a 'guessing' parameter c is the probability of a correct response even if not all subtasks have been completed correctly. An example of the MLTM with $e = 1$ and $g = 0$ models the probability that an examinee will solve an analogy item as the product of succeeding on Rule Construction and Response Evaluation subtasks, each modeled by a dichotomous Rasch model:

$$P(x_{jT} = 1 | \theta_1, \theta_2, \beta_{j1}, \beta_{j2}) = \prod_{m=1}^2 \Psi(\theta_m - \beta_{jm}), \quad (5)$$

where x_{jT} is the overall response, θ_1 and β_{j1} are examinee proficiency and Item j 's difficulty with respect to Rule Construction, and θ_2 and β_{j2} are proficiency and difficulty with respect to Response Evaluation. Under the GLTM, both subtask difficulty parameters β_{j1} and β_{j2} can be further modeled in terms of item features as in the LLTM.

Hybrids of IRT and Latent Class Models

Latent class models are not strictly IRT models because their **latent variables** are discrete rather than continuous. But some cognitively based IRT models incorporate a particular latent class model called 'binary skills' models [7, 16]. In a binary skills model, there is a one-to-one correspondence between the features of items specified in a Q -matrix and elements of peoples' knowledge or skill, or 'cognitive attributes'. A person is likely to answer item j

correctly if she possesses all the attributes an item demands and incorrectly if she lacks some. Junker and Sijtsma [14] describe two ways to model these probabilities: DINA (Deterministic Inputs, Noisy And) and NIDA (Noisy Inputs, Deterministic And). Under DINA, the stochastic model comes after the combination of the attributes; a person with all necessary attributes responds correctly with probability π_{j1} , and a person lacking some attributes responds correctly with probability π_{j0} :

$$\begin{aligned} \Pr(X_{ij} = 1 | \theta_i, q_j, \pi_{j1}, \pi_{j0}) \\ = \pi_{j0} + (\pi_{j1} - \pi_{j0}) \prod_k \theta_{ik}^{q_{jk}}, \end{aligned} \quad (6)$$

where θ_{ik} is 1 if person i possesses attribute k and 0 if not. Under NIDA, there is a stochastic mechanism associated with each attribute; if attribute k is required by item j (i.e., $q_{jk} = 1$), the contribution to person i being able to solve the item will require either her having attribute k or, if not, circumventing the requirement with probability r_{jk} :

$$\Pr(X_{ij} = 1 | \theta_i, q_j, r_j) = \prod_k r_{jk}^{(1-\theta_{ik})q_{jk}}. \quad (7)$$

Latent variable models for cognitive diagnosis such as DINA and NIDA models have the aim of diagnosing the presence or absence of multiple fine-grained skills required for solving problems. In other words, does the examinee show ‘mastery’ or ‘nonmastery’ of each skill?

Two cognitively based models that combine features of IRT models and binary skills models are the Rule Space model [28] and the Fusion model [8]. Under Rule Space, response vectors are mapped into a space of IRT estimates and fit indices. This includes response patterns predicted by particular configurations of knowledge and strategy – both correct and incorrect rules for solving items – as characterized by a vector of cognitive attributes. The distance of a person’s observed responses from these ideal points are estimated, and the best matches suggest that person’s vector of attributes. Under the Fusion model, a NIDA model is posited for the effect of attributes, but an IRT-like upper bound is additionally proposed; a maximum probability π_{j1} times a probability that depends on a continuous person proficiency parameter ϕ_i and item difficulty b_j that concern knowledge and ability other than those modeled in terms of

attributes:

$$\begin{aligned} \Pr(X_{ij} = 1 | \theta_i, \phi_i, \pi_{j1}, q_j, r_j, b_j) \\ = \pi_{j1} \text{Logit}(\theta_i - b_j) \prod_k r_{jk}^{(1-\theta_{ik})q_{jk}}. \end{aligned} \quad (8)$$

Mixtures of IRT Models and Multiple Strategies

Different students can bring different problem-solving strategies to an assessment setting. Further, comparisons of and theories concerning experts’ and novices’ problem-solving suggest that the sophistication with which one chooses and monitors strategy use, develops as expertise grows. Thus, strategy use is a potential target of inference in assessment. Junker [13] distinguishes four cases for modeling differential strategy use

Case 0: No modeling of strategies (basic IRT models).

Case 1: Model strategy changes between persons.

Case 2: Model strategy changes between tasks within persons.

Case 3: Model strategy changes within task within persons.

Multiple-strategy models incorporate multiple $\mathcal{Q}$ -matrices in order to specify the different strategies that are used to solve the test items. IRT mixture models have been proposed for Case 1 by Rost [24] and Mislevy and Verhelst [18]. Consider the case of M strategies; each person applies one of them to all items, and the item difficulty under strategy m depends on features of the task that are relevant under this strategy in accordance with an LLTM structure, so that the difficulty of item i under strategy m is $b_{jm} = \sum_k q_{jmk} \eta_{mk}$. Denoting the proficiency of person i under strategy m as θ_{im} and ϕ_{im} as 1 if she uses strategy m and 0 if not, the response probability under such a model takes a form such as

$$\begin{aligned} \Pr(X_{ij} = 1 | \theta_i, \phi_i, q_j, \Gamma) \\ = \prod_m \phi_i \left[\Psi \left(\theta_{im} - \sum_k q_{jmk} \eta_{mk} \right) \right]^{x_{ij}} \\ \times \left[1 - \Psi \left(\theta_{im} - \sum_k q_{jmk} \eta_{mk} \right) \right]^{1-x_{ij}} \end{aligned} \quad (9)$$

Huang [11] used a Rasch model studied by Andersen [3] for an analysis under Case 2. His example concerns three approaches to force and motion

problems: Newtonian, impetus theory, and nonscientific response. The response of Examinee i to item j is coded as 1, 2, or 3 for the approach used. Each examinee is characterized by three parameters θ_{ik} , indicating propensities to use each approach, and each item is characterized by three parameters indicating propensities to evoke each approach. Strategy choice is modeled as

$$\Pr(X_{ij} = k | \theta_i, \beta_j) = \frac{\exp(\theta_{ik} - \beta_{jk})}{\sum_{m=1}^3 \exp(\theta_{im} - \beta_{jm})}. \quad (10)$$

Conclusion

As assessments address a more diverse and ambitious set of purposes, there comes an increasing need for inferences that can address complex task performance and instructionally useful information. Cognitively based assessment integrates empirically based models of student learning and cognition with methods for designing tasks and observing student performance, with procedures for interpreting the meaning of those observations [19].

Note

1. S.E. Whitely now publishes as S.E. Embretson.

References

- [1] Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring, *Applied Measurement in Education* **7**, 255–278.
- [2] Adams, R., Wilson, M.R. & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model, *Applied Psychological Measurement* **21**, 1–23.
- [3] Andersen, E.B. (1973). *Conditional Inference and Models for Measuring*, Danish Institute for Mental Health, Copenhagen.
- [4] Carroll, J.B. (1976). Psychometric tests as cognitive tasks: a new structure of intellect, in *The Nature of Intelligence*, L.B. Resnick, ed., Erlbaum, Hillsdale, pp. 27–57.
- [5] Embretson, S.E. (1985). A general latent trait model for response processes, *Psychometrika* **49**, 175–186.
- [6] Fischer, G.H. (1973). Logistic latent trait models with linear constraints, *Psychometrika* **48**, 3–26.
- [7] Haertel, E.H. & Wiley, D.E. (1993). Representations of ability structures: implications for testing, in *Test Theory for a New Generation of Tests*, N. Frederiksen, R.J. Mislevy & I.I. Bejar, eds, Erlbaum, Hillsdale, pp. 359–384.
- [8] Hartz, S.M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality*, Doctoral dissertation, Department of Statistics, University of Illinois Champaign-Urbana.
- [9] Hornke, L.F. & Habon, M.W. (1986). Rule-based bank construction and evaluation within the linear logistic framework, *Applied Psychological Measurement* **10**, 369–380.
- [10] Hoskins, M. & De Boeck, P. (1995). Componential IRT models for polytomous items, *Journal of Educational Measurement* **32**, 364–384.
- [11] Huang, C.-W. (2003). *Psychometric analyses based on evidence-centered design and cognitive science of learning to explore students' problem-solving in physics*, Doctoral dissertation, University of Maryland, College Park.
- [12] Janssen, R. & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks, *Applied Psychological Measurement* **21**, 37–50.
- [13] Junker, B.W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment, in *Paper prepared for the Committee on the Foundations of Assessment, National Research Council*.
- [14] Junker, B.W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory, *Applied Psychological Measurement* **25**, 258–272.
- [15] Klein, M.F., Birenbaum, M., Standiford, S.N. & Tatsuka, K.K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions*, Research Report 81–86, Computer-based Education Research Laboratory, University of Illinois, Urbana.
- [16] Maris, E. (1999). Estimating multiple classification latent class models, *Psychometrika* **64**, 187–212.
- [17] Mosenthal, P.B. & Kirsch, I.S. (1991). Toward an explanatory model of document process, *Discourse Processes* **14**, 147–180.
- [18] Mislevy, R.J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies, *Psychometrika* **55**, 195–215.
- [19] National Research Council. (2001). *Knowing what Students Know: The Science and Design of Educational Assessment*, Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser & N. Chudowsky, eds, National Academy Press, Washington.
- [20] Newell, A. & Simon, H.A. (1972). *Human Problem Solving*, Prentice-Hall, Englewood Cliffs.
- [21] Nichols, P.D., Chipman, S.F. & Brennan, R.L., eds (1995). *Cognitively Diagnostic Assessment*, Erlbaum, Hillsdale.
- [22] Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute

-
- for Educational Research, Copenhagen/University of Chicago Press, Chicago (reprint).
- [23] Rijmen, F. & De Boeck, P. (2002). The random weights linear logistic test model, *Applied Psychological Measurement* **26**, 271–285.
- [24] Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis, *Applied Psychological Measurement* **14**, 271–282.
- [25] Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben. (The learning and solution of complex cognitive tasks.), *Zeitschrift für Experimentelle und Angewandte Psychologie* **19**, 476–506.
- [26] Sheehan, K.M. & Mislevy, R.J. (1990). Integrating cognitive and psychometric models in a measure of document literacy, *Journal of Educational Measurement* **27**, 255–272.
- [27] Sternberg, R.J. (1977). *Intelligence, Information-processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*, Erlbaum, Hillsdale.
- [28] Tatsuka, K.K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory, *Journal of Educational Measurement* **20**, 345–354.
- [29] Whitely, S.E. (1976). Solving verbal analogies: some cognitive components of intelligence test items, *Journal of Educational Psychology* **68**, 234–242.

AMY B. HENDRICKSON AND
ROBERT J. MISLEVY

Item Response Theory (IRT) Models for Dichotomous Data

RONALD K. HAMBLETON AND YUE ZHAO

Volume 2, pp. 982–990

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Response Theory (IRT) Models for Dichotomous Data

Classical test theory (CTT) (*see* **Classical Test Models**) models have been valuable to test developers and researchers for over 80 years, but today, because of several important shortcomings, they are being replaced by item response theory (IRT) models in measurement practices. In this entry, several shortcomings of classical test theory are highlighted, and then IRT and several promising IRT models for the analysis of dichotomously scored item response data are introduced. Brief discussions of IRT model parameter estimation, model fit, and software are described. Finally, several important applications are introduced.

Strengths and Weaknesses of Classical Test Theory

CTT has been used in the test development field for over 80 years. The testing and assessment field is full of examples of highly reliable and valid tests based on CTT models and principles, and these tests have been used to produce research findings in thousands and thousands of studies (*see* **Classical Test Models**). References to the CTT model with a focus on true score, observed score, and error score, the use of item difficulty (P values) and item discrimination indices (r values) in test development, corrected split-half reliabilities and coefficient alphas, applications of the Spearman–Brown formula and the formula linking test length and validity, corrections to correlations for range restriction, the standard error of measurement, and much more (*see*, for example, [3, 8]) are easily found in the psychometric methods literature.

There will be no bashing in this entry of classical test theory and common approaches to test development and validation. Were these approaches to be used appropriately by more test developers, tests would be uniformly better than they are today, and the quality of research using educational and psychological tests would be noticeably better too.

At the same time, it is clear that classical test theory and related models and practices have some

shortcomings, and so they are not well suited for some of the demands being placed on measurement models today by two innovations: item banking and **computer adaptive testing** (*see*, for example, [1, 9, 11, 13]). One shortcoming is that classical item statistics are dependent on the particular choice of examinee samples. This shortcoming makes classical item statistics (such as item difficulty levels, and biserial and **point biserial correlations**) problematic in an item bank unless all of the test item statistics are coming from equivalent examinee sample, which of course, is highly unrealistic, and to make such a requirement would lower the utility of item banks in test development (the option must be available to add new test items over time to meet demands to assess new content, and to replace overused test items).

A second shortcoming is that examinee scores are highly dependent on the particular choice of items in a test. Give an ‘easy’ set of items to examinees and they will generally score high, and give a ‘hard’ set of items to them, and they will generally score lower. This dependence of test scores on items in a test creates major problems when computer-adaptive testing (CAT) is used. In principle, in a CAT environment (*see* **Computer-Adaptive Testing**), examinees will see items ‘pitched’ or ‘matched’ to their ability levels, and so, in general, examinees will be administered ‘nonparallel’ tests and the test scores themselves will not provide an adequate basis for comparing examinees to each other or even to a set of norms or performance standards set on a nonequivalent form or version of the test.

These two shortcomings are serious drawbacks to the use of classical test theory item statistics with item banking and computer-adaptive testing but there are more. Typically, classical test models provide only a single estimate of error (*i.e.*, the standard error of measurement) and it is applied to the test scores of all examinees. (It is certainly possible to obtain standard errors conditional on test score but this is rarely done in practice.) But this means that the single error estimate for a test is probably too large for the bulk of ‘middle ability examinees’ and too small for examinees scoring low or high on the ability scale. Also, CTT models the performance of examinees at the test score level (recall ‘test score = true score + error score’) but CAT requires modeling between examinee ability and items so that optimal item selections can be made. Finally, items and examinees are reported on separate, and noncomparable scales

in classical measurement (items are reported on a scale defined over a population of examinees, and examinees are reported on a scale defined for a domain of content). This makes it nearly impossible to implement optimal assessment where items are selected to improve the measurement properties of the test for each examinee, or a prior ability distribution.

IRT Models, Assumptions, and Features

Item response theory is a statistical framework for linking examinee scores to the items on a test to the trait or traits (usually called ‘abilities’) that are measured by that test. Abilities may be narrow or broad; they might be highly influenced by instruction (e.g., math skills) or more general human characteristics (e.g., cognitive functioning); and they might be in the cognitive, affective, or psychomotor domains. These abilities could be unidimensional or multidimensional too but only unidimensional abilities will be considered further in this entry (see **Multidimensional Item Response Theory Models**, and [12]).

A mathematical model must be specified that provides the ‘link’ (for example, the three-parameter logistic model) between these item scores and the trait or ability measured by the test. Nonlinear models have been found to better fit test data, and so there has rarely been interest in linear models, or at least not since about 1950. It is common to assume there is a single unidimensional ability underlying examinee item performance, but models for handling multiple abilities are readily available (see **Multidimensional Item Response Theory Models**, [12]). With a nonlinear model specified, and the examinee item scores available, examinee and item parameters can be estimated.

A popular IRT model for handling dichotomously scored items is the three-parameter logistic test model:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, i = 1, 2, \dots, n. \quad (1)$$

Here, $P_i(\theta)$ is the probability of an examinee providing the correct answer to test item i , and this nonlinear probability function of θ is called an *item characteristic curve* (ICC) or item characteristic function (ICF). The probability is defined either over administrations of items equivalent to test item i

to the examinee, or defined for a randomly drawn examinee with ability level θ . Item parameters in this model are denoted b_i , a_i , and c_i , and described by test developers as item difficulty, item discrimination, and pseudoguessing (or simply, guessing), respectively. The pseudoguessing parameter in the model is the height of the lower asymptote of the ICF introduced in the model to fit the nonzero performance of low performing examinees, perhaps due to random guessing. The b-parameter is located on the ability scale at a point corresponding to a probability of $(1 + c_i)/2$ of a correct response (thus, if $c = .20$, then the b-parameter corresponds to the point on the ability continuum where an examinee has a .60 probability of a correct response). The a-parameter is proportion to the slope of the ICF at the point b on the ability scale. The D is a constant in the model, and was introduced by Birnbaum (see [8]) so that the item parameters in the model closely paralleled the item parameters in the normal-ogive model [7], one of the first IRT models that was introduced into the psychometric methods literature in 1952. (This model was quickly dropped when other more tractable models became available.) ‘ n ’ is the number of items in the test. Figure 1 provides an example of an ICF, and highlights the interpretation of the model parameters.

Figure 2 shows the ICFs for five different test items.

From Figure 2, the impact of the item statistics on the shapes of the ICFs can be seen. The statistics for the five items are typical of those observed in

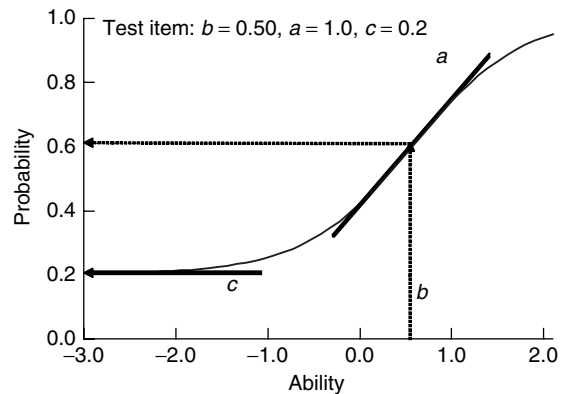


Figure 1 A typical item characteristic function for the three-parameter logistic model (see the S-shaped logistic function)

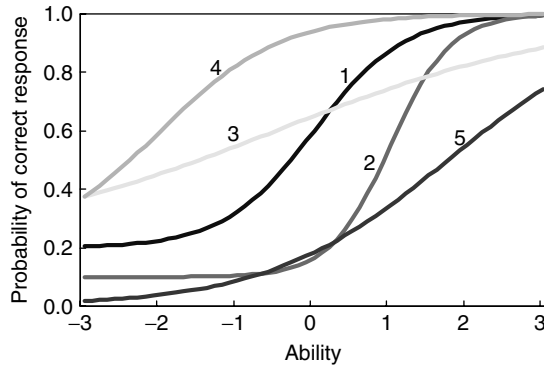


Figure 2 Five item characteristic functions

Item	b	a	c
1	0.00	1.00	0.20
2	1.00	1.50	0.10
3	-0.50	0.30	0.20
4	-2.00	0.75	0.20
5	1.75	0.50	0.00

practice. An item such as item 2, offers great potential for discriminating among examinees, especially among the more capable examinees. An item such as item 3 adds very little to the psychometric quality of the test because the power of the item to discriminate higher from lower ability is low. Item 5 is only a bit more useful. Sometimes, though, test items like items 3 and 5 are used because they assess content that needs to be covered in a test, and no better items, statistically, are available.

There are important special cases of the three-parameter logistic test model for analyzing dichotomously scored test item data. Setting $c = 0$ in the three-parameter logistic model produces the two-parameter logistic test model, a model that often fits free response items scored 0 or 1, or multiple-choice test items where guessing is minimal such as when the test items are easy. Setting $c = 0$ and $a = 1$, produces the one-parameter logistic model (and better known as the '**Rasch model**'), a model that was developed independently of the three-parameter model by Georg Rasch from Denmark in the late 1950s, and has become very common for the analysis of educational and psychological test data (see, for example, [2]).

Until about 15 years ago, most of the work with item response modeling of data was limited to models

that could be applied to dichotomously scored data – the one-, two-, and three-parameter logistic and normal-ogive models [4]. Even when the data were not dichotomous, it was common to recode the data to reduce it to a binary variable (e.g., with a five point scale, perhaps 0 to 2 would be recoded as 0, and 3 to 4 recoded as 1). This was an unfortunate loss of valuable information when estimating examinee ability.

Many nationally normed achievement tests (e.g., California Achievement Tests, Metropolitan Achievement Tests), university admissions tests (e.g., Scholastic Assessment Test, the Graduate Management Admissions Test, and the Graduate Record Exam) and many state proficiency tests and credentialing exams in the United States consist of multiple-choice tests that are scored 0–1. The logistic test models are typically used with these tests in calibrating test items, selecting items, equating scores across forms, and more.

Today, more use is being made of test data arising from new item types that use polytomous-scoring (i.e., item level data that are scored in more than two categories) such as rating scales for recording attitudes and values, and the scoring of writing samples and complex performance tasks (*see Computer-based Test Designs*, and [14]). Numerous IRT models are available today for analyzing polytomous response data. Some of these models were available in the late 1960s (see Samejima's chapter in [12]) but were either too complex to apply, were not sufficiently well developed (for example, parameter estimation was problematic), or software for applying the models was simply not available. These polytomous IRT response models are especially important in the education and psychology fields because more of the test data today is being generated from tasks that are being scored using multicategory scoring rubrics. Polytomous IRT models like the partial credit model, generalized partial credit model, nominal response model, and the graded response model can handle, in principle, an unlimited number of score categories, and these models have all the same features as the logistic test models introduced earlier (*see Item Response Theory (IRT) Models for Polytomous Response Data*).

The test characteristic function (TCF) given below,

$$E(X) = TCF(\theta) = \sum_{i=1}^n P_i(\theta), \quad (2)$$

provides an important link between true scores in CTT and ability scores in IRT. This relation is useful in explaining how examinees can score well on easy tests and less well on hard tests; and is definitely useful in test design because it provides a basis for predicting test score distributions, or mapping performance standards that are often set on the test score scale to the ability scale.

Trait scores, or ability scores, as they are commonly called, are often scaled to a mean of zero and a standard deviation of one for convenience, and without loss of generality. A normal distribution of ability is neither desired nor expected, nor are scores typically transformed to obtain such distribution. Before scores are reported to examinees, it is common to apply a linear transformation to the ability scores (scaled, 0,1) to place them on a more convenient scale (for example, a scale with mean = 100, SD = 10) that does not contain negative numbers and decimals. It is the test developer's responsibilities to conduct validity investigations to determine exactly what a test, and hence, the ability scores, are measuring.

IRT models are based on strong assumptions about the data: In the case of the logistic test models introduced earlier for analyzing dichotomously scored data, they are (a) the assumption of test unidimensionality and (b) the assumption that the ICFs match or fit the actual data [4]. Other models specify different assumptions about the data (see, for example, [12]). Failure to satisfy model assumptions can lead to problems – for example, expected item and ability parameter invariance may not be present, and using ICFs to build tests, when these functions do not match the actual data, will result in tests that will function differently in practice than expected. These two assumptions will be explored in more detail in the next section.

The primary advantage of IRT models is that the parameters of these models are 'invariant'. This means that examinee ability remains constant over various samples of test items from the domain of content measuring the ability of interest. The accuracy of the estimates may vary (e.g., giving easy items to high ability candidates results in more error in the estimates, than giving items that are more closely matched to the examinee's ability level) but the examinee ability parameter being estimated with each sample is the same. This is known as 'ability parameter invariance'. This is accomplished by considering item statistics along with examinee response

data in the ability estimation process. The property of ability parameter invariance, makes it possible, for example, to compare examinees, even though the tests they have taken may differ considerably in test difficulty.

In the case of item parameter invariance, item parameters are independent of the particular choice of examinee samples. This means that for examinees at a particular point on the ability scale, the probability of a correct response is the same, regardless of the subpopulation from which the examinees are sampled [4]. The item parameters, though not the estimates, are 'examinee sample invariant' and this property is achieved by taking into account the characteristics of the sample of examinees in the item parameter estimation process. This is an immensely useful feature in estimating item statistics because the test developer is freed more or less of the concern about the nature of the examinee sample in item calibration work. Good estimation of these ICFs still requires appropriately chosen samples (large enough to produce stable parameter estimates, and broad enough that these ICFs can be properly estimated – recall that they are nonlinear regression functions), but to a great degree the role that the ability levels of the examinee sample plays can be addressed in the calibration process, and what is produced are estimates of the item parameters that are more or less independent of the examinee samples.

IRT Model Parameter Estimation, Model Fit, and Available Software

Several methods are available for estimating IRT model parameters (see, for example, [4, 7]) including new Bayesian (*see Bayesian Item Response Theory Estimation*), and Monte-Carlo Markov-Chain (*see Markov Chain Monte Carlo Item Response Theory Estimation*) procedures (see, [12]). All of these methods capitalize on the principle of local independence, which basically states that a single dominant factor or ability influences item performance, and is expressed mathematically as

$$P(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta) P(U_2 | \theta) \dots P(U_n | \theta) \\ = \prod_{i=1}^n P(U_i | \theta). \quad (3)$$

Here, $U_1, U_2, \dots, U_n$ are the binary scored responses to the items, and n is the total number

of items in the test. In other words, the principle or assumption of local independence, which is a characteristic of all of the popular IRT models, and follows from the assumption of unidimensionality of the test, states that the probability of any particular examinee response vector (of item scores) is given by the product of the separate probabilities associated with those item scores. This means that an examinee's item responses are assumed to be independent of one another, and driven by the examinee's (unknown) ability level. If the item parameters are known, or assumed to be known, the only unknown in the expression is the examinee's ability, and so it is straightforward to solve the equation with one unknown, examinee ability. With maximum likelihood estimation, the value for θ that makes the likelihood of the data that was observed is found (logs of both sides of the equation are taken, the expression for the first derivative with respect to θ is found, and then that expression is set to zero, and the value of θ that solves the differential equation is used as the ability estimate). By working with second derivatives of the likelihood expression, a standard error associated with the ability estimate can be found. For each examinee, item response data is inserted and the resulting equation is solved to find an estimate of θ .

Normally, item parameter estimates are treated as known once they are obtained and placed in an item bank. At the beginning of a testing program, neither item parameters nor abilities are known and then a joint estimation process is used to obtain **maximum likelihood** item parameter estimates and ability estimates (see, [7], for example). Today, more complex estimation procedures are becoming popular and new software is emerging.

Several software packages are available for parameter estimation: BILOG-MG (www.ssicentral.com) can be used with the one-, two-, and three parameter logistic models; and BIGSTEPS (www.winsteps.com), CONQUEST (www.assess.com), LPCN-WIN (www.assess.com), and several other software programs can be used with the Rasch model. The websites provide substantial details about the software.

The assumption of test unidimensionality is that the items in the test are measuring a single dominant trait. In practice, most tests are measuring more than a single trait, but good model fit requires only a

reasonably good approximation to the unidimensionality assumption. One check on unidimensionality that sometimes is applied is this: From a consideration of the items in the test, would it be meaningful to report a single score for examinees? Is there a factor common to the items such as 'mathematics proficiency' or 'reading dimensionality'. Multidimensionality in a dataset might result from several causes: First, the items in a test may cluster into distinct groups of topics that do not correlate highly with each other. Second, the use of multiple item formats (e.g., multiple-choice questions, checklists, rating scales, open-ended questions) may lead to distinct 'method of assessment' factors. Third, multidimensionality might result from dependencies in the data. For example, if responses to one item are conditional on responses to others, multidimensionality is introduced. (Sometimes this type of dimensionality can be eliminated by scoring the set of related items as if it were a 'testlet' or 'super item'.)

Many methods are available to explore the dimensionality of item response data (see [4, 5]). Various reviews of several older methods have found them all in one way or another to have shortcomings. While this point may be discouraging, methods are available that will allow the researcher to draw defensible conclusions regarding unidimensionality. For example, linear factor analysis (e.g., **principal components analysis**), nonlinear factor analysis, and **multidimensional scaling** may be used for this purpose, though not without some problems at the interpretation stage. Further, it is recognized that few constructs are reducible to a strictly unidimensional form and that demonstration of a dominant single trait may be all that is reasonable. For example, using principal components analysis we would expect a dominant first factor to account for roughly 20% or more of the variance in addition to being several times larger than the second factor. Were these conditions met, the assumption of unidimensionality holds to a reasonable degree.

Descriptions of Four IRT Applications

Very brief introductions to four popular applications of IRT follow. Books by Bartram and Hambleton [1], Hambleton, Swaminathan, and Rogers [4], Mills, et al. [9], van der Linden and Glas [11], and Wainer [13] provide many more details on the applications.

Developing Tests

Two of the special features of IRT modeling are item and test information functions. For each item, a test item information function is available indicating where on the reporting scale an item is useful in estimating ability and how much it contributes to an increase in measurement precision. The expression for the item information function (see, [7]) is

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}. \quad (4)$$

Figure 3 provides the item information functions corresponding to the items shown in Figure 2 and identified earlier in the entry. Basically, items provide the most measurement information around their ‘b-value’ or level of difficulty and the amount of information depends on item discriminating power. It is clear from Figure 3 that items 2, 1, and 4 are the most informative, though at different places on the ability continuum. This is an important point, because it highlights that were interest in ability estimation focused at, say, the lower end of the continuum (perhaps the purpose of the test is to identify low performing examinees for a special education program), item 4 would be a considerably more useful item than items 1 and 2 for precisely estimating examinee abilities in the region of special interest on the ability scale. Items 3 and 5 provide very little information at any place on the ability continuum.

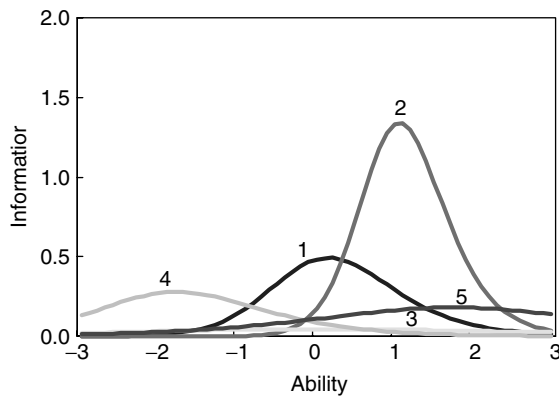


Figure 3 Five item information functions, corresponding to the ICFs reported in Figure 2

The location of maximum information provided by an item is given by the expression:

$$\theta_{i_{\max}} = b_i + \frac{1}{Da_i} \ln \left[0.5(1 + \sqrt{1 + 8c_i}) \right] \quad (5)$$

With the one- and two-parameter logistic models, the point of maximum information is easily seen to be b_i . With $c = 0.0$, the second term simply drops out of the expression.

The test information function (which is a simple sum of the information functions for items in a test) provides an overall impression of how much information a test is providing across the reporting scale:

$$I(\theta) = \sum_{i=1}^n I_i(\theta) \quad (6)$$

The more information a test provides at a point on the reporting scale, the smaller the measurement error will be. In fact, the standard error of measurement at a point on the reporting scale (called the *conditional standard error of measurement* or simply *conditional standard error*) is inversely related to the square root of the test information at that point:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (7)$$

Thus, a goal in test development is to select items such that the standard errors are of a size that will lead to acceptable levels of ability estimation errors.

A test information function is the result of putting a particular set of items into a test. Figure 4 provides

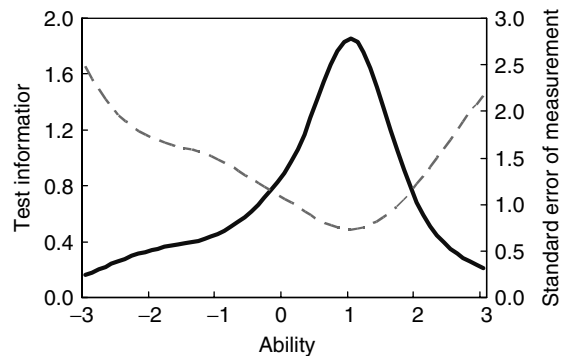


Figure 4 Test information function and corresponding standard errors based on the five items used in Figures 2 and 3

the test information function for the items shown in Figure 3, and the corresponding conditional error function for the five item test. In practice, the goal is usually to push test information function up to at least 10, which roughly corresponds to a classical reliability estimate of about .90.

Sometimes, rather than the test information function being a result of item selection and a product of the test development process, it is specified as the ‘target’ and then items can be selected to produce the test information function of interest. Item selection often becomes a task of selecting items to meet content specifications, and statistical specifications (as reflected in a ‘target information function’). One of the newest IRT topics (called ‘automated test assembly’) is the development of procedures for allowing test developers to define the test of interest in considerable detail, translate those specifications into mathematical equations, and then with the appropriate software, actually solve the resulting equations, and select test items from a bank of calibrated test items to meet the requirements for the test (see, [10, 11]).

Identifying DIF

The property of item parameter invariance is immensely useful in test development work, but it is not something that can be assumed with IRT models. The property must be demonstrated across subpopulations of the population for whom the test is intended. This might be male and females; Blacks, Whites, and Hispanics; well-educated and less well-educated examinees; older, middle age, and younger respondents; etc. (see **Differential Item Functioning**) Basically, IRT DIF analyses involve comparing the item characteristic functions obtained in these subpopulations. Much of the research has investigated different ways to summarize the differences between these ICFs (see, [4, 6]). DIF via IRT modeling is not especially easy to implement (because of the number of steps involved in getting the ICFs from two or more groups on a common scale), but the easy graphing capabilities of ICFs makes DIF interpretation more understandable to many practitioners than the use of reporting DIF statistics only.

Test Score Linking or Equating

In many practical testing situations, such as achievement testing, it is desirable to have multiple versions

or forms of a test. For example, a test such as the Scholastic Assessment Test would quickly become of limited value if the same test items were used over and over again. Every high school senior would be going to a coaching school or scanning the Internet to get an advanced look at the test items. Items would become known to test takers and passed on to others about to take the test. Test validity would drop to zero very quickly. So, while multiple forms of a test may be a necessity, it is also important that these tests be statistically equivalent so that respondents do not benefit or be placed at a disadvantage because of the form of the test they were administered. Proper test development is invaluable in producing near equivalent tests, but it is no guarantee, and so ‘statistical equating’ is carried out to link comparable scores on pairs of tests. Statistical equating can be carried out with classical or IRT modeling, but it tends to be easier to do with IRT models and with a bit more flexibility. There is some evidence too that IRT equating may produce a better matching of scores at the low and high end of the ability scale (see, for example, [4, 7]).

Computer-Adaptive Testing (CAT)

‘Adapting a test’ to the performance of the examinee as he/she is working through the test has always been viewed as a very good idea because of the potential for shortening the length of tests (see **Computer-based Testing; Computer-Adaptive Testing**). The basic idea is to focus on administering items where the particular answer of the examinee is the most uncertain – that is, items of ‘medium difficulty’ for the examinee. When testing is done at a computer, ability can be estimated after the administration of each item, and then the ability estimate provides a basis for the selection of the next test items. Testing can be discontinued when the level of precision that is desired for ability estimates is achieved. It is not uncommon for the length of testing to be cut in half with CAT. The computer provides the mechanism for ability estimation and item selection. Item response theory provides a measurement framework for estimating abilities and choosing items. The property of ‘ability parameter invariance’ makes it possible to compare examinees to each other or standards that may have been set despite the fact that they almost certainly took collections of items that differed substantially in their ‘difficulty’. Without IRT, CAT

would lose many of its advantages. CAT is one of the important applications of IRT (see [1, 13]).

Future Directions and Challenges

The IRT field is advancing quickly. Nonparametric models have been introduced and open up a new direction for IRT model building. New models that build in a hierarchical structure are being advanced too (see **Hierarchical Item Response Theory Modeling**). Also, model parameter estimation methods are being developed based on Bayesian, marginal maximum likelihood, and Monte-Carlo Markov-Chain methods (see **Markov Chain Monte Carlo Item Response Theory Estimation**). Automated test assembly based on IRT models and optimization procedures will eventually be used in item selection (see [10]). Initiatives on just about every aspect of IRT are underway.

Despite the advances, challenges remain. One challenge arises because of the potential for multidimensional data. All of the popular IRT models are based on the assumption of a single dominant factor underlying performance on the test. It remains to be seen to what extent new tests are multidimensional, how that multidimensionality can be detected, and how it might be handled or modeled when it is present. A second challenge is associated with model fit. IRT models are based on strong assumptions about the data, and when they are not met, advantages of IRT modeling are diminished or lost. At the same time, approaches to addressing model fit, remain to be worked out, especially the extent to which model misfit can be present without destroying the validity of the IRT model application.

Finally, IRT is not a magic wand that can be used to fix all of the mistakes in test development such as (a) the failure to properly define the construct of interest, (b) ambiguous items, and (c) flawed test administrations. At the same time, it has been demonstrated many times over that IRT models, when they fit the data, and when other important features of sound measurement are present, IRT models provide

an excellent basis for developing tests, and providing valid scores for making decisions about individuals and groups.

References

- [1] Bartram, D. & Hambleton, R.K. (2005). *Computer-based Testing and the Internet: Issues and Advances*, Wiley, New York.
- [2] Fischer, G.H. & Molenaar, I. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York.
- [3] Gulliksen, H. (1950). *Theory of Mental Tests*, Wiley, New York.
- [4] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.
- [5] Hattie, J.A. (1985). Methodology review: assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9**, 139–164.
- [6] Holland, P.W. & Wainer, H. (1993). *Differential Item Functioning*, Lawrence Erlbaum, Mahwah.
- [7] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Mahwah.
- [8] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [9] Mills, C.N., Potenza, M.T., Fremer, J.J. & Ward, W.C., eds (2002). *Computer Based Testing: Building the Foundation for Future Assessments*, Lawrence Erlbaum, Mahwah.
- [10] Van der Linden, W. (2005). *Linear Models for Optimal Test Design*, Springer, New York.
- [11] Van der Linden, W. & Glas, C. eds (2000). *Computer Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers, Boston.
- [12] Van der Linden, W. & Hambleton, R.K. eds (1997). *Handbook of Modern Item Response Theory*, Springer-Verlag, New York.
- [13] Wainer, H., ed. (2000). *Computerized Adaptive Testing: A Primer*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [14] Zenisky, A.L. & Sireci, S.G. (2002). Technological innovations in large-scale assessment, *Applied Measurement in Education* **15**(4), 337–362.

RONALD K. HAMBLETON AND YUE ZHAO

Item Response Theory (IRT) Models for Polytomous Response Data

LISA A. KELLER

Volume 2, pp. 990–995

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Response Theory (IRT) Models for Polytomous Response Data

In many applications, such as educational testing, attitude surveys, and psychological scales, items are used that can take scores in multiple categories. For example, on a writing test, there could be an essay that is scored on a scale of 0–6. Alternately, on an opinion scale, there could be Likert-type items for which the response is a rating on a scale of strongly agree to strongly disagree. To model these types of item responses, item response theory (IRT) models that take into account these types of responses can be used to analyze the data and provide information regarding the level of latent trait (see **Latent Variable**) of interest that the respondent possesses. There are many IRT models that can be used to model this type of data, and several of the popular models are presented in this section. The models considered are the Graded Response Model, the Rating Scale Model, the Nominal Response Model, the Partial Credit Model, and the Generalized Partial Credit Model (GPCM) (see **Item Response Theory (IRT) Models for Rating Scale Data**).

Notation

To aid in the reading of the different models, some common notation will be used for all models. In all cases, θ will denote the latent variable of interest. The latent variable could be mathematical achievement, or some personality trait like anxiety, achievement motivation, or quality of life. In any case, the same notation will be used. Additionally, the scores of all items considered will be assumed to belong to one of m_i categories, where i indicates the particular item. Each item can have a different number of categories, but in all cases, the number of categories will be denoted m .

Graded Response Model

Samejima introduced the graded response model (GRM) in 1969 [8]. The GRM represents a family of

models that is appropriate to use when dealing with *ordered* categories. Such ordered categories could relate to constructed response items where examinees can receive various levels of scores, such as 0–4 points. The categories in that instance are 0, 1, 2, 3, and 4, and are clearly ordered. Alternately, the categories could be responses to Likert-type items, such as those in attitude surveys, where the response are strongly agree, agree, disagree, and strongly disagree. Again, there is an order inherent in the categories.

The GRM is a generalization of the two-parameter logistic (2PL) model. It uses the 2PL to provide the probability of receiving a certain score *or higher*, given the level of the underlying latent trait. As there are multiple categories to consider for each item, for any given level of θ , there is a probability of obtaining a score of x or higher. That is, the model provides

$$P_{ix}^*(\theta) = P[U_i \geq x|\theta], \quad (1)$$

where U_i is the random variable used to denote the response to an item with m score categories, and x is the specific score category.

Clearly, in the case where the particular score category is 0, the probability of getting a score of zero or higher must be one. That is,

$$P_{i0}^*(\theta) = 1. \quad (2)$$

Similarly, the probability of getting a score of $m_i + 1$ or higher is zero since the largest score is m_i . That is,

$$P_{i(m_i+1)}^*(\theta) = 0. \quad (3)$$

By defining the model in this way, it is possible to obtain the probability of getting a *specific* score, rather than the probability of getting a specific score *or higher*. This is achieved by

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta). \quad (4)$$

As the model is based on 2PL, there is a discrimination parameter involved for each item.

The model is expressed as:

$$P_{ix}^*(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}} \quad (5)$$

and a graphical representation of the same four-category item with equal a -parameters in Figure 1.

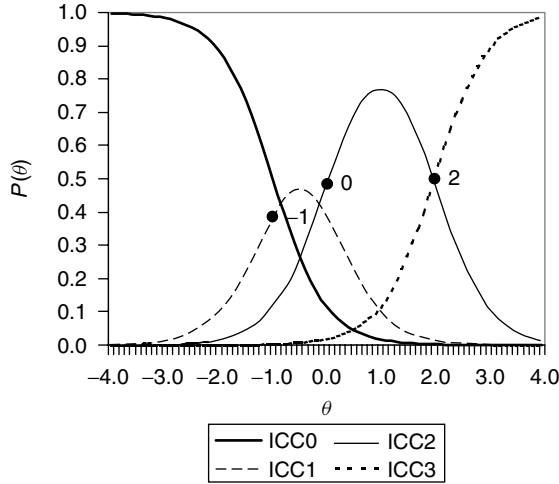


Figure 1 Score category response functions for the GRM ($a_i = 1.2$, $b_{i1} = -1.0$, $b_{i2} = 0.0$, $b_{i3} = 2.0$)

Considering Figure 1, the discrimination parameter for the item is 1.0. The category parameters are given by -1.0 , 0.0 , and 2.0 . These category parameters represent the point on the ability scale where an examinee has a 50% chance of getting a score of 1 or higher, 2 or higher, 3 or higher, etc. For example, at $\theta = 0.0$, the examinee has a 50% probability of obtaining a score of 2 or higher (that is, a score of 2, 3, 4, or 5). For values of $\theta > 0.0$, the probability of scoring a 2 or higher is greater than 50%.

Rating Scale Model

The Rating Scale model is an extension of the GRM, where, instead of having one location parameter for each category, b_{ix} , there is one location parameter *per item*, b_i , and a separate category parameter for each score category, c_{ix} . As a result, the model is expressed using

$$P_{ix}^*(\theta) = \frac{e^{Da_i(\theta - b_i + c_{ix})}}{1 + e^{Da_i(\theta - b_i + c_{ix})}}, \quad (6)$$

where this expression provides the probability of a person with latent trait level θ , receiving a score of x or higher, as in the case of the GRM. This model is often used for scoring Likert items; however, if the items have different numbers of score categories, the rating scale model is not appropriate and the GRM should be used (Muraki & Bock, 1997).

Partial Credit Model

As the GRM was an extension of the two-parameter logistic model, the Partial Credit Model (PCM), introduced by Masters in 1982 [6], is an extension of the one-parameter logistic model, or **Rasch model**, described elsewhere in this volume. As such, there is no discrimination parameter included in the model. The Rasch model provides the probability of getting a score of 1 *rather than* a score of 0, given some level of θ . In the case of the PCM, the Rasch model is used for each pair of adjacent categories. Thus, given an item with m_i response options, the Rasch model is used for each pair. In a four-category item, therefore, there are three pairs of adjacent score categories: 0 and 1, 1 and 2, and 2 and 3. Therefore, given a pair of adjacent response categories, $x - 1$ and x , they can be considered as '0' and '1', respectively, for the Rasch model. As such, the PCM provides the probability of obtaining a score of x rather than $x - 1$. This is in contrast to Samejima's GRM, which models the probability of getting a score of ' x or higher' for each score category x .

The model contains a person parameter, θ , as well as $m_i - 1$ parameters for each item. The $m_i - 1$ parameters correspond to the decision between adjacent categories. Therefore, there will always be one fewer parameters than categories. This is seen in the dichotomous case as well: there is one parameter for each item. This parameter results from the dichotomous decision between score categories 0 and 1.

The model can be easily seen as an extension of the Rasch model. For a particular person with trait level θ , and for a particular item, the Rasch model provides the probability of a correct response (i.e., a score of 1) as:

$$\begin{aligned} P(u = 1|\theta) &= \frac{P(u = 1|\theta)}{P(u = 1|\theta) + P(u = 0|\theta)} \\ &= \frac{\exp(\theta - b)}{1 + \exp(\theta - b)}. \end{aligned} \quad (7)$$

As noted previously, this model is monotonically increasing; as the value of θ increases, so does the probability of obtaining a score of 1.

The PCM extends this to the case of multiple score categories, and as noted above, considers only two categories at a time. Therefore, the probability of an examinee with a given level of θ getting a score of

x rather than a score of $(x - 1)$ on a particular item i , which has m categories is given by

$$P(u = x|\theta) = \frac{P[u = x|\theta]}{P[u = x|\theta] + P[u = (x - 1)|\theta]} = \frac{\exp(\theta - b_x)}{1 + \exp(\theta - b_x)}. \quad (8)$$

In this case, it is necessary to index the parameter b , as there will be many such parameters; there is one b for each comparison made, that is, $m - 1$. Again, as θ increases, the probability of getting score x rather than score $x - 1$ increases; however, the probability of getting score x does not necessarily continue to increase, as, at some point, a score of $x + 1$ becomes more likely than a score of x . Therefore, the model expressed in this way does not consider any other response options other than x and $x - 1$. It is often more desirable, and common, to consider the probability of getting a score of x , not just getting a score of x rather than a score of $x - 1$. As such, the more common way to write the model is

$$P(u = x|\theta) = \frac{\exp \sum_{k=0}^x (\theta - b_k)}{\sum_{h=0}^{m-1} \exp \sum_{k=0}^h (\theta - b_k)}. \quad (9)$$

In the case of the Rasch model, the b parameter provides the difficulty of the item. It is the location on the theta scale at which a person has a .50 probability of answering the item correctly. In the PCM, it is a little more complex, as there are multiple score categories, and only two are taken into consideration when estimating the b_x parameter. The value of b_x is the point where the probability of receiving score x is equal to the probability of receiving score $x + 1$.

A graphical representation of the score categories is provided in Figure 2 for a four-category item with $b_0 = -1.0$, $b_1 = 0.0$, $b_2 = 2.0$:

The b_x values are labeled on the graph. One interpretation of these 'step' parameters is to consider that to have a greater probability of getting a score of 1 over that of getting a score of 0, a θ -level greater than -1.0 is necessary. Similarly, to have a greater probability of getting a score of 1 over a score of 2, a θ -level of greater than

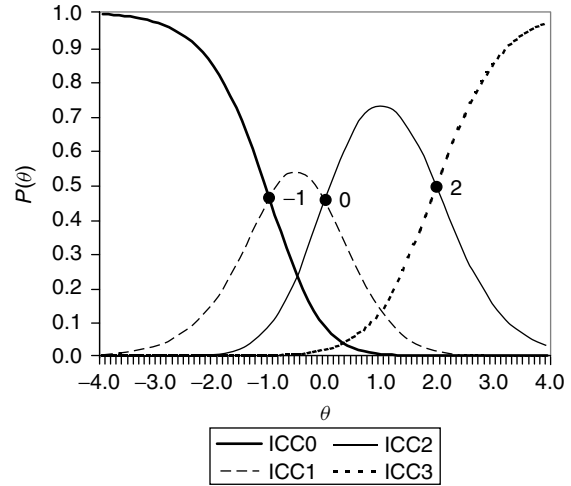


Figure 2 Score category response functions for the PCM ($b_{i1} = -1.0$, $b_{i2} = 0.0$, $b_{i3} = 2.0$)

0.0 is required, and similarly, for having a greater probability of getting a score of 3 over a score of 2.

It is interesting to note that the b_x values need not be ordered such that $b_1 < b_2 < \dots < b_{m-1}$ because the parameter represents the relative magnitude of only *two* adjacent probabilities. When the ordering changes, the interpretation becomes more complicated and has been the source of considerable discussion among researchers.

The Generalized Partial Credit Model

In many cases, different items on the same test/scale do have different discriminations. In 1992, Muraki generalized the Partial Credit Model by allowing items to have different discrimination parameters [7]. As such, this model can be seen as an extension of the two-parameter logistic model in the same way that the PCM can be viewed as an extension of the one-parameter logistic model, or the Rasch model. In this instance, adjacent score categories are still considered, and the model provides the probability of obtaining a score of x rather than a score of $x - 1$, just as in the case of the PCM. However, this probability is given by a different function, which is the two-parameter logistic function that has been presented earlier:

$$P(u = x|\theta) = \frac{P[u = x|\theta]}{P[u = x|\theta] + P[u = (x-1)|\theta]} = \frac{\exp Da(\theta - b_x)}{1 + \exp Da(\theta - b_x)}, \quad (10)$$

where D is a scaling constant equal to 1.7, to put θ on the same metric as the normal ogive model, and a is the discrimination parameter for the particular item. Clearly, the PCM can be seen as a special case of this model where $D = 1$ and $a = 1$ for all items.

Again, it is more desirable to have an expression for the probability of obtaining a score of x , and not the probability of obtaining a score of x rather than a score of $x - 1$. Thus, the model can be expressed as before with

$$P(u = x|\theta) = \frac{\exp \sum_{k=0}^x Da(\theta - b_k)}{\sum_{h=0}^{m-1} \exp \sum_{k=0}^h Da(\theta - b_k)}. \quad (11)$$

The b_x parameters can be interpreted the same way as in the PCM. The only difference is the effect of the a -parameter on the resulting curves. Figure 3 provides a graphical representation of the same four-category item with $b_0 = -1.0$, $b_1 = 0.0$, $b_2 = 2.0$ and $a = 1.20$.

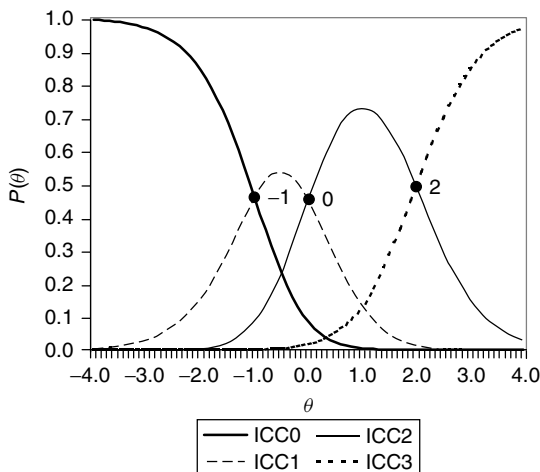


Figure 3 Score category response functions for the GPCM ($a_i = 1.2$, $b_{i1} = -1.0$, $b_{i2} = 0.0$, $b_{i3} = 2.0$)

Although the GPCM and the GRM are conceptualized differently, the results obtained from using the two models are very similar.

The Nominal Response Model

Bock introduced the nominal response model (NMR) in 1972, which is the most general of the unidimensional polytomous IRT models [1]. In this model, there is no assumption that the categories are ordered in any particular way. That is, a score of '2' is not necessarily smaller than a score of '3'. An example of this type of scoring can be seen in scoring multiple-choice items, where each response is given a particular value. Consider a multiple-choice item with four options: A, B, C, and D. This item can be viewed as a polytomously scored item where each option receives a score. There is then a probability associated with each response option. For any given level of θ , there is a probability that a person chooses option A, B, C, and D. And these probabilities change as the level of θ changes.

The nominal response model provides the probability that a person with latent trait level, θ , obtains a score of x . The model is expressed as:

$$P_{ix}(\theta) = \frac{e^{a_{ix}\theta + c_{ix}}}{\sum_{k=1}^{m_i} e^{a_{ix}\theta + c_{ix}}}. \quad (12)$$

In this case, each score category has its own discrimination parameter, a_{ix} . It is for this reason that the response categories do not retain the strict ordering. The other parameter, the c_{ix} , is an intercept parameter for each score category. An alternate parameterization of the model exists as:

$$P_{ix}(\theta) = \frac{e^{a_{ix}(\theta - b_{ix})}}{\sum_{k=1}^{m_i} e^{a_{ix}(\theta - b_{ix})}}. \quad (13)$$

which is similar in form to the other models presented in this section. The b_{ix} parameter is a location parameter and $b_{ix} = -c_{ix}/a_{ix}$. A graphical representation of a four-category item is presented in Figure 4.

The most popular IRT models used for scoring polytomous models were presented in this entry. These models have been applied in many contexts in the social sciences, especially the GRM and PCM.

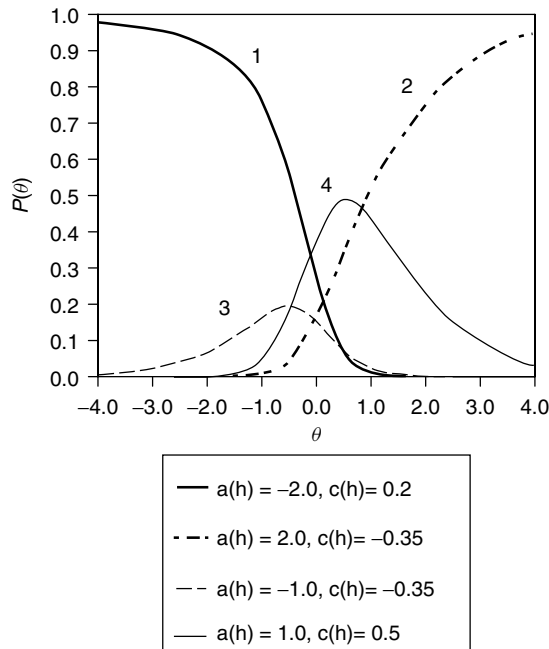


Figure 4 Score category response functions for the NRM

They offer flexibility in the assumptions of the scoring: ordered versus nonordered categories and equal versus nonequal discrimination among items, and as such are capable of modeling a variety of data. The reader is directed to studies such as [2–5, 9, 11, 12] for specific examples. For more detail about the models presented here, as well as other IRT models, the reader is referred to [10].

References

- [1] Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories, *Psychometrika* **37**, 29–51.
- [2] Fraley, R.C., Waller, N.G. & Brennan, K.A. (2000). An item response theory analysis of self-report measures of adult attachment, *Journal of Personality and Social Psychology* **78**(2), 350–365.
- [3] Kalinowski, A.G. (1985). Measuring clinical pain, *Journal of Psychopathology & Behavioral Assessment* **7**(4), 329–349.
- [4] Keller, L.A., Swaminathan, H. & Sireci, S.G. (2003). Evaluating scoring procedures for context-dependent item sets, *Applied Measurement in Education* **16**(3), 207–222.
- [5] Levent, K., Howard, B.M. & Tarter, R.E. (1996). Psychometric evaluation of the situational confidence questionnaire in adolescents: fitting a graded item response model, *Addictive Behaviors* **21**(3), 303–318.
- [6] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.
- [7] Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm, *Applied Psychological Measurement* **16**, 159–176.
- [8] Samejima, F. (1969). Estimation of ability using a response pattern of graded scores, *Psychometrika Monograph* No. 18.
- [9] Singh, J., Howell, R.D. & Rhoads, G.K. (1990). Adaptive designs for Likert-type data: an approach for implementing marketing surveys, *Journal of Marketing Research* **27**(3), 304–322.
- [10] van der Linden, W.J. & Hambleton, R.K., eds 1997). *Handbook of Modern Item Response Theory*, Springer, New York.
- [11] Zhu, W. & Kurz, K.A. (1985). Rasch partial credit analysis of gross motor competence, *Perceptual Motor Skills* **79**(2), 947–962.
- [12] Zickar, M.J. & Robie, C. (1999). Modeling faking good on personality items: an item-level analysis, *Journal of Applied Psychology* **84**(4), 551–563.

LISA A. KELLER

Item Response Theory (IRT) Models for Rating Scale Data

GEORGE ENGELHARD

Volume 2, pp. 995–1003

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Item Response Theory (IRT) Models for Rating Scale Data

IRT Models for Rating Scale Data

Rating scales are the most common formats for collecting observations in the behavioral and social sciences. Rating scale item i provides the opportunity for a person to select a score X in $m_i + 1$ ordered categories ($x = 0, 1, \dots, m_i$). For example, the person might be an examinee indicating an attitude using a Likert Scale (0 = Strongly Disagree, 1 = Disagree, 2 = Uncertain, 3 = Strongly Agree, 4 = Agree) or a rater judging the quality of student writing (0 = novice, 1 = proficient, 2 = accomplished). If there are two response categories (0 = incorrect, 1 = correct), then the ratings yield dichotomous data (see **Item Response Theory (IRT) Models for Dichotomous Data**). If there are three or more categories, then the ratings yield polytomous data. Higher scores generally indicate a higher location on the construct. In IRT, the construct is defined as an unobservable construct or trait. The goal of measurement is locate both persons and items on this underlying construct in order to define measuring instrument.

One of the central problems in psychometrics is the development of models that connect person measures and item calibrations in a meaningful way to represent a construct. The basic idea that motivates the use of IRT models for rating scale data is that the scoring of $m + 1$ ordered categories with ordered integers (0, 1, $\dots$, m) using the assumption that there are equal intervals between the categories may not be justified. IRT models provide a framework to explicitly examine this assumption, and parameterize the categories without this assumption. Specifically, IRT models for rating scale data are used to model category response functions (CRFs) that link the probability of a specific rating with person measures and a set of characteristics reflecting item and category calibrations. The category response functions (CRFs) represent the probability of obtaining a score of x on item i as a function of a person's location on the construct θ . The CRFs can be written as follows:

$$P_{xi}(\theta) = \text{pr}(X_i = x|\theta), \quad (1)$$

for $x = 0, 1, \dots, m$. The IRT models for rating scale data vary in terms of how they define the operating characteristic functions (OCFs). CRFs and OCFs will be defined for each model in the following sections.

There are several different ways to categorize IRT models for analyzing rating scale data. In this entry, the specific models described are unidimensional models for ordered categories. Following Embretson and Reise [6], these models can be viewed as either direct or indirect models. Direct models focus directly on estimating CRFs, while indirect models require two steps that involve first estimating the OCFs and then the CRFs. Thissen and Steinberg [18] referred to these models as divide-by-total and difference models respectively. They also point out that it is possible to go back and forth between the divide-by-total and difference forms, although the derivational form of the model is usually less complex algebraically.

Direct Models

Partial Credit Model

The Partial Credit Model [PCM; 9, 10] is a unidimensional IRT model for ratings in two or more ordered categories. The PCM is a **Rasch model**, and therefore provides the opportunity to realize a variety of desirable measurement characteristics, such as separability of person and item parameters, sufficient statistics for parameters in the model, and specific objectivity [14]. When good model-data fit is obtained, then the PCM, as well as other Rasch models, yields invariant measurement [7]. The PCM is a straightforward generalization of the Rasch model for dichotomous data [15] applied to pairs of increasing adjacent categories.

The Rasch model for dichotomous data can be written as:

$$\frac{P_{i1}(\theta)}{P_{i0}(\theta) + P_{i1}(\theta)} = \frac{\exp(\theta - \delta_{i1})}{1 + \exp(\theta - \delta_{i1})}, \quad (2)$$

where $P(\theta)_{i1}$ is the probability of scoring 1 on item i , $P(\theta)_{i0}$ is the probability of scoring 0 on item i , θ is the location of the person on the construct, and δ_{i1} is the location on the construct where the probability of responding in adjacent categories, 0 and 1, is equal. For the dichotomous case δ_{i1} is defined as the difficulty of item i . Equation (2) represents the operating characteristics function for the PCM. When

2 IRT Models for Rating Scale Data

the data are collected with more than two response categories, then the OCF can be generalized as

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta) + P_{ix}(\theta)} = \frac{\exp(\theta - \delta_{ix})}{1 + \exp(\theta - \delta_{ix})}, \quad (3)$$

where $P(\theta)_{ix}$ is the probability of scoring x on item i , $P(\theta)_{ix-1}$ is the probability of scoring $x - 1$ on item i , θ is the location of the person on the construct, and δ_{ix} is the location on the construct where the probability of responding in adjacent categories, $x - 1$ and 1 , is equal.

The category response function (CRF) for the PCM is:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^x (\theta - \delta_{ij})\right]}{\sum_{r=0}^{mi} \exp\left[\sum_{j=0}^r (\theta - \delta_{ij})\right]}, \quad (4)$$

$x_i = 0, 1, \dots, m_i,$

where $\sum_{j=0}^0 (\theta - \delta_{ij}) \equiv 0$. The δ_{ij} parameter is still interpreted as the intersection between the two consecutive categories where the probabilities of responding in the adjacent categories is equal. The δ_{ij} term is described as a step difficulty by Masters and Wright [10]. Embretson and Reise [6] have suggested calling the δ_{ij} term a category intersection parameter. It is important to recognize that the item parameter δ_{ij} represents the location on the construct where a person has the same probability of responding in categories x and $x - 1$. These conditional probabilities for adjacent categories are expected to increase, but they are not necessarily ordered from low to high on the construct θ . By defining the item parameters locally, it is possible to verify that persons are using the categories as expected. See Andrich [3] for a detailed description of the substantive interpretation of disordered item category parameters.

Generalized Partial Credit Model

Muraki [12, 13] formulated the Generalized Partial Credit Model (GPCM) based on Master's PCM [9]. The GPCM is also a unidimensional IRT model for ratings in two or more ordered categories. One distinguishing feature of the GPCM is the use of the two-parameter IRT model for dichotomous data [4, 8]

as the OCF. As pointed out in the previous section, the OCF for the PCM is the dichotomous Rasch model. Two-parameter IRT model allows for the estimation of two item parameters: item difficulty and item discrimination. These item parameters define location and scale. The OCF for the GPCM is

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta) + P_{ix}(\theta)} = \frac{\exp[\alpha_i(\theta - \delta_{ix})]}{1 + \exp[\alpha_i(\theta - \delta_{ix})]}, \quad (5)$$

where $P(\theta)_{ix}$ is the probability of scoring x on item i , $P(\theta)_{ix-1}$ is the probability of scoring $x - 1$ on item i , θ is the location of the person on the construct, δ_{ix} is the location on the construct where the probability of responding in adjacent categories, $x - 1$ and 1 , is equal, and α_i is the item discrimination or slope parameter. The CRF for the GPCM with two or more ordered response categories is

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^x \alpha_i(\theta - \delta_{ij})\right]}{\sum_{r=0}^{mi} \exp\left[\sum_{j=0}^r \alpha_i(\theta - \delta_{ij})\right]}, \quad (6)$$

$x_i = 0, 1, \dots, m_i,$

where $\delta_{i0} \equiv 0$. Sometimes a scaling constant of 1.7 is added to the model in order to place the θ scale in the same metric as the normal ogive model. The addition of an item discrimination parameter into the OCFs allows for a scale parameter in the CRFs that reflects category spread. The GPCM provides additional information about the characteristics of rating scale data as compared to the PCM. The addition of the scale parameters means that the GPCM is likely to provide better model-data fit than the PCM. The tradeoff is that it is no longer possible to separate person and item parameters; for example, the raw score is no longer a sufficient statistic for locating persons on the construct.

Rating Scale Model

The Rating Scale Model [RSM; 1, 2] is another unidimensional IRT model that can be used analyze ratings in two or more ordered categories. The RSM, as was the PCM, is a member of the Rasch family of IRT models, and therefore shares the desirable measurement characteristics, such as invariant measurement [7] when there is good model-data fit. The

RSM is similar to the PCM, but the RSM was developed to analyze rating scale data with a common or fixed number of response categories across a set of items designed to measure a unidimensional construct or construct. The PCM does not require the same number of categories for each rating scale item. Likert scales are a prime example of this type of format for rating scales. For items with the same number of response categories, the RSM decomposes the category parameter, δ_{ij} , from the PCM into two parameters: a location parameter λ_i that reflects item difficulty and a category parameter δ_j . In other words, the δ_{ij} parameter in the PCM is decomposed into two components, $\delta_{ij} = (\lambda_i + \delta_j)$ where λ_i are the location of the items on the construct and the δ_j are the category parameters across items. The category parameters are considered equivalent across items for the RSM. The OCFs for the RSM is

$$\frac{P_{ix}(\theta)}{P_{ix-1}(\theta) + P_{ix}(\theta)} = \frac{\exp[\theta - (\lambda_i + \delta_x)]}{1 + \exp[\theta - (\lambda_i + \delta_x)]}, \quad (7)$$

where $P(\theta)_{ix}$ is the probability of scoring x on item i , $P(\theta)_{ix-1}$ is the probability of scoring $x - 1$ on item i , θ is the location of the person on the construct, λ_i are the location of the items on the construct, and δ_x is the location on the construct where the probability of responding in adjacent categories, $x - 1$ and 1 , is equal across items. The δ_x parameter is also called the *centralized threshold*. The CRF for the RSM is

$$P_{ix}(\theta) = \frac{\exp \left[\sum_{j=0}^x (\theta - (\lambda_i + \delta_j)) \right]}{\sum_{r=0}^m \exp \left[\sum_{j=0}^r (\theta - (\lambda_i + \delta_j)) \right]}, \quad (8)$$

$x = 0, 1, \dots, m,$

where $\sum_{j=0}^0 (\theta - (\lambda_i + \delta_j)) \equiv 0$. For the RSM, the shape of the OCFs and CRFs across items are the same, while the location varies. Both the RSM and PCM become the dichotomous Rasch model when the rating scale data are collected with two response categories.

Indirect Models

Graded Response Model

The graded response model [GRM; 16, 17] is another unidimensional IRT model for ordered responses. The GRM is an indirect model [6] that requires first the estimation of the OCFs, and then second the subtraction of the OCFs to obtain the CRFs. Thissen and Steinberg [18] call the GRM a difference model. The OCFs for the GRM are two-parameter IRT models for dichotomous data [4, 8]. Although both the GRM and the GPCM share this form for the OCFs, the GRM dichotomizes the categories within the rating scale in a different way. For example, the GRM treats four ordered response categories as a series of three dichotomies as follows: 0 vs. 1, 2, 3; 0, 1 vs. 2, 3; 0, 1, 2 vs. 3. While the PCM, GPCM, and RSM group adjacent categories to form the dichotomies: 0 vs. 1; 1 vs. 2; 2 vs. 3. This difference influences the substantive interpretations of the category parameters in the models. The OCFs for the GRM are modeling the probability of a person scoring x or greater on item i , while the direct models are modeling the probability of x or $x - 1$ for adjacent categories. The OCF for the GRM is

$$P_{ix}^*(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]}. \quad (9)$$

The β_{ij} parameter is interpreted as the location on the construct necessary to respond above the j threshold on item i with a probability of .50, θ is the location of a person on the construct, and α_i is the slope parameter common across OCFs within item i . The slope parameter is constrained to be equal within item i , and this is called the *homogeneous case of the GRM*. The CRF for the GRM is

$$P_{ix}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ix})]}{1 + \exp[\alpha_i(\theta - \beta_{ix})]} - \frac{\exp[\alpha_i(\theta - \beta_{i(x+1)})]}{1 + \exp[\alpha_i(\theta - \beta_{i(x+1)})]}, \quad (10)$$

$x_i = 1, \dots, m_i,$

or

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta), \quad (11)$$

where $P_{i(x=0)}^*(\theta) = 1.0$ and $P_{i(x=mi+1)}^*(\theta) = 0.0$. It is important to highlight the substantive interpretations of category parameters β_{ij} of the GRM. The

4 IRT Models for Rating Scale Data

β_{ij} is not the location on the construct where a person has the same probability of being in adjacent categories – this category parameter is not the point of intersection. The category parameters in the PCM, GPCM, and RSM do represent the intersection points. The category parameters β_{ij} are parameterized to be ordered within an item: $\beta_{i1} \leq \beta_{i2} \dots \leq \beta_{im}$. The GRM has a historical connection to the earlier work of **Thurstone** on the method of successive intervals [5] with the important distinction that a person parameter is added to the GRM model. The β_{ij} parameters are sometimes called *Thurstone thresholds*. These Thurstone thresholds represent the location on the construct where the probability of being rated j or above equals the

probability of being in the categories below j for item i .

Modified Graded Response Model

Muraki [11] proposed a Modified Graded Response Model (MGRM) as a restricted version of the GRM [16]. The MGRM is designed for rating scale data that has a fixed number of common response formats across items. The MGRM is unidimensional IRT model for ordered response data. As with the GRM, the MGRM is an indirect model that requires a two-step process with separate estimation of the OCFs, and then the creation of the CRFs through subtraction. The MGRM is analogous to Thurstone's

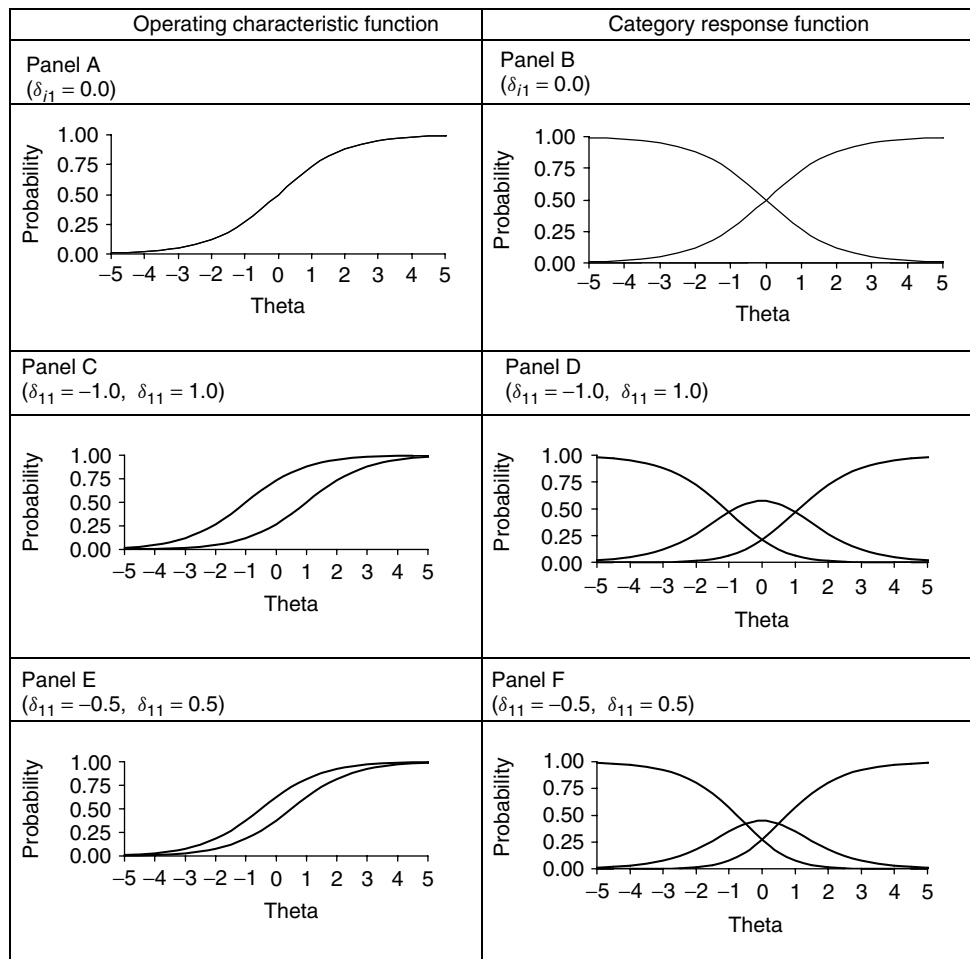


Figure 1 Partial credit model

method of successive intervals or categories [6]. The distinctive feature of the MGRM is the reparameterization of the GRM category parameter β_{ij} into two components: a location parameter β_i and a category threshold parameter τ_j that is common across items: $\beta_{ij} = (\beta_i + \tau_j)$. The OCF for the MGRM is

$$P_{ix}^*(\theta) = \frac{\exp[\alpha_i(\theta - \beta_i + \tau_j)]}{1 + \exp[\alpha_i(\theta - \beta_i + \tau_j)]} \quad (12)$$

and the CRF is

$$P_{ix}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_i + \tau_j)]}{1 + \exp[\alpha_i(\theta - \beta_i + \tau_j)]}$$

$$- \frac{\exp[\alpha_i(\theta - \beta_i + \tau_{(j-1)})]}{1 + \exp[\alpha_i(\theta - \beta_i + \tau_{(j-1)})]},$$

$$x = 1, 2, \dots, m, \quad (13)$$

or

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta), \quad (14)$$

where $P_{i(x=0)}^*(\theta) = 1.0$ and $P_{i(x=mi+1)}^*(\theta) = 0.0$.

Graphical Displays of the Five Models

Figures 1 to 5 provide a graphical framework for comparing some of the essential features of the

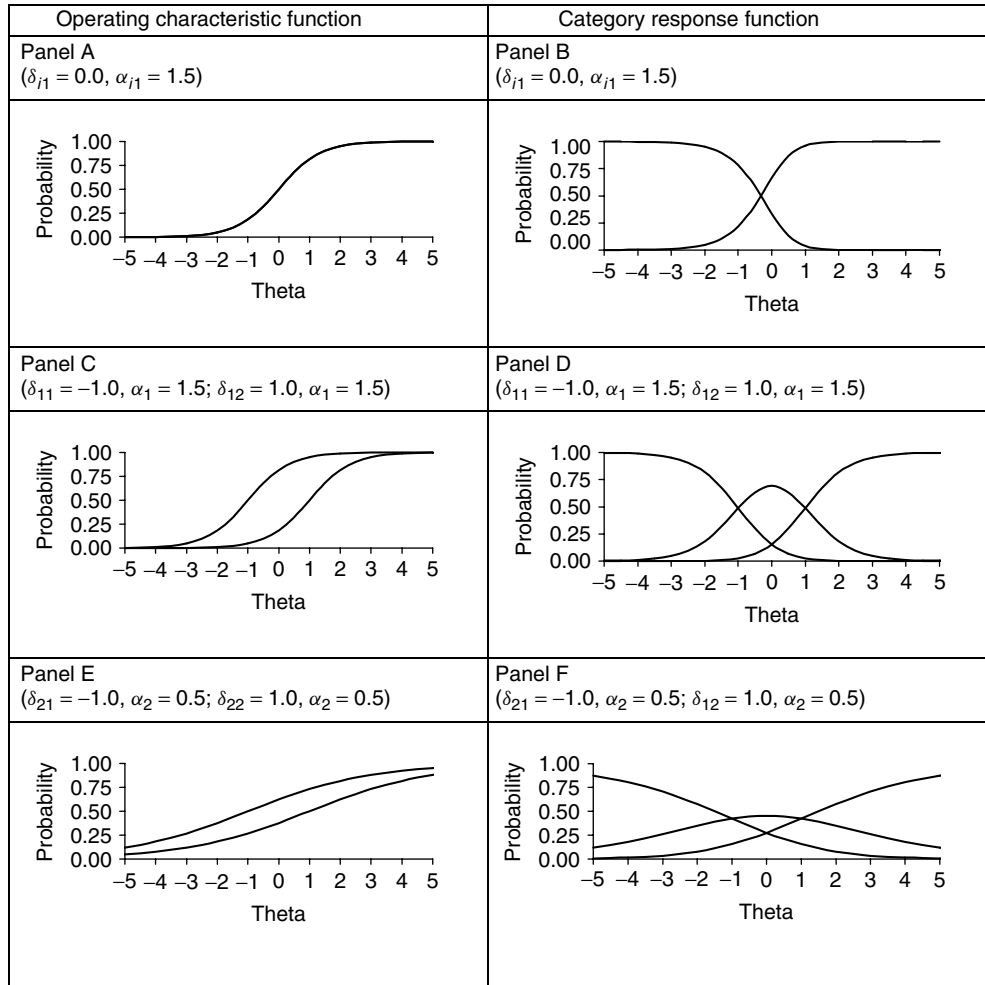


Figure 2 Generalized partial credit model

6 IRT Models for Rating Scale Data

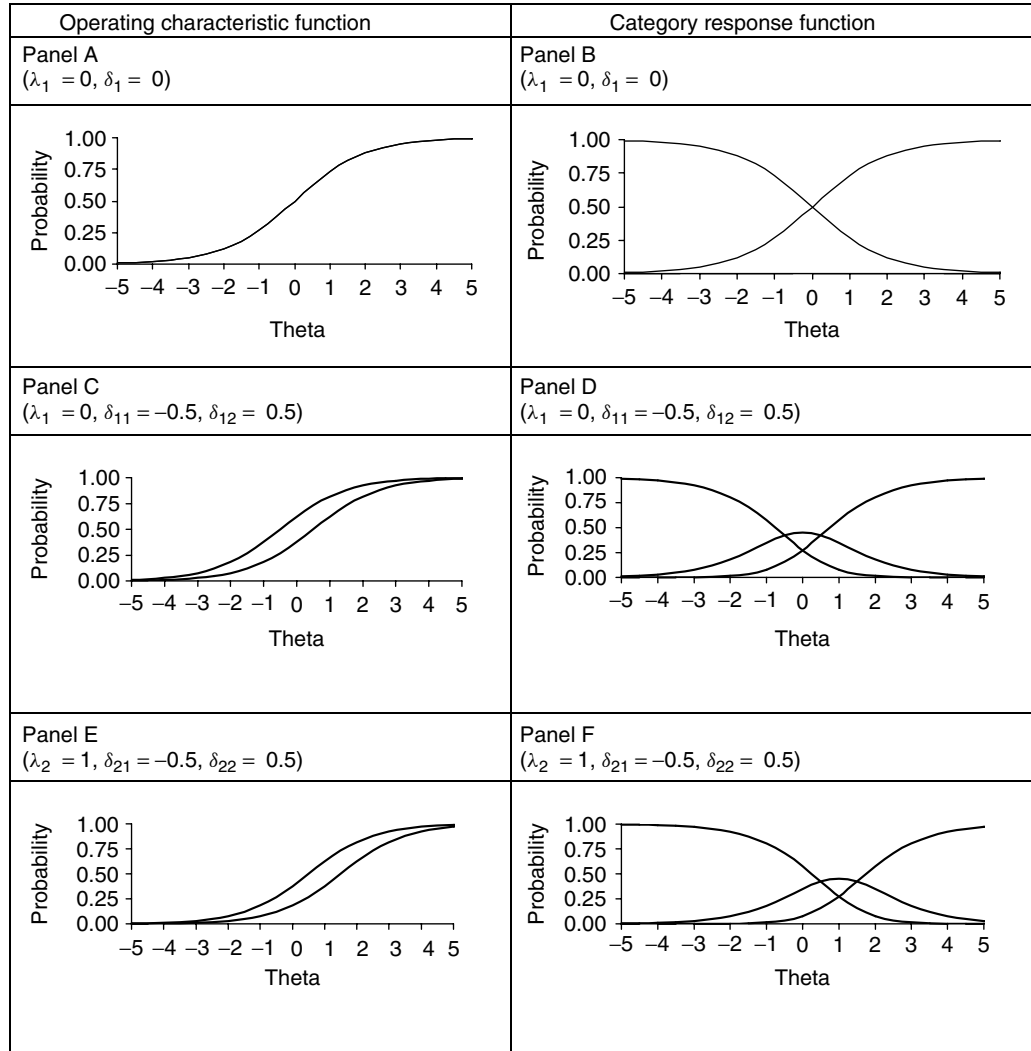


Figure 3 Rating scale model

five IRT models for rating scale data described in this entry. Panels A to F in each figure illustrate the OCFs and CRFs for the five models. The left columns present the OCFs and the right columns present the CRFs. Panels A and B show the OCF and CRF for the dichotomous case, while Panels C through F illustrate these functions for the polytomous case (three ratings) with various values of the parameters for each model. In each figure, the x -axis is the theta value (θ) that represents the location of persons on the latent variable or construct. The y -axis is the conditional probability of

responding in a particular category as a function of the category parameters and the θ location of each person.

Figures 1 to 3 present the direct models (PCM, GPCM, RSM), while Figures 4 and 5 illustrate the indirect models (GRM, MGRM). It is beyond the scope of this entry to provide an extensive comparative discussion of the models, but several points should be noted. First, starting with the PCM in Figure 1, it is easy to see that as the category parameters δ_{ix} become closer together (comparing Panels D and F) the probability of being in

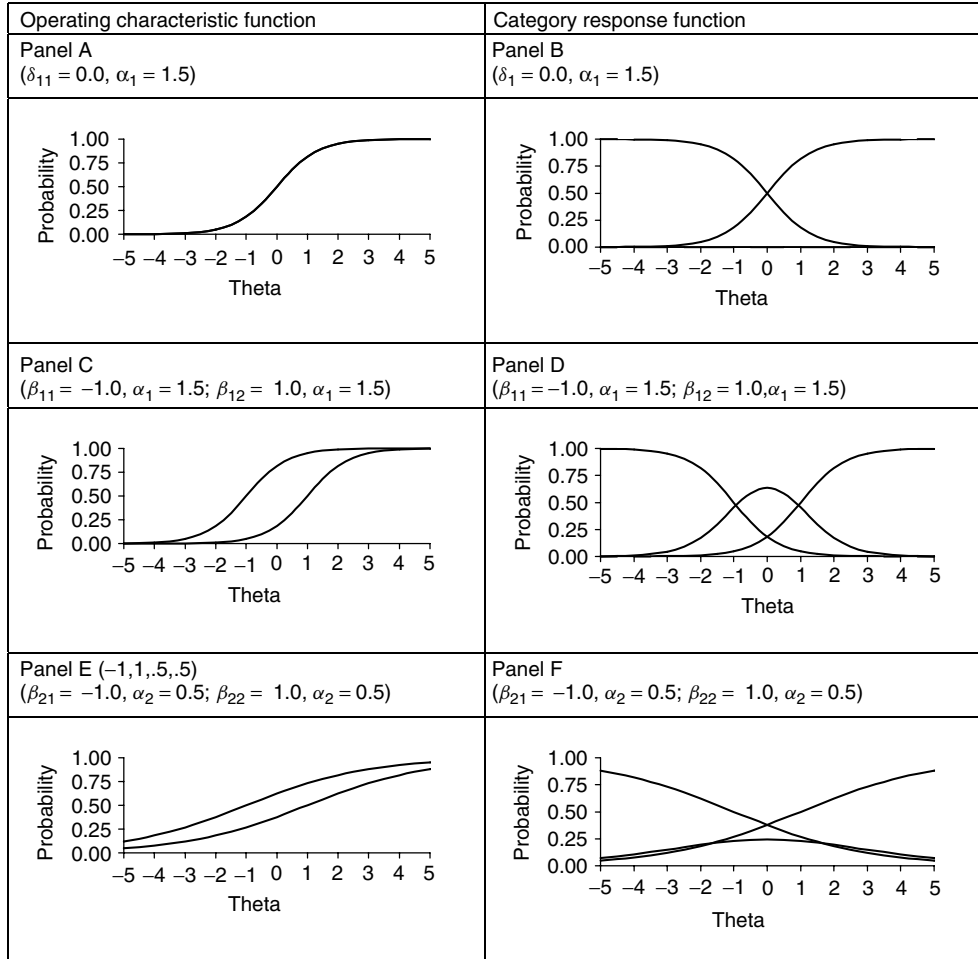


Figure 4 Graded response model

the middle category decreases and the CRFs suggest that less information is being conveyed in this case. Second, Figure 2 for the GPCM illustrates how the addition of a slope or scale parameter, α_i , influences the shape or spread of the OCFs and CRFs. Comparing Panel D in Figure 1 (PCM) with Panels D and F in Figure 2 (GPCM) clearly shows how values of α_i greater than 1 ($\alpha_1 = 1.5$) lead to more peaked CRFs (Figure 2, Panel D), while values of α_i less than 1 ($\alpha_1 = 0.5$) lead to flatter CRFs (Figure 2, Panel F). Third, the RSM shown in Figure 3 illustrates the assumption that even though item locations can change ($\lambda_1 = 0, \lambda_2 = 1$), the rating structure is modeled

to constant across items ($\delta_{11} = \delta_{12} = -0.5, \delta_{21} = \delta_{22} = 0.5$). This is shown in Panels C to F in Figure 3 (RSM).

Figures 4 and 5 present the indirect models (GRM and MGRM). It is important to note that even though the dichotomous cases (Panels A and B) appear graphically equivalent to the direct models in Figures 1 to 3, the substantive definitions and underlying models are distinct. The category parameters for item i , β_{ij} , for the GRM represents the location on the construct where being rated x or above equals the probability of being in the categories below x for item i , while the category parameters for the direct models (PCM, GPCM, and RSM) represent

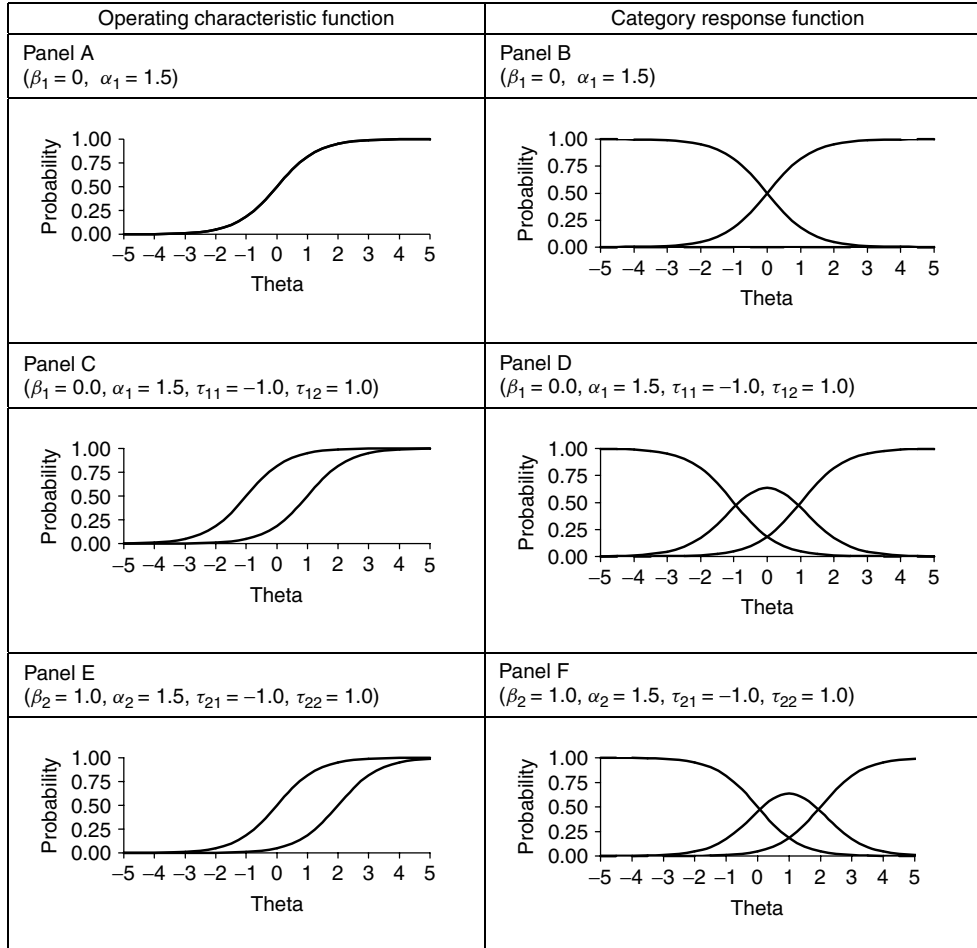


Figure 5 Modified graded response model

the intersections between CRFs. The GRM presented in Figure 4 illustrates how the scale parameter, α_i , influences the shape (flatness or peakedness) of the CRFs as it does for the direct models. Values of α_i greater than 1 ($\alpha_1 = 1.5$) lead to more peaked CRFs (Figure 4, Panel D), while values of α_i less than 1 ($\alpha_1 = 0.5$) lead to flatter CRFs (Figure 4, Panel F). The MGRM shown in Figure 5 models the rating structure ($\beta_{ij} = \beta_i + \tau_j$) as constant across items. The MGRM shown in Figure 5 illustrates the assumption that even though item location can change ($\beta_1 = 0.0, \beta_2 = 1.0$), the rating structure is modeled to constant across items ($\tau_{11} = \tau_{12} = -1.0, \tau_{21} = \tau_{22} = 1.0$). This is shown in Panels C to F in Figure 5 (MGRM).

Summary

This entry provides a very brief introduction to several IRT models for rating scale data. Embretson and Reise [6] and van der Linden and Hambleton [19] should be consulted for detailed descriptions of these models, as well as descriptions of additional models not included in this entry. The development, comparison, and refinement of IRT models for rating scale data is one of the most active areas in psychometrics.

References

- [1] Andrich, D.A. (1978a). Application of a psychometric model to ordered categories which scored with

- successive integers, *Applied Psychological Measurement* **2**, 581–594.
- [2] Andrich, D.A. (1978b). A rating formulation for ordered response categories, *Psychometrika* **43**, 561–573.
- [3] Andrich, D.A. (1996). Measurement criteria for choosing among models for graded responses, in *Analysis of Categorical Variables in Developmental Research*, A. von Eye & C.C. Clogg, eds, Academic Press, Orlando, pp. 3–35.
- [4] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 395–479.
- [5] Edwards, A.L. & Thurstone, L.L. (1952). An internal consistency check for scale values determined by the method of successive intervals, *Psychometrika* **17**, 169–180.
- [6] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Hillsdale.
- [7] Engelhard, G. (1994). Historical views of the concept of invariance in measurement theory, in *Objective Measurement: Theory into Practice*, Vol. 2, M. Wilson, ed., Ablex, Norwood, pp. 73–99.
- [8] Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale.
- [9] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.
- [10] Masters, G.N. & Wright, B.D. (1997). The partial credit model, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer, New York, pp. 101–121.
- [11] Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data, *Applied Psychological Measurement* **14**, 59–71.
- [12] Muraki, E. (1992). A generalized partial credit model. Application of an EM algorithm, *Applied Psychological Measurement* **16**, 159–176.
- [13] Muraki, E. (1997). A generalized partial credit model, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer, New York, pp. 153–164.
- [14] Rasch, G. (1977). On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements, *Danish Yearbook of Philosophy* **14**, 58–94.
- [15] Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press, Chicago.
- [16] Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- [17] Samejima, F. (1997). The graded response model, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer, New York, pp. 85–100.
- [18] Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models, *Psychometrika* **51**(4), 567–577.
- [19] van der Linden, W.J. & Hambleton, R.K., eds (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

GEORGE ENGELHARD

Jackknife

JOSEPH LEE RODGERS

Volume 2, pp. 1005–1007

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Jackknife

The jackknife is a resampling procedure that is parallel to several other resampling methods, including the **bootstrap**, the **randomization test**, and **cross-validation**. These methods are designed to generate an empirical **sampling distribution**, which can be used to do hypothesis testing, to estimate standard errors, and to define **confidence intervals**. They have particular value in statistical settings in which the assumptions of parametric statistical procedures are not appropriate. These methods are alternatives to the whole class of hundreds of parametric statistical procedures, including t Tests, the **analysis of variance (ANOVA)**, regression (see **Multiple Linear Regression**), and so on.

Parametric procedures such as the t Test and ANOVA (developed by **Fisher**, **Gosset**, and others in the early twentieth century), were based on theoretical sampling distributions. The mathematical development of these distributions was, at least in part, a necessary response to the computational limitations of the time, which precluded using computationally intensive methods involving resampling. In fact, Fisher invented the first resampling procedure, the randomization test, but could not use it extensively because of computational limitations [3]. He was convinced enough of the value of resampling methods to note that ‘the statistician does not carry out this very tedious process but his conclusions have no justification beyond the fact that they could have been arrived at by this very elementary method (quoted in Edgington [1]).’ Edgington characterized resampling methods as ‘the substitution of computational power for theoretical analysis’ (p. 3). Efron [2] noted that ‘From a traditional point of view, [resampling] methods . . . are prodigious computational spendthrifts. We blithely ask the reader to consider techniques which require the usual statistical calculations to be multiplied a thousand times over. None of this would have been feasible twenty-five years ago, before the era of cheap and fast computation (p. 2)’.

Even with the development of the computer, and the almost unlimited computational availability of the modern computing era, the use of resampling procedures in applied research was slow to develop (see

discussion in [1]). By the beginning of the twenty-first century, however, there are many software systems that support and implement resampling methods (see [7] for a discussion of these programs). Further, the use of resampling methods is becoming more integrated into applied statistics and research settings.

The jackknife was originally developed by Quenouille [6] and **Tukey** [8]. Tukey named it the ‘jackknife’ to ‘suggest the broad usefulness of a technique as a substitute for specialized tools that may not be available, just as the Boy Scout’s trusty tool serves so variedly [5].’ Mosteller and Tukey follow with a more specific description: ‘The jackknife offers ways to set sensible confidence limits in complex situations (p. 133)’. Efron [2], the developer of the bootstrap, named the jackknife method after its developers, the ‘Quenouille–Tukey jackknife’.

In fact, the jackknife can be considered a whole set of procedures, because the original development has been expanded, and new versions of the jackknife have been proposed. For example, Efron [2] discusses the ‘infinitesimal jackknife,’ which provides an alternative to the traditional jackknifed standard deviation. Efron refers the reader to several highly technical articles for sophisticated treatment of the jackknife, including its asymptotic theory (e.g., [4]).

How is the jackknife applied? According to Mosteller and Tukey [5, p. 133], ‘The basic idea is to assess the effect of each of the groups into which the data have been divided . . . through the effect upon the body of data that results from *omitting* that group.’ However, the use of the method has been reformulated since that description. In particular, instead of viewing the effect of ‘omitting that group,’ as suggested in Tukey’s work, a more modern view is to consider resampling without replacement to create a subsample of the original data (which is practically equivalent to Tukey’s conceptualization, but allows more conceptual breadth).

Rodgers [7] defined a taxonomy that organizes the various resampling methods in relation to two conceptual dimensions relevant to defining the empirical sampling distribution – resampling with or without replacement, and resampling to re-create the complete original sample or resampling to create a subsample. The jackknife method involves repeatedly drawing samples without replacement that are smaller than the original sample (whereas, for example, the bootstrap involves resampling with replacement to re-create the complete sample, and the randomization test involves

resampling without replacement to re-create the complete sample).

Mosteller and Tukey [5] provided four different examples of how the jackknife can be used, and a description of their examples is instructive. The first example (p. 139) used the jackknife to compute a confidence interval for a standard deviation obtained from 11 scores that were highly skewed. The second example (p. 142) applied the jackknife to estimate the 10% point in a whole population of measurements when the data were five repeated measures from 11 different individuals. The third example (p. 145) applied the jackknife to the log transforms of a sample of the populations of 43 large US cities for two purposes, to reduce bias in estimating the average size, and to compute the standard error of this estimate. Finally, the fourth example, a more complex application of the jackknife, obtained jackknifed estimators of the stability of a discriminant function (see **Discriminant Analysis**) defined to distinguish which of the *Federalist* papers were written by Hamilton and which ones by Madison.

As a more pragmatic and detailed example, consider a research design with random assignment of 10 subjects to a treatment or a control group ($n = 5$ per group). Did the treatment result in a reliable increase? Under the null hypothesis of no group difference, the 10 scores are sorted into one or the other group purely by chance. The usual formal evaluation of the null hypothesis involves computation of a two-independent-group t -statistic, which is a scaled version of a standardized mean difference between the two groups. This measure is compared to the distribution under the null hypothesis that is modeled by the theoretical t distribution. The validity of this statistical test rests on a number of parametric statistical assumptions, including normality and independence of errors, and homogeneity of variance. As an alternative that rests on fewer assumptions, the null distribution can be defined as an empirical t distribution using a jackknife or other resampling procedure.

To illustrate the use of the jackknife in this problem, we might decide to delete one observation from each group, and build an empirical sampling distribution from the many t statistics that would be obtained. One such null sampling model would involve sampling without replacement from the original ten observations until each group has four observations. There are $10!/[4! \times 4! \times 2!] = 3150$ different resamples that can be formed in this way. The

t -statistic is computed in each of these 3150 resampled datasets, providing a null sampling distribution for the t -statistic. (Alternatively, a random sample of these 3150 resampled datasets could be used to approximate this distribution [1]). If these 3150 t statistics are ordered, the position of the one that was actually observed in the sample can be found and an empirical P value computed by observing how extreme the actually observed t -value is in relation to this empirical null distribution. This P value – equivalent conceptually to the P value obtained from the appropriate parametric test, a two-independent group t Test – can be used to draw a statistical conclusion about group differences. Appropriately-chosen **quantiles** of the empirical t distribution can be used together with a jackknife estimate of the standard error of the mean difference to estimate the bounds of a confidence interval for the mean difference.

An example of a jackknife in a correlation and regression setting – one that follows exactly from Quenouille's original development of the jackknife – is presented in Efron [2, p. 9]. There, he uses the jackknife to estimate bias in the sample correlation between 1973 LSAT and GPAs of entering law school students.

How would a researcher decide when to use the jackknife, instead of one of the alternative resampling schemes, or the equivalent parametric procedure? There are several considerations, but no single definitive answer to this question. First, in some cases, the appropriate sampling model (e.g., resampling with or without replacement) is dictated by the nature of the problem. Settings in which the natural sampling model appears to involve sampling without replacement would naturally lead to the jackknife or the randomization test. Second, when the assumptions of parametric tests are met, there are both practical and statistical advantages to using those; when they are clearly and strongly violated, there are obvious advantages to the resampling methods. Third, the history and background of the jackknife originated from problems in the estimation and evaluation of bias; Efron [2, Chapter 2] gives a careful treatment of this focus. Finally, there are statistical reasons to prefer the bootstrap over the other resampling procedures in general hypothesis testing settings. Although the jackknife works well in such settings, the bootstrap will often outperform the jackknife.

References

- [1] Edgington, E.S. (1987). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [2] Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia.
- [3] Fisher, R.A. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.
- [4] Miller, R.G. (1974). The jackknife - a review, *Biometrika* **61**, 1–17.
- [5] Mosteller, F. & Tukey, J. (1977). *Data analysis and Regression*, Addison-Wesley, Reading.
- [6] Quenouille, M. (1949). Approximate tests of correlation in time series, *Journal of the Royal Statistical Society. Series B* **11**, 18–84.
- [7] Rodgers, J.L. (1999). The bootstrap, the jackknife, and the randomization test: a sampling taxonomy, *Multivariate Behavioral Research* **34**, 441–456.
- [8] Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract), *Annals of Mathematical Statistics* **29**, 614.

JOSEPH LEE RODGERS

Jonckheere–Terpstra Test

CLIFFORD E. LUNNEBORG

Volume 2, pp. 1007–1008

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Jonckheere–Terpstra Test

The m treatment levels must be ordered prior to any examination of the data. Assume independent random samples of response to treatment having sizes $n_1, n_2, \dots, n_m$ for the m ordered treatment levels. Let $x_{u,j}$ indicate the treatment response of the j th case receiving treatment level u .

For each of the $M = m(m-1)/2$ possible pairs of treatment levels, $u < v$, compute the quantity

$$J(u, v) = \sum_{j=1, \dots, n_u} \sum_{k=1, \dots, n_v} c(x_{u,j}, x_{v,k}), \quad (1)$$

where $c(x_{u,j}, x_{v,k})$ is 1, $1/2$, or 0, depending on whether $x_{u,j}$ is less than, equal to, or greater than $x_{v,k}$.

$J(u, v)$ will take its minimum value of 0 if every response at treatment level v is smaller than the smallest response at level u . It will take its maximum value of $(n_u \times n_v)$ if every response at treatment level v is greater than the largest response at level u . Larger values of $J(u, v)$ are consistent with response magnitude increasing between treatment levels u and v .

The Jonckheere–Terpstra test statistic [2] is the sum of the $J(u, v)$ terms over the M distinct pairs of treatment levels:

$$J = \sum_{u=1, (m-1)} \sum_{v=u, m} J(u, v). \quad (2)$$

Large values of J favor the alternative hypothesis.

Under the null hypothesis, the treatment responses are exchangeable among the treatment levels. Thus, the Jonckheere–Terpstra test is a permutation test (see **Permutation Based Inference**) – its null reference distribution consists of the values of J for all possible permutations of the $N = n_1 + n_2 + \dots + n_m$ responses among the m treatments, preserving the treatment level group sizes. For larger sample sizes, a Monte Carlo (see **Monte Carlo Simulation**) approximation to the null reference

distribution can be produced by randomly sampling a very large number of the possible permutations, for example, 5000. This test is available in the XactStat (www.cytel.com) and SC (www.mole-soft.demon.co.uk) packages.

For a range of small sample sizes, tables of critical values of J are provided in [1]. These are strictly valid only in the absence of ties. An asymptotic approximation for large sample sizes is also described in [1].

Example

The data in the table below are taken from [3]. A sample of 30 student volunteers were taught a finger-tapping task and then randomized to one of three treatments. They received a drink containing 0 mg caffeine, 100 mg caffeine, or 200 mg caffeine and were then retested on the finger-tapping task. The substantive hypothesis was that increased dosages of caffeine would be accompanied by increased tapping speeds (Table 1).

The jonckex function in SC package reports a value of J of 229.5. The P value for the exact (permutation) test was 0.0008, and for the large-sample normal approximation the value was 0.0011. Tapping speed clearly increases with caffeine consumption.

References

- [1] Hollander, M. & Wolfe, D.A. (2001). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.
- [2] Jonckheere, A.R. (1954). A distribution-free k -sample test against ordered alternatives, *Biometrika* **41**, 133–145.
- [3] Lunneborg, C.E. (2000). *Data Analysis by Resampling*, Duxbury, Pacific Grove.

CLIFFORD E. LUNNEBORG

Table 1 Tapping rate as a function of caffeine intake

Caffeine (mg)	Rate (taps per minute)
0	242, 245, 244, 248, 247, 248, 242, 244, 246, 242
100	248, 246, 245, 247, 248, 250, 247, 246, 243, 244
200	246, 248, 250, 252, 248, 250, 246, 248, 246, 250

Kendall, Maurice George

BRIAN S. EVERITT

Volume 2, pp. 1009–1010

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kendall, Maurice George

Born: September 6, 1907, in Northamptonshire, UK.

Died: March 29, 1983, in Surrey, UK.

At his first secondary school, Derby Central School, Kendall was initially more interested in languages than in mathematics. But near the end of his secondary schooling his mathematical aptitude began to appear, and the improvement was so great that he was awarded a scholarship to study at St John's College, Cambridge, where he graduated as a mathematics Wrangler in 1929. From Cambridge, Kendall entered and passed the examinations to enter the administrative class of the Civil Service and joined the Ministry of Agriculture and Fisheries. It was here that Kendall first became involved in statistical work.

It was in 1935 that Kendall, while spending part of his holiday reading statistics books at the St John's College library, met **G. Udny Yule** for the first time. The result of this meeting was eventually the recruitment of Kendall by Yule to be joint author of a new edition of *An Introduction to the Theory of Statistics*, first published in 1911, and appearing in its 14th edition in the 1950s [6].

The work with Yule clearly whetted Kendall's appetite for mathematical statistics and he attended lectures at University College, London and began publishing papers on statistical topics. Early during World War II, Kendall left the Civil Service to take up the post of statistician to the British Chamber of Shipping, and despite the obvious pressures of such a war time post, Kendall managed, virtually single-handedly, to produce Volume One of *The Advanced Theory of Statistics* in 1943 and Volume Two in 1946. For the next 50 years, this text remained the standard work for generations of students of mathematical statistics and their lecturers and professors.

In 1949, Kendall moved into academia, becoming Professor of Statistics at the London School of Economics where he remained until 1961. During this time, he published a stream of high-quality papers on a variety of topics including the theory of

k-statistics, **time series** and rank correlation methods (*see* **Rank Based Inference**). His monograph on the latter remains in print in the twenty-first century [2]. Kendall also helped organize a number of large sample survey projects in collaboration with governmental and commercial agencies. Later in his career, Kendall became Managing Director and Chairman of the computer consultancy, SCI-CON. In the 1960s, he completed the rewriting of his major book into three volumes, which was published in 1966 [1]. In 1972, he became Director of the World Fertility Survey. Kendall was awarded the Royal Statistical Society's Guy Medal in Gold and in 1974 a knighthood for his services to the theory of statistics. Other honors bestowed on him include the presidencies of the Royal Statistical Society, the Operational Research Society, and the Market Research Society.

Sixty years on from the publication of the first edition of *The Advanced Theory of Statistics* the material of the book remains (although in a somewhat altered format [3–5]) the single most important source of basic information for students of and researchers in mathematical statistics. The book has remained in print for as long as it has, largely because of Kendall's immaculate organization of its contents, and the quality and style of his writing.

References

- [1] Kendall, M.G. & Stuart, A. (1966). *The Advanced Theory of Statistics*, Griffin Publishing, London.
- [2] Kendall, M.G. & Dickinson, J. (1990). *Rank Correlation Method*, 5th Edition, Oxford University Press, New York.
- [3] McDonald, S., Wells, G. & Arnold, H. (1998). *Kendall's Advanced Theory of Statistics*, Vol. 2A, *Classical Inference*, 6th Edition, Arnold, London.
- [4] O'Hagan, T. & Forster, T. (2004). *Kendall's Advanced Theory of Statistics*, Vol. 2B, *Bayesian Inference*, 2nd Edition, Arnold, London.
- [5] O'Hagan, T., Ord, K., Stuart, A. & Arnold, S. (2004). *Kendall's Advanced Theory of Statistics. 3-Volume Set*, Arnold, London.
- [6] Yule, G.U. & Kendall, M.G. (1951). *An Introduction to the Theory of Statistics*, 14th Edition, Griffin Publishing, London.

BRIAN S. EVERITT

Kendall's Coefficient of Concordance

ANDY P. FIELD

Volume 2, pp. 1010–1011

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kendall's Coefficient of Concordance

Kendall's Coefficient of Concordance, W , is a measure of the agreement between several judges who have rank ordered a set of entities. In this sense, it is similar to an **intraclass correlation** for ranked data.

As an illustration, we use the same example as for intraclass correlation (from [1]) in which lecturers were asked to mark eight different essays on a percentage scale. Imagine, however, that the lecturers had rank ordered the essays (assigning a rank of 1 to the best and a rank of 8 to the worst) as in Table 1.

If the lecturers agree, then the total ranks for each essay will vary. For example, if all four lecturers thought essay 6 was the best, then the sum of ranks would be only 4, and if they thought essay 3 was the worst, then it would have a total rank of $8 \times 4 = 32$. However, if there is a lot of disagreement then particular essays will have both high and low ranks assigned to them and the resulting totals will be roughly equal. In fact, if there is maximal disagreement between judges, the totals for each essay will be the same.

Kendall's statistic represents the ratio of the observed variance of the total ranks of the ranked entities to the maximum possible variance of the total ranks. In this example, it would be the variance of total ranks for the essays divided by the maximum possible variance in total ranks of the essays. The first step is, therefore, to calculate the variance of total ranks for essays. To estimate this variance, we use a sum of squared error. In general terms, this is the squared difference between an observation and

the mean of all observations:

$$ss = \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1)$$

The mean of the total ranks for each essay can be calculated in the usual way, but it is also equivalent to $k(n+1)/2$ in which k is the number of judges and n is the number of things being ranked.

This gives us:

$$\begin{aligned} SS_{\text{Rank Totals}} &= (19 - 18)^2 + (15 - 18)^2 + (25 - 18)^2 \\ &\quad + (21 - 18)^2 + (21 - 18)^2 + (14 - 18)^2 \\ &\quad + (17 - 18)^2 + (12 - 18)^2 \\ &= 170. \end{aligned} \quad (2)$$

This value is then divided by an estimate of the total possible variance. We could get this value by simply working out the sum of squared errors for row totals when all raters agree. In this case, the row totals would be 4 (all agree on a rank of 1), 8 (all agree a rank of 2), 12, 16, 20, 24, 28, 32 (all agree a rank of 8). The resulting sum of squares would be:

$$\begin{aligned} SS_{\text{Max}} &= (4 - 18)^2 + (8 - 18)^2 + (12 - 18)^2 \\ &\quad + (16 - 18)^2 + (20 - 18)^2 + (24 - 18)^2 \\ &\quad + (28 - 18)^2 + (32 - 18)^2 \\ &= 672. \end{aligned} \quad (3)$$

The resulting coefficient would be:

$$W = \frac{SS_{\text{Rank Totals}}}{SS_{\text{Max}}} = \frac{170}{672} = 0.25. \quad (4)$$

I have done this just to show what the coefficient represents. However, it is a rather cumbersome method because of the need to calculate the maximum possible sums of squares. Thanks to some clever mathematics, we can avoid calculating the maximum possible sums of squares in this long-winded way and simply use the following equation:

$$W = \frac{12 \times SS_{\text{Rank Totals}}}{k^2, (n^3 - n)}, \quad (5)$$

where k is the number of judges and n is the number of things being judged. We would obtain the same answer:

$$W = \frac{12 \times 170}{16(512 - 8)} = \frac{2040}{8064} = 0.25. \quad (6)$$

Table 1 Eight essays ranked by four lecturers

Essay	Dr. Field	Dr. Smith	Dr. Scrote	Dr. Death	Total
1	7	7	3	2	19
2	6	6	2	1	15
3	8	8	6	3	25
4	5	5	7	4	21
5	4	4	8	5	21
6	1	3	4	6	14
7	3	2	5	7	17
8	2	1	1	8	12
					Mean = 18

2 Kendall's Coefficient of Concordance

A significance test can be carried out using a chi-square statistic with $n - 1$ degrees of freedom $\chi_W^2 = k(n - 1)W$. In this case, the test statistic of 7 is not significant.

W is constrained to lie between 0 (no agreement) and 1 (perfect agreement), but interpretation is difficult because it is unclear how much sense can be made of a statement such as 'the variance of the total ranks of entities was 25% of the maximum possible'. Significance tests are also relatively meaningless because the levels of agreement usually viewed as good in the social sciences are way above what would be required for significance. However, W can be converted into the mean **Spearman correlation coefficient** (see [2]):

$$\bar{r}_s = \frac{kW - 1}{k - 1}. \quad (7)$$

If we computed the Spearman correlation coefficient between all pairs of judges, this would be the average value. In this case, we would get $((4 \times 0.25) - 1)/3 = 0$. The clear interpretation here is that on average pairs of judges did not agree!

References

- [1] Field, A.P. (2005). *Discovering Statistics Using SPSS*, 2nd Edition, Sage, London.
- [2] Howell, D.C. (2002). *Statistical Methods for Psychology*, 5th Edition, Duxbury, Belmont.

ANDY P. FIELD

Kendall's Tau – τ

DIANA KORNBROT

Volume 2, pp. 1011–1012

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kendall's Tau – τ

Kendall's tau, τ , like **Spearman's rho**, ρ is a measure of correlation based on ranks (see **Rank Based Inference**). It is useful when the raw data are themselves ranks, as for example job applicants, or when the data is ordinal. Examples of ordinal data include common rating scales based on responses ranging from 'strongly disagree' to 'strongly agree'.

Kendall's τ is based on the number of inversions (swaps) needed to get the ranks of both variables in the same or exactly the opposite order. For example, consider a data set with raw data pairs (20,17; 23,24; 28,25; 31,40; 90,26). Ranks of the X values are in order (1, 2, 3, 4, 5) and ranks of the corresponding Y values are (1, 2, 3, 5, 4). If the last two Y values are inverted (swapped), then ranks for X and Y will be identical.

Calculation

In order to calculate τ , it is first necessary to rank both the X and the Y variable. The number of inversions of ranks can then be counted using a simple diagram. The ranks of the X variables are ordered with the corresponding Y ranks underneath them. Then a line is drawn from each X rank to its corresponding Y rank. The number of inversions, I , will be equal to the number of crossing points. Two $N = 5$ examples are shown in Figure 1, with the relevant sequence of inversions in bold.

Example 1. Number of inversions, $I = 1$, $\tau = 0.80$

Raw X	20	23	28	31	90
Raw Y	17	24	25	40	26
Rank X	1	2	3	4	5
Rank Y	1	2	3	5	4
Invert 1	1	2	3	4	5

If the number of inversions is I , then τ is given by (1).

$$\tau = 1 - \frac{2I}{N(N-1)/2} \quad (1)$$

Equation (1) overestimates τ if there are ties. Procedures (cumbersome) for adjusting ties are given in [2]. Corrections for ties are implemented in standard statistical packages.

Hypothesis Testing

For the null hypothesis of no association, that is, $\tau = 0$ and $N > 10$, τ is approximately normally distributed with standard error, s_τ , given by (2)

$$s_\tau = \sqrt{\frac{2(2N+5)}{9N(N-1)}} \quad (2)$$

Accurate tables for $N \leq 10$ are provided by Kendall & Gibbons [2], but not used by many standard packages, for example, SPSS and JMP.

Confidence intervals and hypotheses about values of τ other than 0 can be obtained by noting that the Fisher transformation gives a statistic, z_r , that is normally distributed with variance $1/(N-3)$.

$$z_r = \frac{1}{2} \ln \left[\frac{1+\tau}{1-\tau} \right] \quad (3)$$

Comparison with Pearson's r and Spearman's ρ

For normally distributed data with a linear relation, parametric tests based on r (see **Pearson Product**

Example 2. Number of inversions, $I = 7$, $\tau = -0.60$

Raw X	20	23	28	31	90
Raw Y	40	24	25	26	17
Rank X	1	2	3	4	5
Rank Y	5	2	3	4	1
Invert 1	2	5	3	4	1
Invert 2	2	3	5	4	1
Invert 3	2	3	4	5	1
Invert 4	2	3	4	1	5
Invert 5	2	3	1	4	5
Invert 6	2	1	3	4	5
Invert 7	1	2	2	4	5

Figure 1 Examples of how to calculate the number of rank inversions in order to calculate τ

2 Kendall's Tau – τ

Moment Correlation) are usually more powerful than rank tests based on either τ or r_s . Authorities [1, 2] recommend τ over r_s as the best rank-based procedure, but r_s is far easier to calculate if a computer is not available. Kendall's τ has a simple interpretation – it is the degree to which the rankings of the two variables, X and Y , agree.

- [2] Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th Edition, Edward Arnold, London & Oxford.

DIANA KORBROT

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.

Kernel Smoothing

MARTIN L. HAZELTON

Volume 2, pp. 1012–1017

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kernel Smoothing

Kernel smoothing is a methodology for exposing structure or trend in data. It can be applied in a number of settings, including *kernel density estimation*, where the aim is to estimate the probability density underlying a variable, and *kernel regression*, where the aim is to estimate the conditional mean of a response variable given one or more covariates. Kernel smoothing is a nonparametric technique, in that it does not rely on the assumption of some particular parametric model for the data at hand. For instance, kernel regression does not rely on the existence of a linear (or other low degree polynomial) relationship existing between mean response and covariate. This makes kernel smoothing a powerful tool for **exploratory data analysis** and a highly applicable technique in circumstances where classical parametric models are clearly inappropriate.

Kernel Density Estimation

The problem of estimating a probability density function, or *density* for short, is a fundamental one in statistics. Suppose that we observe data $x_1, \dots, x_n$ on a single variable from which we wish to estimate the underlying density f . If we are willing to assume that f is of some given parametric form (e.g. normal, gamma), then the problem reduces to one of estimating the model parameters (e.g. mean μ and variance σ^2 in the normal case). This is a rather rigid approach in that the data play a limited role in determining the overall shape of the estimated density. Much greater flexibility is afforded by the **histogram**, which can be regarded as a simple nonparametric density estimator (when scaled to have unit area below the bars).

Although the histogram is widely used for data visualization, this methodology has a number of deficiencies. First, a histogram is not smooth, (typically) in contrast to the underlying density. Second, histograms are dependent on the arbitrary choice of *anchor point*; that is, the position on the x-axis at which the left hand edge of the first bin (counting from the left) is fixed. Third, histograms are dependent on the choice of bin width. We illustrate these problems using data on the age (in years) at onset of schizophrenia for a sample of 99 women. Four

histograms are displayed in Figure 1, using a variety of anchor points and bin widths. The importance of the anchor point is demonstrated by Figures 1(a) and 1(b), which differ only in a 2.5-year shift in the bin edges. The histogram shown in Figure 1(a) suggests that the distribution is skewed with a long right tail, whereas the histogram in Figure 1(b) indicates a bimodal structure. The effect of bin width is illustrated by comparing Figures 1(c) and 1(d) with the first pair of histograms. Notice that the thin bins in Figure 1(c) result in a rather noisy histogram, whereas the large bin width in Figure 1(d) produces a histogram that lacks the detail of Figures 1(a) and 1(b).

One may think of a histogram as being composed of rectangular building blocks. Each datum carries a block of area (or weight) $1/n$, which is centered at the middle of the bin in which the datum falls and stacked with other blocks in that bin. It follows that the histogram density estimate can be written as

$$\hat{f}_{\text{hist}}(x) = \sum_{i=1}^n \frac{1}{n} K(x - t_i) \quad (1)$$

where

$$K(x) = \begin{cases} \frac{1}{b} & -\frac{b}{2} < x \leq \frac{b}{2} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

in which b is the bin width and t_i is the midpoint of the bin in which x_i falls. An obvious generalization of (1) would be to replace t_i with x_i , so that each building block is centered at its respective data point rather than at a bin midpoint. The effect of this modification is illustrated in Figure 2(a), which shows the resulting estimate for the age at onset of schizophrenia data. (The data themselves are displayed as a *rug* on this plot.) Although this change solves the anchor point problem, the estimate itself remains jagged. We can obtain a smooth estimate by replacing the rectangular blocks by smooth functions or *kernels*. We then obtain the *kernel density estimate*,

$$\hat{f}(x) = \sum_{i=1}^n \frac{1}{nh} K\left(\frac{x - x_i}{h}\right), \quad (3)$$

where the kernel K is itself a smooth probability density function. Typically K is assumed to be unimodal and satisfy the conditions $K(x) = K(-x)$ and $\int K(x)x^2 dx = 1$; the normal density is a popular choice. The parameter h in (3) is usually called the

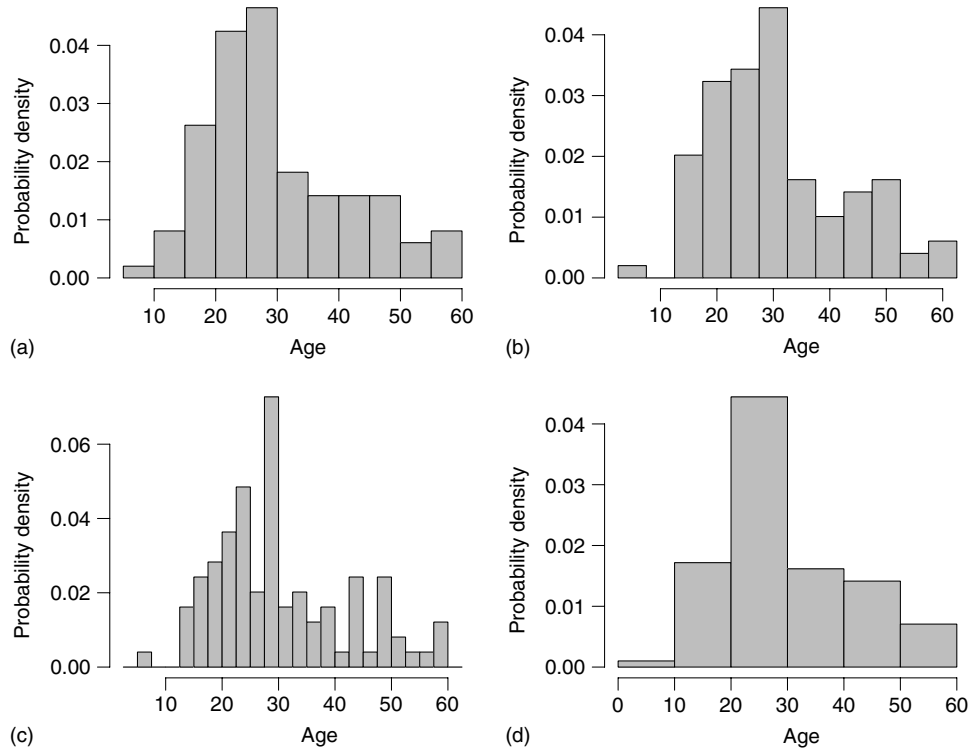


Figure 1 Histograms of age at onset of schizophrenia for a sample of 99 women. Histograms (a) and (b) both use a bin width of 5 years, but differ in anchor point. Histograms (c) and (d) use bin widths of 2.5 and 10 years respectively

bandwidth. It scales the kernel function and hence controls the smoothness of the density estimate in much the same manner as the bin width does in a histogram. Figures 2(b), 2(c), and 2(d) show kernel density estimates for the age data constructed using bandwidths $h = 10$, $h = 2$ and $h = 3.7$ respectively. The estimate in Figure 2(b) is oversmoothed, obscuring important structure in the data (particularly in the range 40–50 years), while estimate shown in Figure 2(c) is undersmoothed, leaving residual wiggles in the estimate. The degree of smoothing in the estimate in Figure 2(d) appears to be more appropriate. The shape of this final estimate (Figure 2(d)) suggests that the underlying distribution may be a mixture of two unimodal components, one centered at about 25 years and the other around 45 years. A commentary on the data and an analysis based on this type of model is given in **finite mixture distributions**.

As we have seen, kernel density estimation overcomes the deficiencies of the histogram in terms of smoothness and anchor point dependence, but

choosing an appropriate value for the bandwidth remains an issue. A popular approach to bandwidth selection is to seek to minimize the mean integrated squared error of $\hat{f}$:

$$\text{MISE}(h) = E \int \{\hat{f}(x; h) - f(x)\}^2 dx. \quad (4)$$

Here the notation $\hat{f}(x; h)$ emphasizes the dependence of the density estimate on h . Finding the minimizer, h_0 , of (4) is difficult in practice because $\text{MISE}(h)$ is a functional of the unknown target density f , and is awkward to manipulate from a mathematical perspective. A number of sophisticated methods for resolving these problems have been developed, amongst which the Sheather–Jones bandwidth selection technique is particularly well regarded [4]. The Sheather–Jones methodology was used to select the bandwidth for the density estimate in Figure 2(d).

Kernel density estimation has a wide variety of uses. It is a powerful tool for data visualization and exploration, and should arguably replace the

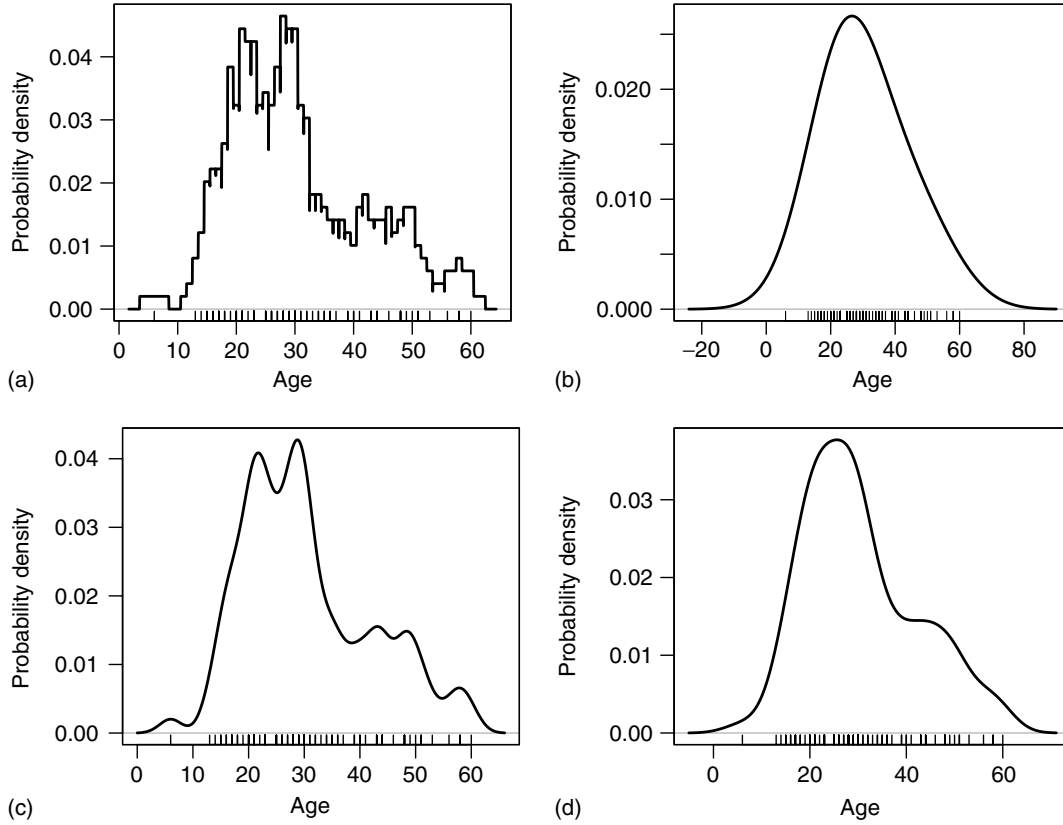


Figure 2 Kernel density estimates of age at onset of schizophrenia for a sample of 99 women. Estimate (a) is constructed using a rectangular kernel, while estimates (b), (c), and (d) use a normal kernel. The bandwidths in the last three estimates are $h = 10$, $h = 2$ and $h = 3.75$ respectively

histogram for this type of purpose. Kernel density estimates can also be used in **discriminant analysis**, **goodness-of-fit testing**, and testing for multimodality (or *bump hunting*) [5]. These applications often involve the multivariate version of (3); see [3] for details.

Kernel Regression

The data displayed in the **scatterplots** in Figure 3 are the annual number of live births in the United Kingdom from 1865 to 1914. Classical parametric regression models would struggle to represent the trend in these data because a linear or low-order polynomial (e.g. quadratic or cubic) fit is clearly inappropriate. The trouble with this approach to regression for these data is that one is required to

specify a functional form that describes the global behavior of the data, whereas the scatterplot displays a number of interesting features that are localized to particular intervals on the x-axis. For example, there is an approximately linear increase in births between 1865 and 1875, and a bump in the number of births between 1895 and 1910. This suggests that any model for the trend in the data should be local in nature, in that estimation at a particular year x_0 should be based only on data from years close to x_0 .

The preceding discussion motivates interest in the **nonparametric regression model**

$$Y_i = m(x_i) + \varepsilon_i \quad (i = 1, \dots, n), \quad (5)$$

where $x_1, \dots, x_n$ are covariate values, $Y_1, \dots, Y_n$ are corresponding responses, and $\varepsilon_1, \dots, \varepsilon_n$ are independent error terms with zero mean and common variance σ^2 . The regression function m describes

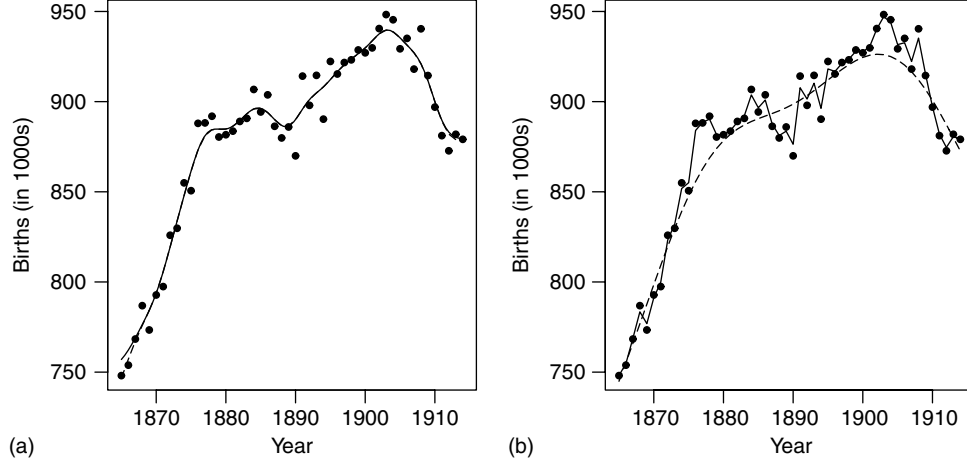


Figure 3 Kernel regression estimates for the UK birth data. (a) Displays the Nadaraya–Watson and local linear kernel estimates as solid and dashed lines respectively; both estimated were constructed with bandwidth $h = 1.61$. (b) Displays local linear kernel estimates obtained using bandwidths $h = 5$ (dashed line) and $h = 0.5$ (solid line)

the conditional mean of Y given x , and is a smooth function for which no particular parametric form is assumed. Local estimation of m may be achieved using a weighted average of the responses,

$$\hat{m}(x) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}, \quad (6)$$

where the weights $w_1, \dots, w_n$ are selected so that w_i decreases as x_i becomes more distant from the estimation point x . The *Nadaraya–Watson* kernel regression estimator (originally proposed in [2] and [7]) defines the weights using kernel functions through

$$w_i = \frac{1}{h} K\left(\frac{x - x_i}{h}\right), \quad (7)$$

where K is typically a unimodal probability density with $K(x) = K(-x)$ and $\int K(x)x^2 dx = 1$, such as the standard normal density. Clearly (7) achieves the desired effect that data points with covariate distant to the estimation point x have negligible influence on the estimate. The bandwidth h scales the kernel in the same fashion as for density estimation. The Nadaraya–Watson regression estimate for the birth data is displayed as the solid line in Figure 3(a). It picks up the general trend in birth numbers through

time while smoothing out the haphazard year-to-year variation.

The Nadaraya–Watson estimator, for which we will use the specific notation $\hat{m}_0(x)$, may be derived in terms of a weighted least squares problem. Specifically, $\hat{m}_0(x) = \hat{\beta}_0$ where $\hat{\beta}_0$ minimizes

$$\sum_{i=1}^n (Y_i - \beta_0)^2 \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \quad (8)$$

One may therefore think of the Nadaraya–Watson estimate at the point x as the result of fitting a constant regression function (i.e. just an intercept, β_0) by weighted least squares. We can generalize this approach by fitting a linear regression (see **Multiple Linear Regression**) at the point x , minimizing the revised weighted sum of squares

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1(x_i - x))^2 \frac{1}{h} K\left(\frac{x - x_i}{h}\right). \quad (9)$$

If $\hat{\beta}_0, \hat{\beta}_1$ are the minimizers of (9), then the *local linear kernel regression estimate* at the point x , $\hat{m}_1(x)$, is defined by $\hat{m}_1(x) = \hat{\beta}_0$. The Nadaraya–Watson (or *local constant kernel*) estimate and the local linear kernel estimate often give rather similar results, but the latter tends to perform better close to edges of the data set. The local linear kernel estimate is displayed as a dashed regression line in Figure 3(a).

Although it is difficult to distinguish it from the Nadaraya–Watson estimate over most of the range of the data, close inspection shows that the local linear estimate follows the path of the data more appropriately at the extreme left hand end of the plot.

As with density estimation, the choice of bandwidth is crucial in determining the quality of the fitted kernel regression. An inappropriately large bandwidth will oversmooth the data and obscure important features in the trend, while a bandwidth that is too small will result in a rough, noisy kernel estimate. This is illustrated in Figure 3(b), where local linear kernel estimates were constructed using bandwidths $h = 5$ and $h = 0.5$ giving the dashed and solid lines respectively. A number of data driven methods for selecting h have been proposed, including a simple **cross-validation** approach that we now describe. Suppose we fit a kernel regression using all the data except the i th observation, and get an estimate $\hat{m}^{[-i]}(x; h)$, where the dependence on the bandwidth h has been emphasized in the notation. If h is a good value for the bandwidth, then $\hat{m}^{[-i]}(x_i; h)$ should be a good predictor of Y_i . It follows that the cross-validation function

$$CV(h) = \sum_{i=1}^n (\hat{m}^{[-i]}(x_i; h) - Y_i)^2 \quad (10)$$

is a reasonable performance criterion for h , and the minimizer $\hat{h}_{CV}$ of $CV(h)$ should be a suitable bandwidth to employ. This cross-validation approach was used to provide the bandwidth $h = 1.61$ for the local linear (and local constant) kernel estimates in Figure 3(a).

We have focused our attention on models with a single covariate, but kernel regression methods may be extended to cope with multiple covariates. For example, if the response Y is related to covariates x and z , then we might use a **general additive model** (GAM),

$$Y_i = m(x_i) + l(z_i) + \varepsilon_i \quad (i = 1, \dots, n), \quad (11)$$

where m and l are smooth regression functions which may be estimated using kernel methods. Alternatively, we may mix nonparametric and parametric terms to give a semiparametric regression model, such as

$$Y_i = m(x_i) + \beta z_i + \varepsilon_i \quad (i = 1, \dots, n). \quad (12)$$

Kernel methods may also be applied in **generalized linear models** to fit nonlinear terms in the linear predictor; see [1], for example.

Conclusions

Kernel methods can be used for nonparametric estimation of smooth functions in a wide variety of situations. We have concentrated on estimation of probability densities and regression functions, but kernel smoothing can also be applied to estimation of other functions in statistics such as spectral densities and hazard functions. There are now a number of books that provide a comprehensive coverage of kernel smoothing. Bowman and Azzalini's monograph [1] is recommended for those seeking an applications-based look at the subject, while Wand and Jones [6] provide a more theoretical (though still accessible) perspective. Both kernel density estimation and kernel regression methods are implemented in many software packages, including R, S-Plus, SAS, and Stata.

References

- [1] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Clarendon Press, Oxford.
- [2] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Applications* **9**, 141–142.
- [3] Scott, D.W. (1992). *Multivariate Density Estimation; Theory, Practice and Visualization*, Wiley-Interscience, New York.
- [4] Sheather, S.J. & Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society. Series B* **53**, 683–690.
- [5] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer, New York.
- [6] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*, Chapman & Hall, London.
- [7] Watson, G.S. (1964). Smooth regression analysis, *Sankhyā. Series A* **26**, 359–372.

(See also **Scatterplot Smoothers**)

MARTIN L. HAZELTON

***k*-means Analysis**

DAVID WISHART

Volume 2, pp. 1017–1022

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

***k*-means Analysis**

Cluster analysis is a term for a group of multivariate methods that explore the similarities or differences between cases in order to find subgroups containing relatively homogenous cases (see **Cluster Analysis: Overview**). The cases may be, for example, patients with various symptoms, ideas arising from a focus group, clinics with different types of patient. There are two main types of cluster analysis: *optimization methods* that produce a single partition of the data, and **hierarchical clustering** which forms a nested series of mutually exclusive partitions or subgroups of the data. Divisions or fusions in hierarchical clustering, once made, are irrevocable. When an agglomerative algorithm has joined two cases they cannot subsequently be separated, and when a divisive algorithm has made a split they cannot be reunited. By contrast, optimisation cluster methods reallocate cases iteratively to produce a single *optimal* partition of the data into mutually exclusive clusters, and can therefore be used *inter alia* to correct for inefficient assignments made during hierarchical clustering.

Probably the most common optimisation method is *k*-means analysis, first described by Forgey [5], MacQueen [8] and Ball & Hall [2]. It has a compelling simplicity, namely, to find a partition of the cases into *k* clusters such that each case is closer to the mean of the cluster to which it is assigned than to any other cluster. However, it will be shown below that this simplicity conceals a computational flaw that does not guarantee convergence. For this reason we refer to it as ‘naïve’ *k*-means analysis. The method requires that a case *i* is moved from a cluster *p*, the cluster to which it is currently assigned, to another cluster *q* if the case is closer to the mean of cluster *q* than it is to the mean of cluster *p*; that is, if

$$d_{ip} > d_{iq}. \quad (1)$$

Naïve *k*-means analysis is usually implemented in software as follows:

1. Choose some initial partition of the cases into *k* clusters. This can be a tree section, for example, by hierarchical clustering, *k* selected ‘seed’ cases, or a random assignment of the cases to *k* clusters.
2. Compute the distance from every case to the mean of each cluster, and assign the cases to their nearest clusters.

3. Recompute the cluster means following any change of cluster membership at step 2.
4. Repeat steps 2 and 3 until no further changes of cluster membership occur in a complete iteration. The procedure has now converged to a stable *k*-cluster partition.

An Exact *k*-means Algorithm

It is a common misconception that *k*-means analysis, as described above, optimizes (i.e. minimizes) the *Euclidean Sum of Squares*, or trace (**W**), where **W** is the within-clusters dispersion matrix. The Euclidean Sum of Squares E_p for a cluster *p* is defined as:

$$E_p = \sum_{i \in p} n_i \sum_j (x_{ij} - \mu_{pj})^2, \quad (2)$$

where x_{ij} is the value of the *j*th variable for case *i* in cluster *p*, μ_{pj} is the mean of the *j*th variable for the *p*th cluster, and n_i is a weight for case *i* (usually 1).

The total Euclidean Sum of Squares is

$$E = \sum_p E_p. \quad (3)$$

The computational flaw in naïve *k*-means analysis arises from an incorrect assumption that if a case is moved to a cluster whose mean is closer than the mean of the cluster to which it is currently assigned, following test (1), then the move reduces *E*. This neglects to take account of the changes to the means of the clusters caused by reassigning the case. To minimize *E* unequivocally, a case should only be moved from a cluster *p* to another cluster *q* *if and only if the move reduces E* in (3). That is, iff:

$$E_p + E_q > E_{p-i} + E_{q+i}. \quad (4)$$

We call (4) the ‘exact assignment test’ for minimum *E*. It is not the same as assigning a case *i* to the nearest cluster mean, as in naïve *k*-means analysis, because moving any case from cluster *p* to cluster *q* also changes the means of *p* and *q*; and in certain circumstances, these changes may actually increase *E*. Rearranging (4) thus,

$$E_p - E_{p-i} > E_{q+i} - E_q, \quad (5)$$

which is equivalent to moving a case *i* from cluster *p* to cluster *q* iff:

$$I_{p-i,i} > I_{q,i}, \quad (6)$$

2 k -means Analysis

where $I_{p-i,i}$ is the increase in E resulting from the addition of case i to the complement $p-i$ of cluster p , and $I_{q,i}$ is the increase in E resulting from the addition of case i to cluster q . Beale [3] gives the increase I_{pq} in E on the union of two clusters p and q as follows:

$$I_{pq} = n_p n_q \sum_j \frac{(\mu_{pj} - \mu_{qj})^2}{(n_p + n_q)}, \quad (7)$$

where n_p and n_q are the sum of case weights of clusters p and q (usually their sizes), and μ_{pj} and μ_{qj} are the means of the clusters p and q for variable j . Upon substituting (7) for the comparison of a case i with clusters $p-i$ and q , (6) becomes

$$\frac{n_p n_i d_{pi}^2}{n_p - n_i} > \frac{n_q n_i d_{qi}^2}{n_q + n_i}. \quad (8)$$

Where the cases have equal weight, (8) simplifies to

$$\frac{n_p d_{pi}^2}{n_p - 1} > \frac{n_q d_{qi}^2}{n_q + 1}. \quad (9)$$

We refer to (8) as the ‘exact assignment test’ for k -means analysis. Since E is a sum of squares, and each reassignment that satisfies (8) actually reduces E by the difference between the two sides of (8), the procedure must converge in finite time. This is because E , being bounded by zero, cannot be indefinitely reduced.

It will be evident that the exact assignment test (8) or (9) is not equivalent to moving a case i from cluster p to cluster q if $d_{ip} > d_{iq}$, as in naïve k -means analysis, yet the naïve test is actually how k -means analysis is implemented in many software packages. In geometrical terms, moving a case i from cluster p to cluster q pulls the mean $\mu_{q+i,j}$ of q towards i and pushes the mean $\mu_{p-i,j}$ of p away from i . This causes the distances from the means of some cases in clusters p and q to increase, such that E may actually increase as a result of moving a case from cluster p to cluster q even if $d_{ip} > d_{iq}$. It does not usually happen with small data sets, but it can occur with large data sets where boundary cases may oscillate between two or more clusters in successive iterations and thereby the naïve procedure fails to converge.

This is why some standard k -means software also employ an *ad hoc* ‘movement of means’ convergence criterion. It stops the procedure when a complete iteration does not move a cluster centre by more than

a specified percentage, which must be greater than zero, of the smallest distance between any of the initial cluster centres. This is not a convergence criterion at all, but merely a computational device to prevent the program from oscillating between two or more solutions and hence iterating indefinitely. MacQueen [8] recognized this defect in his naïve k -means algorithm but failed to formulate the exact assignment test (8) that corrects for it. Anderberg [1] reports only the naïve k -means procedure. Hartigan [7] reports the exact assignment test for minimum E correctly, but curiously implements the naïve k -means test in the program listed in his appendix. Späth [9] correctly describes the exact assignment test for minimum E and implements it correctly in his program. Everitt et al. [4] give a worked example of naïve k -means analysis without mentioning the exact assignment test.

This important modification of the assignment test ensures that *exact* k -means analysis will always converge in finite time to a specific minimum E value, whereas *naïve* k -means analysis may only ever achieve interim higher values of E without actually converging. Furthermore, those k -means programs that utilize the ‘movement of means’ convergence criterion do not report the value of E because it can differ from one run to the next depending upon the order of the cases, the starting conditions and the direction in which the initial cluster means migrate.

A General Optimization Procedure

The exact assignment test can be generalized for any well-defined clustering criterion that is amenable to optimization, for example, to minimize the sum of within-group squared distances, similar to average-link cluster analysis in hierarchical clustering, or UPGMA. Another clustering criterion that has been proposed for binary variables is an information content statistic to be minimized by the relocation of a case from one cluster to another. The general procedure for optimization is now defined for any clustering criterion f to be minimized, as follows: Given any interim classification C of the cases, a case i currently assigned to a cluster p should be moved to another cluster q if:

$$f(C|i \in p) > f(C|i \in q). \quad (10)$$

The generalized exact assignment test (10) will always converge to a minimum f for a classification

C provided that there is a lower bound below which f cannot reduce further, such as zero as for the Euclidean Sum of Squares E specified in (2) and (3) because E is a sum of squares and for the within-cluster sum of squared distances.

Convergence

This does not necessarily mean that the final classification C obtained at convergence is *optimal*, since the *k*-means procedure will also converge to sub-optimal classifications that, nevertheless, satisfy the exact assignment test, see Table 1 below. For this reason, it is important to evaluate several different classifications obtained by *k*-means analysis for the same data, as advocated by Gordon [6]. Some writers have proposed fairly complex techniques for choosing initial seed points as the starting classification for *k*-means, and for combining and splitting clusters during the assignment process. Their general approach is to merge similar clusters or split heterogeneous clusters, with the object of finding a value of k , the best number of clusters, that yields a stable homogeneous classification. However, the choice of initial seed points, or the starting classification, can also be adequately tested by generating several random initial classifications, and it is open to the user to evaluate different values of k for the **number of clusters**.

It is well known that *k*-means analysis is also sensitive to the order in which the cases are tested for relocation, and some writers advocate varying this order randomly. Wishart [10] describes a FocalPoint *k*-means procedure that randomly rearranges the case order in each of a series of independent runs of the procedure. FocalPoint also generates a different random initial classification for each run. Randomisation

of the case order and the initial classification will generally yield different final classifications over the course of a number of trials. For example, in a study of 25 mammals described in terms of the composition of their milk, FocalPoint obtained 7 different stable classifications at the five-cluster level within 500 randomization trials, with values of E ranging from 2.58 to 3.45, as listed in Table 1. This is why it is important, firstly, to implement *k*-means analysis with the exact assignment test; and secondly, to pursue the optimisation procedure until it converges to a stable minimum E value that can be replicated. *k*-means programs that stop when a cluster centre does not move by more than a specified percentage of the smallest distance between any of the initial cluster centres will not necessarily converge to a stable, definitive E value. FocalPoint, by comparison, maintains a record of each classification obtained in a series of random trials, arranges them in ascending order of their E values, and counts the frequency with which each classification is replicated.

The results for 500 trials with the mammals milk data are given in Table 1, in which seven different stable classifications were found. The one corresponding to the smallest E value occurred in two-thirds of the trials and is almost certainly the global optimum solution for five clusters with these data. It is, however, instructive to examine the six suboptimal classifications, which find small clusters of one or two cases. Further examination reveals these to be outliers, so that the procedure has in effect removed the outliers and reduced the value of k resulting in larger clusters of 10 or 11 cases elsewhere. There is always the possibility that *k*-means analysis will find singleton clusters containing outliers, because such clusters contribute zero to the sum of squares E and therefore once separated they cannot be reassigned. There is, therefore, a case for creating a separate residue set of outliers.

In Clustan software, Wishart [10, 11], a residue set can be created by specifying a threshold distance d^* above which a case cannot be assigned to a cluster. Thus, if $d_{ip} > d^*$, the case is deemed too remote to be assigned to cluster p and is instead placed in the residue. Another possibility is to specify a percentage r , such as $r = 5\%$, and to place the 5% of the cases that are most remote from any cluster in the residue set. This requires additional work to

Table 1 Seven stable classifications found for 25 mammals by *k*-means analysis in 500 random trials, with cluster sizes, final E values and frequencies

Classification	Cluster sizes	E	Frequency
1	8,6,5,4,2	2.58	336
2	10,7,5,2,1	2.81	49
3	10,7,4,2,2	2.93	4
4	9,7,6,2,1	2.946	101
5	11,6,4,2,2	2.952	2
6	10,8,5,1,1	3.38	3
7	11,6,6,1,1	3.45	5

store the distances of the cases to their nearest cluster and at the completion of each iteration to order the distances, reallocate those cases corresponding to $r\%$ largest distances to the residue set, and recalculate the cluster means.

Comparison of *k*-means with Other Optimisation Methods

Some writers argue that *k*-means analysis produces spherical clusters that can be distorted by the presence of outliers, and advocate instead a search for elliptical clusters by fitting an underlying mixture of multivariate normal distributions (*see Finite Mixture Distributions; Model Based Cluster Analysis*). The choice of method really depends upon the object of the analysis – for example, whether one is seeking ‘natural’ classes or a more administrative classification. Suppose, for example, the cases are individuals described solely by their height and weight. A search for ‘natural’ clusters should only find one class, with the data fitting a single elliptical bivariate normal distribution (*see Catalogue of Probability Density Functions*). By comparison, *k*-means analysis will find any number of spherical

clusters throughout the distribution, a kind of mapping of the distribution. It will therefore separate, for example, ballerinas and sumo wrestlers, infants and adults, a summarization of the data rather than a ‘natural’ classification. The decision to use *k*-means analysis therefore depends upon the type of clusters that are sought – should they be such that any cluster of individuals are *all* similar in terms of all of the variables; or is variation allowed in some of the variables. Model-based clustering is suitable for the latter, whereas *k*-means analysis would suffice for the former.

Another goal for *k*-means analysis is the summarization of a large quantity of data, such as occur in data mining or remote sensing applications. The goal here is to reduce a large amount of data to manageable proportions by simplifying 200,000 cases in terms of 1000 clusters averaging 200 cases each. In this example, each cluster would comprise a set of individuals that are all very similar to each other, and a single exemplar can be selected to represent each cluster. The data are now described in terms of 1000 different cases that map the whole population in terms of all of the variables. Further modeling can more easily be undertaken on the 1000 exemplars than would be feasible on 200,000 individuals.

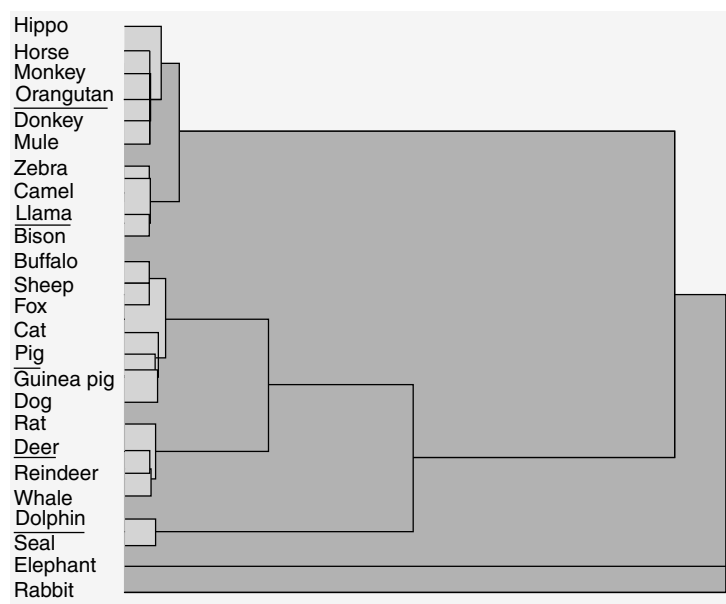


Figure 1 Dendrogram for *k*-means analysis at five clusters with two outliers (elephant and rabbit)

Dendrogram for *k*-means

Optimization methods such as *k*-means analysis generally yield a single partition of a set of n cases into k clusters. This says nothing about the relationships between the k clusters, or about the relationships between the cases assigned to any one cluster. A summary tree has therefore been implemented in ClustanGraphics [10] that shows these relationships, as illustrated in Figure 1.

A five-cluster solution for the 25 mammals data was obtained using ‘exact’ *k*-means analysis, as shown in Table 1. An outlier threshold was specified that caused two cases (elephant and rabbit) to be excluded from the classification as outliers. A subtree is now constructed for each cluster that shows how the members of that cluster would agglomerate to form it; and a supertree is constructed to show how the clusters would agglomerate hierarchically by Ward’s method. The resulting tree is shown in Figure 1, with the cluster exemplars underlined.

When using *k*-means analysis it is important also to use Ward’s method to construct the agglomerative tree for exact *k*-means analysis, because both methods optimize the Euclidean Sum of Squares. This is an excellent way of representing the structure around a *k*-means cluster model for a small data set. However, with larger data sets a full tree can become unwieldy, and therefore it is easier to restrict it to the final supertree agglomeration of k clusters.

References

- [1] Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press, London.
- [2] Ball, G.H. & Hall, D.J. (1967). A clustering technique for summarizing multivariate data, *Behavioural Science* **12**, 153–155.
- [3] Beale, E.M.L. (1969). Euclidean cluster analysis, *Bulletin of the International Statistical Institute* **43**(2), 92–94.
- [4] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th Edition, Edward Arnold, London.
- [5] Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classification, *Biometrics* **21**, 768–769.
- [6] Gordon, A.D. (1999). *Classification*, 2nd Edition, Chapman & Hall, London.
- [7] Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, London.
- [8] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings 5th Berkeley Symposium on Mathematics, Statistics and Probability* **I**, 281–297.
- [9] Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood, Chichester.
- [10] Wishart, D. (2004a). *FocalPoint User Guide*, Clustan, Edinburgh.
- [11] Wishart, D. (2004b). *ClustanGraphics Primer: A Guide to Cluster Analysis*, Clustan, Edinburgh.

DAVID WISHART

Kolmogorov, Andrey Nikolaevich

DAVID C. HOWELL

Volume 2, pp. 1022–1023

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kolmogorov, Andrey Nikolaevich

Born: April 25, 1903, in Tambov, Russia.

Died: October 20, 1987, in Moscow, Russia.

Andrey Nikolaevich Kolmogorov was born in Russia in 1903 and was raised by his aunt when his mother died in childbirth and his father was in exile. After a traditional schooling, Kolmogorov worked as a conductor on the railway. His educational background must have been excellent, however, because in his spare time he wrote a treatise on Newton's laws of mechanics, though he had not yet attended university. Kolmogorov enrolled in Moscow State University in 1920 and studied a number of subjects, including metallurgy, Russian history, and mathematics. He published his first paper two years later, and at the time of his graduation in 1925 he published eight more, including his first paper on probability.

By the time Kolmogorov completed his doctorate in 1929, he had 18 publications to his credit. Two years later, he was appointed a professor at Moscow University. The same year, he published a paper entitled *Analytical methods in probability theory*, in which he laid out the basis for the modern theory of Markov processes (see **Markov Chains**). Two years after that, in 1933, he published an axiomatic treatment of probability theory. Over the next few years, Kolmogorov and his collaborators published a long series of papers. Among these was a paper in which he developed what is now known as the **Kolmogorov-Smirnov test**, which is used to test the goodness of fit of a given set of data to a theoretical distribution. Though Kolmogorov's contributions to statistics are numerous and important, he was a mathematician, and not a statistician. Within mathematics, his range was astounding. Vitanyi [1] cited what he called

a 'nonexhaustive' list of the fields in which Kolmogorov worked:

'The theory of trigonometric series, measure theory, set theory, the theory of integration, constructive logic (intuitionism), topology, approximation theory, probability theory, the theory of random processes, information theory, mathematical statistics, dynamical systems, automata theory, theory of algorithms, mathematical linguistics, turbulence theory, celestial mechanics, differential equations, Hilbert's 13th problem, ballistics, and applications of mathematics to problems of biology, geology, and the crystallization of metals.

In over 300 research papers, textbooks and monographs, Kolmogorov covered almost every area of mathematics except number theory. In all of these areas even his short contributions did not just study an isolated question, but in contrast exposed fundamental insights and deep relations, and started whole new fields of investigations.'

Kolmogorov received numerous awards throughout his life and earned honorary degrees from many universities. He died in Moscow in 1987.

Sources of additional information

- *Andrey Nikolaevich Kolmogorov*: Available at <http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Kolmogorov.html>
- Vitanyi, P. M. B. A short biography of A. N. Kolmogorov. Available at <http://homepages.cwi.nl/~paulv/KOLMOGOROV.BIOGRAPHY.html>

Reference

- [1] Vitanyi, P.M.B. (2004). A short biography of A. N. Kolmogorov. Available at <http://homepages.cwi.nl/~paulv/KOLMOGOROV.BIOGRAPHY.html>

DAVID C. HOWELL

Kolmogorov–Smirnov Tests

VANCE W. BERGER AND YANYAN ZHOU

Volume 2, pp. 1023–1026

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kolmogorov–Smirnov Tests

The Kolmogorov–Smirnov and Smirnov Tests

The Kolmogorov–Smirnov test is used to test the goodness of fit of a given set of data to a theoretical distribution, while the Smirnov test is used to determine if two samples appear to follow the same distribution. Both tests compare cumulative distribution functions (*cdfs*). The problem of determining whether a given set of data appear to have been drawn from a known distribution is often of interest only because it has implications for the subsequent statistical analysis. For example, one may test a given distribution for normality, and if one fails to reject this hypothesis, then one may proceed as if normality were proven and use a parametric analysis that relies on normality for validity. This two-stage approach is problematic, because it confuses failure to reject a null hypothesis (in this case, normality) with proving its truth [10]. The consequence of this error is that the true Type I error rate may then exceed the nominal Type I error rate [1, 6], which is often taken to be 0.05.

The Smirnov Test

The problem of determining whether two sets are drawn from the same distribution functions arises naturally in many areas of research and, in contrast with the problem of fitting a set of data to a known theoretical distribution, is often of intrinsic interest. For example, one may ask if scores on a standardized test are the same in two different states or in two different counties. Or one may ask if within a family the incidence of chicken pox is the same for the first-born child and the second-born child. Formally, we can state the problem as follows:

Let $x:(x_1, x_2, \dots, x_m)$ and $y:(y_1, y_2, \dots, y_n)$ be independent random samples of size m and n , respectively, from continuous or ordered categorical populations with *cdfs* of F and G , respectively. We wish to test the null hypothesis of equality of distribution functions:

$$H_0: F(t) = G(t), \text{ for every } t. \quad (1)$$

This null hypothesis can be tested against the omnibus two-sided alternative hypothesis:

$$H_a: F(t) \neq G(t) \text{ for at least one value of } t, \quad (2)$$

or it can be tested against a one-sided alternative hypothesis:

$$H_a: F(t) \geq G(t) \text{ for all values of } t, \text{ strictly greater for at least one value of } t, \quad (3)$$

or

$$H_a: F(t) \leq G(t) \text{ for all values of } t, \text{ strictly smaller for at least one value of } t. \quad (4)$$

To compute the Smirnov test statistic, we first need to obtain the empirical *cdfs* for the x and y samples. These are defined by

$$F_m(t) = \frac{\text{number of sample } x\text{'s } \leq t}{m} \quad (5)$$

and

$$G_n(t) = \frac{\text{number of sample } y\text{'s } \leq t}{n}. \quad (6)$$

Note that these empirical *cdfs* take the value 0 for all values of t below the smallest value in the combined samples and the value 1 for all values of t at or above the largest value in the combined samples. Thus, it is the behavior between the smallest and largest sample values that distinguishes the empirical *cdfs*. For any value of t in this range, the difference between the two distributions can be measured by the signed differences, $[F_m(t) - G_n(t)]$ or $[G_n(t) - F_m(t)]$, or by the absolute value of the difference, $|F_m(t) - G_n(t)|$. The absolute difference would be the appropriate choice where the alternative hypothesis is nondirectional, $F(t) \neq G(t)$, while the choice between the two signed differences depends on the direction of the alternative hypothesis. The value of the difference or absolute difference may change with the value of t . As a result, there would be a potentially huge loss of information in comparing distributions at only one value of t [3, 14, 17].

The Smirnov test resolves this problem by employing as its test statistic, D , the maximum value of the selected difference between the two empirical cumulative distribution functions:

$$D = \max |F_m(t) - G_n(t)|, \quad \min(x, y) \leq t \leq \max(x, y), \quad (7)$$

2 Kolmogorov–Smirnov Tests

where the alternative hypothesis is that $F(t) \neq G(t)$,

$$D^+ = \max[F_m(t) - G_n(t)],$$

$$\min(x, y) \leq t \leq \max(x, y), \quad (8)$$

where the alternative hypothesis is that $F(t) > G(t)$ for some value(s) of t , and

$$D^- = \max[G_n(t) - F_m(t)],$$

$$\min(x, y) \leq t \leq \max(x, y), \quad (9)$$

where the alternative hypothesis is that $F(t) < G(t)$ for some value(s) of t . Note, again, that the difference in empirical *cdf*s need be evaluated only for the set of unique values in the samples x and y .

To test H_0 at the α level of significance, one would reject H_0 if the test statistic is large, specifically, if it is equal to or larger than D_α . The Smirnov test has been discussed by several authors, among them Hilton, Mehta, and Patel [11], Nikiforov [15], and Berger, Permutt, and Ivanova [7]. One key point is that the exact permutation reference distribution (see **Exact Methods for Categorical Data**) should be used instead of an approximation, because the approximation is often quite poor, [1, 6]. In fact, in analyzing a real set of data, Berger [1] found enormous discrepancies between the exact and approximate Smirnov tests for both the one-sided and the two-sided P values. Because of the discreteness of the distribution of sample data, it generally will not be possible to choose D_α to create a test whose significance level is exactly equal to a prespecified value of α . Thus, D_α should be chosen from the permutation reference distribution so as to make the Type I error probability no greater than α .

For any value of t , the two-sample data can be fit to a 2×2 table of frequencies. The rows of the table correspond to the two populations sampled, X and Y , while the columns correspond to response magnitudes – responses that are no larger than t in one column and responses that exceed t in a second column. Taken together, these tables are referred to as the *Lancaster decomposition* [3, 16]. Such a set of tables is unwieldy where the response is continuously measured, but may be of interest where the response is one of a small number of ordered categories, for example, unimproved, slightly improved, and markedly improved. In this case, the Smirnov D can be viewed as the maximized Fisher exact

test statistic – the largest difference in success proportions across the two groups, as the definition of success varies over the set of Lancaster decomposition tables [16]. Is D maximized when success is defined as either slightly or markedly improved, or when success is defined as markedly improved?

The row margins, sample sizes, are fixed by the design of the study. The column margins are complete and sufficient statistics under the equality null hypothesis. Hence, the exact analysis would condition on the two sets of margins [8]. The effect of conditioning on the margins is that the permutation sample space is reduced compared to what it would be with an unconditional approach. Berger and Ivanova [4] provide S-PLUS code for the exact conditional computations; see also [13, 18, 19]. The Smirnov test is invariant under reparameterization of x . That is, the maximum difference D will not change if x undergoes a monotonic transformation. So the same test statistic D results if x is analyzed or if $\log(x)$ or some other transformation, such as the square of x , is analyzed; hence, the P value remains the same too.

The Smirnov test tends to be most sensitive around the median value where the cumulative probability equals 0.5. As a result, the Smirnov test is good at detecting a location shift, especially changes in the median values, but it is not always as good at detecting a scale shift (spreads), which affects the tails of the probability distribution more, and which may leave the median unchanged. The **power** of the Smirnov test to detect specific alternatives can be improved in any of several ways. For example, the power can be improved by refining the Smirnov test so that any ties are broken. Ties result from different tables in the permutation reference distribution having the same value of the test statistic.

There are several approaches to breaking the ties, so that at least some of the tables that had been tied are now assigned distinct values of the test statistic [12, 16]. Only one of these approaches represents a true improvement in the sense that even the exact randomized version of the test becomes uniformly more powerful [8]. Another modification of the Smirnov test is based on recognizing its test statistic to be a maximized difference of a constrained set of linear rank test statistics, and then removing the constraint, so that a wider class of linear rank test statistics is considered in the maximization [5].

The power of this adaptive omnibus test is excellent.

While there do exist tests that are generally or in some cases even uniformly more powerful than the Smirnov test, the Smirnov test still retains its appeal as probably the best among the simple tests that are available in standard software packages (the exact Smirnov test is easily conducted in StatXact). One could manage the conservatism of the Smirnov test by reporting not just its P value but also the entire P value interval [2], whose upper end point is the usual P value but whose lower end point is what the P value would have been without any conservatism.

The Kolmogorov–Smirnov Test

As noted above, the Kolmogorov–Smirnov test assesses whether a single sample could have been sampled from a specified probability distribution. Letting $G_n(t)$ be the empirical cdf for the single sample and $F(t)$ the theoretical cdf – for example, Normal with mean μ and variance σ^2 – the Kolmogorov–Smirnov test statistic takes one of these forms:

$$D_k = \max |F(t) - G_n(t)|, \\ \min(x) \leq t \leq \max(x), \quad (10)$$

where the alternative hypothesis is that $F(t) \neq G(t)$,

$$D_k^+ = \max[F(t) - G_n(t)], \\ \min(x) \leq t \leq \max(y), \quad (11)$$

where the alternative hypothesis is that $F(t) > G(t)$ for some value(s) of t , and

$$D_k^- = \max[G_n(t) - F(t)], \\ \min(x) \leq t \leq \max(x), \quad (12)$$

where the alternative hypothesis is that $F(t) < G(t)$ for some value(s) of t .

The Kolmogorov–Smirnov test has an exact null distribution for the two directional alternatives but the distribution must be approximated for the nondirectional case [9]. Regardless of the alternative, the test is less accurate if the parameters of the theoretical distribution have been estimated from the sample.

References

- [1] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [2] Berger, V.W. (2001). The p -value interval as an inferential tool, *Journal of the Royal Statistical Society. Series D, (The Statistician)* **50**, 79–85.
- [3] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [4] Berger, V.W. & Ivanova, A. (2001). Permutation tests for phase III clinical trials. Chapter 14, in *Applied Statistics in the Pharmaceutical Industry with Case Studies Using S-PLUS*, S.P. Millard & A. Krause, eds, Springer Verlag, New York.
- [5] Berger, V.W. & Ivanova, A. (2002). Adaptive tests for ordered categorical data, *Journal of Modern Applied Statistical Methods* **1**, 269–280.
- [6] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**, 74–82.
- [7] Berger, V.W., Permutt, T. & Ivanova, A. (1998). The convex hull test for ordered categorical data, *Biometrics* **54**, 1541–1550.
- [8] Berger, V. & Sackrowitz, H. (1997). Improving tests for superior treatment in contingency tables, *Journal of the American Statistical Association* **92**, 700–705.
- [9] Conover, W.J. (1999). *Practical Nonparametric Statistics*, Wiley, New York.
- [10] Greene, W.L., Concato, J. & Feinstein, A.R. (2000). Claims of equivalence in medical research: are they supported by the evidence? *Annals of Internal Medicine* **132**, 715–722.
- [11] Hilton, J.F., Mehta, C.R. & Patel, N.R. (1994). An algorithm for conducting exact Smirnov tests, *Computational Statistics and Data Analysis* **17**, 351–361.
- [12] Ivanova, A. & Berger, V.W. (2001). Drawbacks to integer scoring for ordered categorical data, *Biometrics* **57**, 567–570.
- [13] Kim, P.J. & Jennrich, R.I. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov–Smirnov criterion, in *Selected Tables in Mathematical Statistics*, Vol. I, H.L. Harter & D.B. Owen, eds, American Mathematical Society, Providence.
- [14] Moses, L.E., Emerson, J.D. & Hosseini, H. (1984). Analyzing data from ordered categories, *New England Journal of Medicine* **311**, 442–448.
- [15] Nikiforov, A.M. (1994). Exact Smirnov two-sample tests for arbitrary distributions, *Applied Statistics* **43**, 265–284.
- [16] Permutt, T. & Berger, V.W. (2000). A new look at rank tests in ordered $2 \times k$ contingency tables, *Communications in Statistics – Theory and Methods* **29**, 989–1003.
- [17] Rahlfs, V.W. & Zimmermann, H. (1993). Scores: ordinal data with few categories – how should they be analyzed? *Drug Information Journal* **27**, 1227–1240.

4 Kolmogorov–Smirnov Tests

- [18] Schroer, G. & Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov–Smirnov tests for two or three samples, *Computational Statistics and Data Analysis* **20**, 185–202.
- [19] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego.

VANCE W. BERGER AND YANYAN ZHOU

Kruskal–Wallis Test

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Volume 2, pp. 1026–1028

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kruskal–Wallis Test

The nonparametric Kruskal-Wallis test [5] and [6] is an extension of the **Wilcoxon-Mann-Whitney test**. The null hypothesis is that the k populations sampled have the same average (median). The alternative hypothesis is that at least one sample is from a distribution with a different average (median). This test is an alternative to the parametric one-way **analysis of variance F test**.

Assumptions

The samples are assumed to be independent of each other, the populations are continuously distributed, and there are no tied values. Monte Carlo results by Fay and Sawilowsky [3] and Fay [4] indicated that the best techniques for resolving tied values are (a) assignment of midranks or (b) random assignment of tied values for or against the null hypothesis.

Procedure

Rank all the observations in the combined samples, keeping track of the sample membership. Compute the rank sums of each sample. Let R_i equal the sum of the ranks of the i th sample of sample size n_i . The logic of the test is that the ranks should be randomly distributed among the k samples.

Test Statistic

The computational formula for the test statistic, H , is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (1)$$

where N is the total sample size, n_i is the size of the i th group, k is the number of groups, and R_i is the rank-sum of the i th group. Reject H_0 when $H \geq$ critical value.

Large Sample Sizes

For large sample sizes, the null distribution is approximated by the χ^2 distribution with $k - 1$ degrees

of freedom. Thus, the rejection rule is to reject H_0 if $H \geq \chi_{\alpha, k-1}^2$, where $\chi_{\alpha, k-1}^2$ is the value of χ^2 at nominal α with $k - 1$ degrees of freedom. Monte Carlo simulations conducted by Fahoome and Sawilowsky [2] and Fahoome [1] indicated that the large sample approximation requires a minimum sample size of 7 when $k = 3$, and 14 when $k = 6$, for $\alpha = 0.05$.

Example

The Kruskal-Wallis statistic is calculated using the following five samples, $n_1 = n_2 = n_3 = n_4 = n_5 = 15$ (Table 1).

Table 1 Data set for five samples

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
4	5	5	4	2
6	8	7	10	4
9	9	7	10	9
13	11	8	13	10
13	11	8	13	11
16	11	9	13	12
17	18	11	15	15
20	21	14	20	15
26	23	14	21	19
29	24	18	26	20
33	27	20	34	21
33	32	20	35	31
34	33	22	38	32
36	34	33	41	33
39	37	33	43	33

Table 2 Ranked data set for five samples

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
3	5.5	5.5	3	1
7	11	8.5	18	3
14.5	14.5	8.5	18	14.5
28	22	11	28	18
28	22	11	28	22
36	22	14.5	28	25
37	38.5	22	34	34
43	47	31.5	43	34
52.5	50	31.5	47	40
55	51	38.5	52.5	43
62	54	43	67	47
62	57.5	43	69	56
67	62	49	72	57.5
70	67	62	74	62
73	71	62	75	62
638	595	441.5	656.5	519

2 Kruskal–Wallis Test

Table 3 Ranked sums for five samples

	R_i	$(R_i)^2$
R_1	638.0	407,044.00
R_2	595.0	354,025.00
R_3	441.5	194,922.25
R_4	656.5	430,992.25
R_5	519.0	269,361.00
		1,656,344.50

The combined samples were ranked, and tied ranks were assigned average ranks (Table 2).

The rank sums are: $R_1 = 638$, $R_2 = 595$, $R_3 = 441.5$, $R_4 = 656.5$, and $R_5 = 519$. The sum of $R_i^2 = 1,656,344.50$ (Table 3).

$H = 12/75 \cdot 76 (1,656,344.5/15) - 3 \cdot 76 = 0.00211(110, 422.9667) - 228 = 4.4694$. The H statistic is 4.4694. Because H is less than the critical value for H at $\alpha = 0.05$, $n = 15$, and $k = 5$, the null hypothesis cannot be rejected. The large sample approximation is 9.488, which is Chi-square with $5 - 1 = 4$ degrees of freedom at $\alpha = 0.05$. Because $4.4694 < 9.488$, the null hypothesis cannot be rejected on the basis of the evidence from these samples.

References

- [1] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.
- [2] Fahoome, G. & Sawilowsky, S. (2000). *Twenty Nonparametric Statistics. Annual meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.
- [3] Fay, B.R. (2003). *A Monte Carlo computer study of the power properties of six distribution-free and/or nonparametric statistical tests under various methods of resolving tied ranks when applied to normal and nonnormal data distributions*, Unpublished doctoral dissertation, Wayne State University, Detroit.
- [4] Fay, B.R. & Sawilowsky, S. (2004). The Effect on Type I Error and Power of Various Methods of Resolving Ties for Six Distribution-Free Tests of Location. Manuscript under review.
- [5] Kruskal, W.H. (1957). Historical notes on the Wilcoxon unpaired two-sample test, *Journal of American Statistical Association* **52**, 356–360.
- [6] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks on one-criterion analysis of variance, *Journal of American Statistical Association* **47**, 583–621.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Kurtosis

KARL L. WUENSCH

Volume 2, pp. 1028–1029

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Kurtosis

Karl Pearson [6] defined a distribution's degree of kurtosis as $\eta = \beta_2 - 3$, where $\beta_2 = \sum (X - \mu)^4 / n\sigma^4$, the expected value of the distribution of Z scores which have been raised to the 4th power. β_2 is often referred to as 'Pearson's kurtosis', and $\beta_2 - 3$ (often symbolized with γ_2) as 'kurtosis excess' or 'Fisher's kurtosis', even though it was Pearson who defined kurtosis as $\beta_2 - 3$. An unbiased estimator for γ_2 is $g_2 = n(n+1) \sum Z^4 / (n-1)(n-2)(n-3) - 3(n-1)^2 / (n-2)(n-3)$. For large sample sizes ($n > 1000$), g_2 may be distributed approximately normally, with a standard error of approximately $\sqrt{(24/n)}$ [7]. While one could use this sampling distribution to construct confidence intervals for or tests of hypotheses about γ_2 , there is rarely any value in doing so.

Pearson [6] introduced kurtosis as a measure of how flat the top of a symmetric distribution is when compared to a normal distribution of the same variance. He referred to distributions that are more flat-topped than normal distributions ($\gamma_2 < 0$) as 'platykurtic', those less flat-topped than normal ($\gamma_2 > 0$) as 'leptokurtic', and those whose tops are about as flat as the tops of normal distributions as 'mesokurtic' ($\gamma_2 \approx 0$).

Kurtosis is actually more influenced by scores in the tails of the distribution than scores in the center of a distribution [3]. Accordingly, it is often appropriate to describe a leptokurtic distribution as 'fat in the tails' and a platykurtic distribution as 'thin in the tails.'

Moors [5] demonstrated that $\beta_2 = \text{Var}(Z^2) + 1$. Accordingly, it may be best to treat kurtosis as the extent to which scores are dispersed away from the shoulders of a distribution, where the shoulders are the points where $Z^2 = 1$, that is, $Z = \pm 1$. If one starts with a normal distribution and moves scores from the shoulders into the center and the tails, keeping variance constant, kurtosis is increased. The distribution will probably appear more peaked in the center and fatter in the tails, like a Laplace distribution ($\gamma_2 = 3$) or a Student's t with few degrees of freedom ($\gamma_2 = 6/(df - 4)$). Starting again with a normal distribution, moving scores to the shoulders from the center and the tails will decrease kurtosis. A uniform distribution certainly has a flat top

with $\gamma_2 = 1.2$, but γ_2 can reach a minimum value of -2 when two score values are equally probable and all other score values have probability zero (a rectangular U distribution, that is, a binomial distribution with $n = 1, p = .5$). One might object that the rectangular U distribution has all of its scores in the tails, but closer inspection will reveal that it has no tails, and that all of its scores are in its shoulders, exactly one standard deviation from its mean. Values of g_2 less than that expected for a uniform distribution (-1.2) may suggest that the distribution is bimodal [2], but bimodal distributions can have high kurtosis if the modes are distant from the shoulders.

Kurtosis is usually of interest only when dealing with approximately symmetric distributions. **Skewed** distributions are always leptokurtic [4]. Among the several alternative measures of kurtosis that have been proposed (none of which has often been employed) is one which adjusts the measurement of kurtosis to remove the effect of skewness [1].

While it is unlikely that a behavioral researcher will be interested in questions that focus on the kurtosis of a distribution, estimates of kurtosis, in combination with other information about the shape of a distribution, can be useful. DeCarlo [3] described several uses for the g_2 statistic. When considering the shape of a distribution of scores, it is useful to have at hand measures of skewness and kurtosis, as well as graphical displays. These statistics can help one decide which estimators or tests should perform best with data distributed like those on hand. High kurtosis should alert the researcher to investigate outliers in one or both tails of the distribution.

References

- [1] Blest, D.C. (2003). A new measure of kurtosis adjusted for skewness, *Australian & New Zealand Journal of Statistics* **45**, 175-179.
- [2] Darlington, R.B. (1970). Is kurtosis really "peakedness?", *The American Statistician* **24**(2), 19-22.
- [3] DeCarlo, L.T. (1997). On the meaning and use of kurtosis, *Psychological Methods* **2**, 292-307.
- [4] Hopkins, K.D. & Weeks, D.L. (1990). Tests for normality and measures of skewness and kurtosis: their place in research reporting, *Educational and Psychological Measurement* **50**, 717-729.

2 Kurtosis

- [5] Moors, J.J.A. (1986). The meaning of kurtosis: Darlington reexamined, *The American Statistician* **40**, 283–284.
- [6] Pearson, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A rejoinder, *Biometrika* **4**, 169–212.
- [7] Snedecor, G.W. & Cochran, W.G. (1967). *Statistical Methods*, 6th Edition, Iowa State University Press, Ames.

KARL L. WUENSCH

Laplace, Pierre Simon (Marquis de)

DIANA FABER

Volume 2, pp. 1031–1032

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Laplace, Pierre Simon (Marquis de)

Born: March 23, 1749, in Normandy, France.

Died: March 5, 1827, in Paris, France.

From the early years of the 19th century, Laplace became recognized as a mathematical genius, astronomer, and pioneer in the growing discipline of statistics. Napoleon had a genuine interest in science and mathematics, which helped to increase their standing in France. Such support from Napoleon no doubt enhanced the reputation of Laplace in particular. Given these circumstances, **Adolphe Quételet** travelled from Belgium to Paris in 1823. He learned observational astronomy and, from Laplace, received some informal instruction on mathematical probability and its application.

At the age of 17, Laplace attended the University of Caen where he distinguished himself as a mathematician. In 1767, he went to Paris, and here he was appointed professor at the Ecole Militaire. He became a member of the Bureau des Longitudes from its inception in 1795 and a member of the Observatory, which dated from the 17th century. Both were prominent research institutions for mathematical and astronomical studies. Laplace took a leading part in the Bureau, where he encouraged others to investigate his own research problems. He also became an examiner at the Ecole Normale, another prestigious intellectual French institution in Paris. Laplace became a member of the Académie Royale des Sciences, which from 1795 was named the *l'Institut National* [1]. Such was the prestige of the sciences that they were categorized as 'First Class', and in 1816 the Académie des Sciences were granted the power to recommend and appoint further members. Laplace became one of its key figures and, moreover, enjoyed the patronage of Napoleon, which benefited not only Laplace himself but also the standing of science in general. Thus, Laplace was enabled to exert his own influence. For example, he persuaded Baron

Montyon that statistics deserved special attention. When the Académie decided to award a statistics prize, this relatively new subject became recognized as an additional independent branch of science. Furthermore, as Joseph Fourier (1768–1830) pointed out, statistics was not to be confused with political economy.

One of the first works of Laplace was 'Mémoire sur la Probabilité des Causes par les Evénements' (1774); in this work, he announced his 'principe' as that of inverse probability, giving recognition to the earlier work of **Jean Bernoulli**, which he had not seen. This principle provided a solution to the problem of deciding upon a mean on the basis of several observations of the same phenomenon. In 1780, he presented to the Académie 'Mémoire sur la Probabilité', published in 1781. According to Stigler, these publications 'are among the most important and most difficult works in the history of mathematical probability' ([3, p. 100], see also [2]). In 1809, he then read 'Sur les Approximations des Formules qui sont Fonctions de très-grands Nombres', published the following year. The 'Supplément' to this memoir contained his statement of a **central limit theorem**. Among other works for which Laplace is best known is his 'Exposition du Système du Monde' of 1796, a semipopular work. He dedicated his work on astronomical mathematics, 'Traité de Mécanique Céleste' to Napoleon, which was published in four volumes between 1799 and 1805. In 1812, he presented 'Théorie Analytique des Probabilités' to the First Class of the Académie. This was followed by 'Essai Philosophique' in 1814.

References

- [1] Crosland, M. (1992). *Science Under Control: The French Academy of Sciences (1795–1914)*, Cambridge University Press, Cambridge.
- [2] Porter, T.M. (1986). *The Rise of Statistical Thinking 1820–1900*, Princeton University Press, Princeton.
- [3] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, The Belknap Press of Harvard University Press, Harvard.

DIANA FABER

Latent Class Analysis

BRIAN S. EVERITT

Volume 2, pp. 1032–1033

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Latent Class Analysis

Latent class analysis provides a model for data involving categorical variables that assumes that the observed relationships between the variables arise from the presence of an underlying categorical or ‘grouping’ variable for the observations, with, within each group, the variables being independent of one another, the so-called *conditional independence* assumption. Overviews of latent class analysis are given in [1, 4], but essentially the model involves postulating a **finite mixture distribution** for the data, in which the component densities are multivariate Bernoulli (see **Catalogue of Probability Density Functions**). For binary observed variables, the latent class model assumes that the underlying latent variable has c categories and that in the i th category, the probability of a ‘1’ response on variable j is given by

$$\Pr(x_{ij} = 1 | \text{group } i) = \theta_{ij}, \quad (1)$$

where x_{ij} is the value taken by the j th variable in group i . From the conditional independence assumption, it follows that the probability of an observed vector of responses, $\mathbf{x}' = [x_{i1}, x_{i2}, \dots, x_{iq}]$, where q is the number of observed variables is given by

$$\Pr(\mathbf{x} | \text{group } i) = \prod_{j=1}^q \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{1-x_{ij}}. \quad (2)$$

The proportions of each category in the population are assumed to be $p_1, p_2, \dots, p_c$, with $\sum_{i=1}^c p_i = 1$, leading to the following unconditional density function for the observed vector of responses $\mathbf{x}$

$$\Pr(\mathbf{x}) = \sum_{i=1}^c p_i \prod_{j=1}^q \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{1-x_{ij}}. \quad (3)$$

The parameters of this density function, the p_i and the θ_{ij} are estimated by **maximum likelihood**, and the division of observations into the c categories made by using the maximum values of the estimated posterior probabilities (see **Finite Mixture Distributions**). For examples of the application of latent class analysis see, for example, [2, 3].

References

- [1] Clogg, C. (1992). The impact of sociological methodology on statistical methodology (with discussion), *Statistical Science* **7**, 183–196.
- [2] Coway, M.R. (1998). Latent class analysis, in *Encyclopedia of Biostatistics*, Vol. 3, P. Armitage & T. Colton, eds, Wiley, Chichester.
- [3] Everitt, B.S. & Dunn, G. (2001). *Applied Multivariate Data Analysis*, Arnold, London.
- [4] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton-Mifflin, Boston.

(See also **Cluster Analysis: Overview**)

BRIAN S. EVERITT

Latent Transition Models

FRANK VAN DE POL AND ROLF LANGEHEINE

Volume 2, pp. 1033–1036

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Latent Transition Models

During the twentieth century, a considerable number of **panel studies** have been carried out, making short **time series** of three or more measurements available for large numbers of subjects. This raised the question how to describe stability and change in the categorical variables typical in these surveys, relating to socioeconomic, medical, marketing, and other issues. The description of flows between the states of these variables for discrete periods of time turned out to be an intuitively appealing approach.

The foundation for these models was laid by the Russian mathematician **Andrei A. Markov (1856–1922)** who formulated a simple model for a sequence of discrete distributions in time. He supposed that history is irrelevant except for the previous state, and showed that a system with predictable behavior emerges. As an example, we will take employment status, with three states: employed, E, unemployed, U, and out of the labor force, O. Suppose this variable is measured at several successive quarters in a random sample of size 10 000. Ignoring births and deaths for simplicity, turnover as in Table 1 could be observed.

The assumption of the **Markov chain** is that the process has no memory. Only the initial state proportions matter for the prediction of the distribution at the next occasion. Prediction is possible by using the transition probabilities, the probabilities of being in each of the possible states at the next time point, given the state currently occupied. After multiplication of the initial state proportions with the corresponding transition probabilities and summation of resulting products that relate to the same employment status at the next time point, the distribution of the next time point is obtained; that is, $0.4 \times 0.97 + 0.1 \times 0.05 + 0.5 \times 0.01$ for the number of employed at the next occasion, a slight decrease to 39.8% compared to the initial 40%. Once the distribution for the next occasion is obtained, the distribution for one period ahead in time can be predicted, provided that the Markov chain is stationary, that is, transition probabilities are still the same.

When the Markov assumption holds for a stationary chain, the initial distribution will gradually converge to equilibrium as more and more periods pass. This means that, in the long run, the probability of being in a specific state becomes independent of the initial state, that is, the matrix of probabilities that describe the transition from initial states to states many periods later forms a matrix with equal rows. It can be computed by matrix multiplication of the matrix with one-period transition probabilities, **T**, with **T** itself and the result again with **T**, and so on until all rows of the resulting matrix are the same. For the example data, it will take several dozens of years to reach the equilibrium state, 33% employed, 8% unemployed, and 59% out of the labor force.

Social science data generally exhibit less change in the long run than a Markov chain predicts. With respect to labor market transitions, the Markov property is not realistic because of heterogeneity within states with respect to the probability of changing state. People out of the labor force are not homogeneous; retired people will generally have a low probability of reentering the labor force, and students have a much higher probability of entering it. People who are employed for several periods have a higher probability of remaining employed than those who have recently changed jobs, and so on. Blumen et al. [1] introduced a model with a distinction between people who change state according to a Markov chain, the movers, and people who remain in the same state during the observation periods, the stayers.

The example in Table 2 shows a population that can be split up in 25% movers and 75% stayers. However, the distinction between movers and stayers is not observed. It is a **latent variable**, whose distribution can be estimated with panel data on three or more occasions and an appropriate model [4]. As the proportion of stayers is a model parameter, this mover–stayer model can accommodate data with little change, especially little long-term change.

Another more realistic model is the latent Markov model. As the mover–stayer model, it predicts less long-term change than the simple Markov chain. It is due to Wiggins [13], based on his 1955 thesis, and Lazarsfeld and Henry [6]. According to the latent Markov model, observed change is partly due to **measurement error**, also called *response uncertainty* or *unreliability*. The model separates measurement error from latent change, also called *indirectly measured change*. In Table 3, a cross-table is conceived that

2 Latent Transition Models

Table 1 A cross-table of successive measurements viewed as one transition of a Markov chain

Initial state	Initial state proportions	State next quarter					
		Frequencies			Transition probabilities		
		E	U	O	E	U	O
Employed, E	0.40	3880	40	80	0.97	0.01	0.02
Unemployed, U	0.10	50	880	70	0.05	0.88	0.07
Out of labor force, O	0.50	50	50	4900	0.01	0.01	0.98

Table 2 Separation of movers and stayers, a latent, indirectly measured, distinction of types

		State next quarter					
		Frequencies			Transition probabilities		
Initial state	Initial state proportions	E	U	O	E	U	O
Movers, 25%							
Employed, E	0.10	380	40	80	0.76	0.08	0.16
Unemployed, U	0.05	50	380	70	0.10	0.76	0.14
Out of labor force, o	0.10	50	50	400	0.10	0.10	0.80
Stayers, 75%							
Employed, E	0.30	3500	0	0	1	0	0
Unemployed, U	0.05	0	500	0	0	1	0
Out of labor force, o	0.40	0	0	4500	0	0	1

Table 3 Measurement error and latent, indirectly measured, transitions both explain observed change

Latent initial state	Latent initial state proportions	Observed (same quarter)			Latent state next quarter		
		Response probabilities			Transition probabilities		
		E	U	O	E	U	O
Proportions							
Employed, E	0.40	0.992	0.004	0.004	0.97	0.01	0.02
Unemployed, U	0.10	0.04	0.92	0.04	0.11	0.76	0.13
Out of labor force, o	0.50	0.003	0.005	0.992	0.01	0.01	0.98
Frequencies							
Employed, E	4000	3972	15	14	3899	29	72
Unemployed, U	500	18	312	20	55	382	63
Out of labor force, o	5000	17	23	4960	59	41	4900

relates the indirectly measured, latent, initial labor market status with the observed labor market status. This cross-table holds the probabilities of a specific response, given the indirectly measured, latent status. For instance, the probability that unemployed will erroneously be scored as out of the labor force is 20/500 or 0.04.

The latent Markov model has a local independence assumption. Response probabilities only depend on

the present latent state, not on previous states or on other variables. Moreover, a Markovian process is assumed for the latent variables. With these assumptions, the latent Markov model can be estimated from panel data on three or more occasions, that is, from many observations on few time points. MacDonald and Zucchini [8] show how a similar ‘hidden Markov’ model can be estimated from only one time series with many time points.

Some data have a strong measurement error component and other data clearly have heterogeneity in the inclination to change. If both are the case, it makes sense to set up a partially latent mover–stayer model, that is, to allow a certain proportion of perfectly measured stayers in the model, in addition to movers whose status was indirectly measured [4]. It is possible to estimate this model with data from four or more occasions. In fact, the example data for the latent Markov model of Table 3 were generated from a population that is a mixture of 75% stayers with 25% movers, only the latter responding with some measurement error. This means that model choice was not correct in the above example.

With real life data, model choice is not only a matter of making assumptions but also of testing whether the model fits the data. Computer programs like PanMark [10], WinLTA [14] and LEM [7] offer **maximum likelihood** (ML) estimation, that is, such parameters that make the data most likely, given the model. Then fit measures may be computed, most of which summarize the difference between fitted frequencies and observed frequencies. A plausible model may be selected on the basis of fit measures, but this is no guarantee for choosing the correct model. For valid model selection and for accurate parameter estimates, it is good to have rich data, for instance, with more than one indicator of the same latent concept at each occasion.

The models described so far are latent class models with specific Markovian restrictions. They offer a model-based extension to descriptive statistics. In some disciplines they are used for testing hypotheses, for instance, that learning goes through stages and that no one goes back to a lower stage. Collins and Wugalter [2] showed how a model for this theory should be built by setting some transition probabilities to zero, taking measurement error into account.

More restrictions can be tested when modeling joint and marginal distributions in one analysis. For instance, the hypothesis of no change in the marginal distribution, marginal homogeneity, can be tested in combination with a Markov chain [12]. In addition to marginal homogeneity, cross-tables of subsequent measurements can be tested for symmetry, or for quasi-symmetry if marginal homogeneity does not apply.

Once a model has been selected as an appropriate representation of reality, one may want to explain

its structure by relating latent classes to exogenous (predictor) variables. This can be done by considering the (posterior) distribution of latent classes for each response pattern. In the labor market example, for instance, one could have a latent distribution at the first occasion of 99.9% E, 0.05% U and 0.05% O for response pattern EEE, and so on, up to response pattern OOO. One can add up these distinct latent distributions to obtain the overall distribution of latent classes or, more usefully, to cross-tabulate some of the latent classes with exogenous variables [11]. After cross-tabulation, a logit analysis (*see Logistic Regression*) could follow. In this approach, parameter estimation of the latent Markov model precedes, and is independent of, the analysis with exogenous variables. Alternatively, full information ML approaches enable simultaneous estimation of all relevant parameters [3, 9] (*see Quasi-symmetry in Contingency Tables*).

With a cross-tabulation of increasingly many categorical variables, the number of cells eventually will be larger than the number of observations. Thus, sample tables are sparse, that is, many cells will be empty and therefore fit measures that follow a chi-square distribution for infinitely large samples are not applicable. In the context of latent Markov models, a **bootstrap** approach has been developed to find an appropriate distribution for these fit measures, which is implemented in PanMark and Latent Gold [5]. WinLTA uses a Gibbs sampling-based method to address the problem of sparse tables (*see Markov Chain Monte Carlo and Bayesian Statistics*).

Finally, it should be noted that the discrete time approach in this contribution is restrictive in the sense that information on events between panel measurement occasions is ignored. Usually, panel data do not have this sort of information. However, when information on the exact timing of events is somehow available, transition probabilities in discrete time can be exchanged for more basic flow parameters in continuous time, called *transition rates* or *hazard rates*, which apply in an infinitesimally small period of time. Relating these rates to exogenous variables is known as *event history analysis*, **survival analysis** or *duration analysis*.

References

- [1] Blumen, I., Kogan, M. & McCarthy, P.J. (1966). Probability models for mobility. in *Readings in Mathematical*

- Social Science*, P.F. Lazarsfeld & N.W. Henry, eds, MIT Press, Cambridge, pp. 318–334; Reprint from: idem (1955). *The Industrial Mobility of Labor as a Probability Process*, Cornell University Press, Ithaca.
- [2] Collins, L.M. & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables, *Multivariate Behavioral Research* **27**, 131–157.
 - [3] Humphreys, K. & Janson, H. (2000). Latent transition analysis with covariates, nonresponse, summary statistics and diagnostics: modelling children's drawing development, *Multivariate Behavioral Research* **35**, 89–118.
 - [4] Langeheine, R. & van de Pol, F.J.R. (1990). A unifying framework for Markov modeling in discrete space and discrete time, *Sociological Methods & Research* **18**, 416–441.
 - [5] Latent Gold. (2003). <http://www.statisticalinnovations.com/products/latentgold.html>
 - [6] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton Mifflin, Boston.
 - [7] Lem. (1998). <http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html>
 - [8] MacDonald, I.L. & Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall, London.
 - [9] Muthén, B. (2002). Beyond SEM: general latent variable modeling, *Behaviormetrika* **29**, 81–117.
 - [10] PanMark. (2003). <http://www.assess.com/Software/Panmark.htm>
 - [11] Van de Pol, F.J.R. & Mannan, H. (2002). Questions of a novice in latent class modelling, *Methods of Psychological Research Online* **17**, <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue17/>.
 - [12] Vermunt, J.K., Rodrigo, M.F. & Ato-Garcia, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data, *Sociological Methods & Research* **30**, 170–196.
 - [13] Wiggins, L.M. (1973). *Panel Analysis: Latent Probability Models for Attitudes and Behavior Processes*, Elsevier, Amsterdam.
 - [14] WinLTA. (2002). <http://methodology.psu.edu/downloads/winlta.html>.

FRANK VAN DE POL AND ROLF LANGEHEINE

Latent Variable

JEROEN K. VERMUNT AND JAY MAGIDSON

Volume 2, pp. 1036–1037

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Latent Variable

Many constructs that are of interest to social scientists cannot be observed directly. Examples are preferences, attitudes, behavioral intentions, and personality traits. Such constructs can only be measured indirectly by means of observable indicators, such as questionnaire items designed to elicit responses related to an attitude or preference. Various types of scaling techniques have been developed for deriving information on unobservable constructs of interest from the indicators. An important family of scaling methods is formed by latent variable models.

A latent variable model is a, possibly nonlinear, path analysis or graphical model (see **Path Analysis and Path Diagrams; Graphical Chain Models**). In addition to the manifest variables, the model includes one or more unobserved or latent variables representing the constructs of interest. Two assumptions define the causal mechanisms underlying the responses. First, it is assumed that the responses on the indicators are the result of an individual's position on the latent variable(s). The second assumption is that the manifest variables have nothing in common after controlling for the latent variable(s), which is often referred to as the *axiom of local independence*.

The two remaining assumptions concern the distributions of the latent and manifest variables. Depending on these assumptions, one obtains different kinds of latent variable models. According to Bartholomew (see [1, 3]), the four main kinds are factor analysis (FA), latent trait analysis (LTA), latent profile analysis (LPA), and **latent class analysis** (LCA) (see Table 1).

In FA and LTA, the latent variables are treated as continuous normally distributed variables. In LPA and LCA on the other hand, the latent variable is

discrete, and therefore assumed to come from a multinomial distribution. The manifest variables in FA and LPA are continuous. In most cases, their conditional distribution given the latent variables is assumed to be normal. In LTA and LCA, the indicators are dichotomous, ordinal, or nominal categorical variables, and their conditional distributions are assumed to be binomial or multinomial.

The more fundamental distinction in Bartholomew's typology is the one between continuous and discrete latent variables. A researcher has to decide whether it is more natural to treat the underlying latent variable(s) as continuous or discrete. However, as shown by Heinen [2], the distribution of a continuous latent variable model can be approximated by a discrete distribution. This shows that the distinction between continuous and discrete latent variables is less fundamental than one might initially think.

The distinction between models for continuous and discrete indicators turns out not to be fundamental at all. The specification of the conditional distributions of the indicators follows naturally from their scale types. The most recent development in latent variable modeling is to allow for a different distributional form for each indicator. These can, for example, be normal, student, lognormal, gamma, or exponential distributions for continuous variables, binomial for dichotomous variables, multinomial for ordinal and nominal variables, and Poisson, binomial, or negative-binomial for counts (see **Catalogue of Probability Density Functions**). Depending on whether the latent variable is treated as continuous or discrete, one obtains a generalized form of LTA or LCA.

References

[1] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, Arnold, London.

[2] Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*, Sage Publications, Thousand Oaks.

[3] Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman & Hall/CRC, London.

JEROEN K. VERMUNT AND JAY MAGIDSON

Table 1 Four main kinds of latent variable models

Manifest variables	Latent variable(s)	
	Continuous	Categorical
Continuous	Factor analysis	Latent profile analysis
Categorical	Latent trait analysis	Latent class analysis

Latin Squares Designs

JOHN W. COTTON

Volume 2, pp. 1037–1040

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Latin Squares Designs

Definition and History of Latin Squares

A Latin square has t rows, t columns, and t treatments. Each treatment appears once in every row and once in every column. See Jones and Kenward [13] and Kirk [14] for a description of how to construct and randomly select Latin squares of size t . Closely related magic squares may have appeared as early as 2200 B.C. [8]. Dürer's engraving 'Melancholia I' contained such a square in 1514 [9], and Arnaut Daniel, an eleventh-century troubador, employed a *sestina* form in poetry that persists to the present. Typically, the *sestina* has six six-line stanzas plus a three-line *semistanza* that is ignored here. The same six words are placed one at a time at the ends of lines with each stanza having each word once and the words being balanced across stanzas in Latin square fashion [2]. For further theory, history, and examples of *sestinas*, see [19, pp. 231–242]. In 1782, Euler examined Latin squares mathematically [20]. Fisher [7] first reported using Latin square designs in 1925 in his analyses of agricultural experiments.

Model of a Latin Square

Let the response in row i , column j , and treatment k be

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \lambda_{k'} + e_{ijk} \quad (1)$$

where α_i is the effect of row i , β_j is the effect of column j , γ_k is the effect of treatment k , $\lambda_{k'}$ is a carryover (residual) effect from the immediately prior treatment k' , and e_{ijk} is the random error for this response. Also assume that each e_{ijk} is i.i.d. $N(0, \sigma^2)$, that is, independently and identically normally distributed with mean zero and variance σ^2 . This model is called a **fixed effects model**.

To develop a mixed model (see **Linear Multi-level Models; Generalized Linear Mixed Models**), replace α_i with a random row effect, a_i , assumed to be i.i.d. $N(0, \sigma_a^2)$ and independent of the error scores. Some authors [4, 16] prefer **randomization tests** or resampling tests (also discussed in [13]) to mixed model analyses unless the persons studied were randomly selected from a specific population.

Behavioral scientists most frequently are interested in a repeated measurements Latin square with rows as persons, columns as periods (stages, trials, etc.), and Latin labels as treatments. For Period 1 with no prior treatments, $\lambda_{k'}$ is defined as 0. Equation (1) is called a *carryover model* (see **Carryover and Sequence Effects**); without $\lambda_{k'}$ it is the *classical Latin square model*. Notice, for example, for the classical model, that having every treatment once in every row and column implies that $\bar{Y}_A = \mu + \alpha_A + \bar{\beta} + \bar{\gamma} + \bar{e}_A$ and $\bar{Y}_B = \mu + \alpha_B + \bar{\beta} + \bar{\gamma} + \bar{e}_B$, making $\bar{Y}_A - \bar{Y}_B$ an unbiased estimate of the treatment effect difference $\alpha_A - \alpha_B$. *Warning: If interactions (see **Interaction Effects**) (not in Equation (1)) are present, a Latin square analysis is misleading.*

A Latin square design and a three-way factorial design both have three independent variables. Consider a 2 by 2 by 2 factorial design with factors A, B, and C. Such a design has eight cells: *A1B1C1*, *A1B1C2*, *A1B2C1*, *A1B2C2*, *A2B1C1*, *A2B1C2*, *A2B2C1*, and *A2B2C2*. When replicated, this design permits the assessment of three two-way interactions and one three-way interaction. However, one possible 2 by 2 Latin square for three variables only has four cells that correspond to the italicized cells above. Consequently, the difference in row means in this Latin square could be due to row effects, to the interaction between column and treatment effects, or to both. For other difficulties specific to 2 by 2 Latin squares, see [1, 10–12].

Analyzing a 4 by 4 Latin Square Data Set

Table 1 presents an efficient [2, pp. 185–188; 21, 22] Latin square design for carryover models, often termed a *Williams square*. The data in Table 1 come from a simulated experiment [2] on listener response to four different television episodes (treatments 1 through 4). The responses reported are applause-meter scores. Ordinarily, a conventional **analysis of variance** (ANOVA) for Latin square data yields the same results, regardless of whether persons are fixed or random effects. So, Table 2 summarizes just two (classical and carryover) ANOVAs for the data in Table 1. The classical analysis concludes that persons and treatments are significant ($p < 0.05$); the carryover analysis finds all four effects significant ($p < 0.001$). The latter

2 Latin Squares Designs

Table 1 Results of a hypothetical experiment on the effects of television episodes (T1 to T4) on subject applause-meter scores ratings (Reprinted from [2], p. 177, Table 5.1, by courtesy of Marcel Dekker, Inc.)

	Period				Row total
	1	2	3	4	
Person 1	68 (T1)	74 (T2)	93 (T4)	94 (T3)	329
Person 2	60 (T2)	66 (T3)	59 (T1)	79 (T4)	264
Person 3	69 (T3)	85 (T4)	78 (T2)	69 (T1)	301
Person 4	90 (T4)	80 (T1)	80 (T3)	86 (T2)	336
Column Total	287	305	310	326	1230

Treatment 1 Total = 276, Treatment 2 Total = 298, Treatment 3 Total = 309, Treatment 4 Total = 347.

Table 2 Classical and carryover analyses of variance for table 1 data (combining information from [2], Tables 5.2, 5.4. Reprinted by courtesy of Marcel Dekker, Inc.)

Source	<i>df</i>	Classical		Carryover	
		<i>MS</i> (All four types)	<i>F</i>	<i>MS</i> (Any type but I)	<i>F</i>
Persons (Rows)	3	267.45	10.22**	207.29	1776.79****
Periods (Cols)	3 (2) ⁺	71.08	2.72	36.58	313.57***
Treatments (Latins)	3	220.42	8.42*	266.75	2286.40****
Carryovers	3	—	—	52.22	447.57***
Error	6 (3) ⁺	26.17	—	0.117	—

* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$. **** $p < 0.0001$.

⁺ df_{Periods} drops to 2 with crossover analysis because of partial confounding of periods and carryover; df_{Error} becomes 3 because of the inclusion of carryover effects in the model.

conclusion is consistent with the simulated model parameters.

Expected Mean Squares and Procedures Related to Orthogonality and Other Design Features

In a Type I analysis (see **Type I, Type II and Type III Sums of Squares**) [18, Vol. 1, Ch. 9], one computes a sum of squares (*SS*) for the first effect, say persons, extracting all variability apparently associated with that effect. The next one extracts from the remaining *SS* a sum of squares for a second effect, reflecting any possible effect that remains. This continues until all sources of effects have been assessed. Nonorthogonal effects that result from a lack of balance of variables as in our carryover model lead to a contaminated Type I mean square for a variable that was extracted early. Thus, testing the persons effect first, then periods, treatments, and carryovers leads to [2, p. 181] $E(MS_{\text{Persons}}) = \sigma^2 + t\sigma_a^2 + Q(\text{Carryovers})$. Because $E(MS_{\text{Error}}) = \sigma^2$,

$F = MS_{\text{Persons}}/MS_{\text{Error}}$ has the *F* distribution only if $t\sigma_a^2 + Q(\text{Carryovers}) = 0$. This gives $E(MS_{\text{Persons}})$ the same value as $E(MS_{\text{Error}})$. A better option is to use a Type II analysis in which each sum of squares is what is left for one effect, given all other effects. Various sources, for example, [3, pp. 8–91] and [14, Ch. 8] provide formulas and programming procedures for a variety of statistical procedures for repeated measurement experiments such as that in Table 1. See [2, 6] for comparisons of BMDP, SAS, SPSS, and SYSTAT (see **Software for Statistical Analyses**) computer programs and analyses using the carryover model for data as in Table 1.

Analysis of Replicated Repeated Measurements Latin Square Data

See Edwards [5, pp. 372–381] for an example of a classical fixed-model analysis of a 5 by 5 Latin square experiment with five replications.

Jones and Kenward [13] reanalyzed a visual perception experiment [17] with 16 students such that two persons served in each row of an 8 by 8 Latin square for Periods 1 to 8 and also for Periods 9 to 16. They performed an extensive series of SAS PROC MIXED [15] carryover ANOVAs, some for complicated period effects and some with an antedependence correction for possible nonsphericity.

References

- [1] Cotton, J.W. (1989). Interpreting data from two-period cross-over design (also termed the replicated 2×2 Latin square design), *Psychological Bulletin* **106**, 503–515.
- [2] Cotton, J.W. (1993). Latin square designs, in *Applied Analysis of Variance in Behavioral Science*, L.K. Edwards, ed., Dekker, New York, pp. 147–196.
- [3] Cotton, J.W. (1998). *Analyzing Within-Subjects Experiments*, Erlbaum, Mahway.
- [4] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Dekker, New York.
- [5] Edwards, A.L. (1985). *Experimental Design in Psychological Research*, 5th Edition, Harper & Row, New York.
- [6] Edwards, L.K. & Bland, P.C. (1993). Some computer programs for selected ANOVA problems, in *Applied Analysis of Variance in Behavioral Science*, L.K. Edwards, ed., Dekker, New York, pp. 573–604.
- [7] Fisher, R.A. (1970). *Statistical Methods for Research Workers*, 14th Edition, Oliver & Boyd, Edinburgh, Originally published in 1925.
- [8] Fults, J.L. (1974). *Magic Squares*, Open Court, La Salle.
- [9] Gardner, M. (1961). *The 2nd Scientific American Book of Mathematical Puzzles and Diversions*, Simon and Schuster, New York.
- [10] Grizzle, J.E. (1965). The two-period change-over design and its use in clinical trials, *Biometrics* **21**, 467–480.
- [11] Grizzle, J.E. (1974). Correction, *Biometrics* **30**, 727.
- [12] Hills, M. & Armitage, P. (1979). The two-period cross-over clinical trial, *British Journal of Clinical Pharmacology* **8**, 7–20.
- [13] Jones, B. & Kenward, M.K. (2003). *Design and Analysis of Cross-over Trials*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.
- [14] Kirk, R. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [15] Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). *SAS System for Mixed Models*, SAS Institute, Inc., Cary.
- [16] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury, Pacific Grove.
- [17] McNulty, P.A. (1986). Spatial velocity induction and reference mark density, Unpublished Ph.D. thesis, University of California, Santa Barbara.
- [18] SAS Institute Inc. (1990). *SAS/STAT User's Guide. Version 6*, 4th Edition, SAS Institute, Inc., Cary.
- [19] Strand, M. & Boland, E. (2000). *The Making of a Poem: A Norton Anthology of Poetic Forms*, Norton, New York.
- [20] Street, A.P. & Street, D.J. (1987). *Combinatorics of Experimental Design*, Clarendon Press, Oxford.
- [21] Wagenaar, W. (1969). Note on the construction of digram-balanced Latin squares, *Psychological Bulletin* **72**, 384–386.
- [22] Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments, *Australian Journal of Scientific Research* **A2**, 149–168.

(See also **Educational Psychology: Measuring Change Over Time**)

JOHN W. COTTON

Laws of Large Numbers

JIM W. KAY

Volume 2, pp. 1040–1041

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Laws of Large Numbers

Laws of large numbers concern the behavior of the average of n random variables, $S_n = 1/n \sum_{i=1}^n X_i$, as n grows large. This topic has an interesting early history dating back to the work of **Bernoulli** and **Poisson** in the eighteenth and nineteenth centuries; see [6] and [9] for details. For a fairly straightforward modern introduction, see [7]. There are two types of laws, namely, *weak* and *strong*, depending on the mode of convergence of the average S_n to its limit: *weak* refers to convergence in probability and *strong* denotes convergence with probability 1. The difference between these modes is perhaps best appreciated by those with a solid grounding in measure theory; see [3] or [8], for example. There are several versions of the law depending on assumptions regarding (a) the independence of the X_i , and (b) whether the X_i follow the same probability distribution. Here, we consider the simplest case in which the X_i are assumed to be mutually independent and follow the same probability distribution with finite mean μ [4, 8]. More general versions exist in which either, or both, of assumptions (a) and (b) are relaxed, but then, other conditions are imposed. For example: if the random variables have different means and different variances, but certain moment conditions hold, then the Kolmogorov strong law is obtained [8]; if the random variables form a stationary stochastic process (see **Markov Chains**), then a strong law is available (the ‘ergodic’ theorem, [1]); if both assumptions (a) and (b) are weakened while the random variables have a martingale structure (see **Martingales**), and a moment condition is satisfied, then a strong law is available [5, 8].

The weak law states that S_n converges in probability to μ as $n \rightarrow \infty$; put more precisely, this states that for any small positive number ϵ , $\Pr\{|S_n - \mu| < \epsilon\} \rightarrow 1$ as $n \rightarrow \infty$. The strong law states that S_n converges to μ with probability 1 as $n \rightarrow \infty$; put more precisely, this states that for any small positive number ϵ , $\lim_{n \rightarrow \infty} \Pr\{|S_n - \mu| < \epsilon\} = 1$. We now consider three applications of the law.

Application 1 The frequentist concept of probability is based on the notion of the stabilization of relative frequency (see **Probability: An Introduction**). Suppose that n independent experiments

are conducted under identical conditions, and suppose that, in each experiment, the event A either occurs, with probability π , or does not occur, with probability $1 - \pi$. In the i th experiment, let X_i take the value 1 if A occurs, and 0 if A does not occur ($i = 1, \dots, n$). Then S_n , the proportion of times that the event A occurs, is the relative frequency of the event A in the first n experiments and by the strong law, as $n \rightarrow \infty$, S_n converges with probability 1 to π , the true probability of the event A .

Application 2 Suppose that it is of interest to estimate the mean and variance of the reaction time in some large population of subjects, who are faced with a forced-choice perceptual task, and that reaction times are available from a simple random sample of n subjects from the population. Let X_i be a random variable representing the reaction time for the i th subject and suppose that the X_i have finite mean μ and variance σ^2 ($i = 1, \dots, n$). Consider the usual unbiased estimators S_n of μ and $T_n = n/(n - 1)\{1/n \sum_{i=1}^n X_i^2 - \bar{X}^2\}$ of σ^2 . Then, from the strong law, it follows that S_n converges with probability 1 to the population mean μ , that $1/n \sum_{i=1}^n X_i^2$ converges with probability 1 to the expectation of the X_i^2 (which is $\mu^2 + \sigma^2$), and, therefore, that T_n converges with probability 1 to the population variance σ^2 . These results provide theoretical support for the strong consistency of statistical estimation for the sample estimators S_n and T_n (see **Estimation**).

Application 3 We now illustrate how the law of large numbers provides theoretical support for the estimation of the P value associated with a Monte Carlo test. An experiment involving associative memory was conducted with a black-capped chickadee; see [2].

The bird was shown that there was food in a particular feeder and later had to find the food when presented with it and four other empty feeders. The baited feeder was determined randomly and the process was repeated 30 times. In the absence of learning, the order of the feeders that the bird looks in will be independent of where the food is, and so the bird is equally likely to find the food in the first, second, third, fourth, and fifth feeder it examines. The chickadee found the food 12 times in the first feeder it looked in, 8 times in the second, 5 times in the third, 5 times in the fourth and it never had to examine five feeders in order to get

2 Laws of Large Numbers

the food. Initially, it is of interest to test the null hypothesis that the chickadee is guessing randomly against the alternative hypothesis that this is not the case. Under the null hypothesis, no learning had occurred, and we would expect that the bird would find the food six times in each position, and one could calculate a chi-square test statistic to be 13. The P value associated with the test is the probability that the chi-squared test statistic takes a value of 13 or greater, assuming that the null hypothesis is true. Suppose now that n data sets of size 30 are generated assuming the null hypothesis to be true; that is, 1, 2, 3, 4, or 5 feeders are equally likely to be examined in any trial. Then, for each data set, the value of the chi-squared statistic is computed. Let X_i take the value 1, if the i th value of the test statistic is greater than or equal to 13, and 0 otherwise ($i = 1, \dots, n$). Then, S_n gives the proportion of times that the test statistic is greater than or equal to 13, given that the null hypothesis is true. By the strong law, S_n converges with probability 1 to the true P value associated with the test, and the estimate can be made as precise as required by generating the required number of data sets under the null hypothesis.

References

- [1] Breiman, L. (1968). *Probability*, Addison-Wesley, Reading, MA.
- [2] Brodbeck, D.R., Burack, O.R. & Shettleworth, S.J. (1992). One-trial associative memory in black-capped chickadees, *Journal of Experimental Psychology: Animal Behaviour Processes* **18**, 12–21.
- [3] Chung, K.L. (2001). *A Course in Probability Theory*, 3rd Edition, Academic Press, San Diego.
- [4] Feller, W. (1968). *An Introduction to Probability and Its Applications: Volume I*, 3rd Edition, Wiley, New York.
- [5] Feller, W. (1971). *An Introduction to Probability and Its Applications: Volume II*, 2nd Edition, Wiley, New York.
- [6] Hacking, I. (1990). *The Taming of Chance*, Cambridge University Press, Cambridge, UK.
- [7] Isaac, R. (1995). *The Pleasures of Probability*, Springer-Verlag, New York.
- [8] Sen, P.K. & Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*, Chapman & Hall, London.
- [9] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*, The Belknap Press of Harvard University Press, Cambridge, MA.

(See also **Probability: Foundations of**)

JIM W. KAY

Least Squares Estimation

SARA A. VAN DE GEER

Volume 2, pp. 1041–1045

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Least Squares Estimation

The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other (see **Optimization Methods**). We will study the method in the context of a regression problem, where the variation in one variable, called the response variable Y , can be partly explained by the variation in the other variables, called covariables X (see **Multiple Linear Regression**). For example, variation in exam results Y are mainly caused by variation in abilities and diligence X of the students, or variation in survival times Y (see **Survival Analysis**) are primarily due to variations in environmental conditions X . Given the value of X , the best prediction of Y (in terms of mean square error – see **Estimation**) is the mean $f(X)$ of Y given X . We say that Y is a function of X plus noise:

$$Y = f(X) + \text{noise}.$$

The function f is called a regression function. It is to be estimated from sampling n covariables and their responses $(x_1, y_1), \dots, (x_n, y_n)$.

Suppose f is known up to a finite number $p \leq n$ of parameters $\beta = (\beta_1, \dots, \beta_p)'$, that is, $f = f_\beta$. We estimate β by the value $\hat{\beta}$ that gives the best fit to the data. The least squares estimator, denoted by $\hat{\beta}$, is that value of b that minimizes

$$\sum_{i=1}^n (y_i - f_\beta(x_i))^2, \quad (1)$$

over all possible b .

The least squares criterion is a computationally convenient measure of fit. It corresponds to **maximum likelihood estimation** when the noise is normally distributed with equal variances. Other measures of fit are sometimes used, for example, least absolute deviations, which is more robust against **outliers**. (See **Robust Testing Procedures**).

Linear Regression. Consider the case where f_β is a linear function of β , that is,

$$f_\beta(X) = X_1\beta_1 + \dots + X_p\beta_p. \quad (2)$$

Here $(X_1, \dots, X_p)$ stand for the observed variables used in $f_\beta(X)$.

To write down the least squares estimator for the linear regression model, it will be convenient to use matrix notation. Let $\mathbf{y} = (y_1, \dots, y_n)'$ and let $\mathbf{X}$ be the $n \times p$ data matrix of the n observations on the p variables

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,p} \\ \vdots & \dots & \vdots \\ \mathbf{x}_{n,1} & \dots & \mathbf{x}_{n,p} \end{pmatrix} = (\mathbf{x}_1 \dots \mathbf{x}_p), \quad (3)$$

where $\mathbf{x}_j$ is the column vector containing the n observations on variable j , $j = 1, \dots, p$. Denote the squared length of an n -dimensional vector $\mathbf{v}$ by $\|\mathbf{v}\|^2 = \mathbf{v}'\mathbf{v} = \sum_{i=1}^n v_i^2$. Then expression (1) can be written as

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2,$$

which is the squared distance between the vector $\mathbf{y}$ and the linear combination $\mathbf{b}$ of the columns of the matrix $\mathbf{X}$. The distance is minimized by taking the *projection* of $\mathbf{y}$ on the space spanned by the columns of $\mathbf{X}$ (see Figure 1).

Suppose now that $\mathbf{X}$ has full column rank, that is, no column in $\mathbf{X}$ can be written as a linear combination of the other columns. Then, the least squares estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}. \quad (4)$$

The Variance of the Least Squares Estimator.

In order to construct **confidence intervals** for the components of $\hat{\beta}$, or linear combinations of these components, one needs an estimator of the covariance

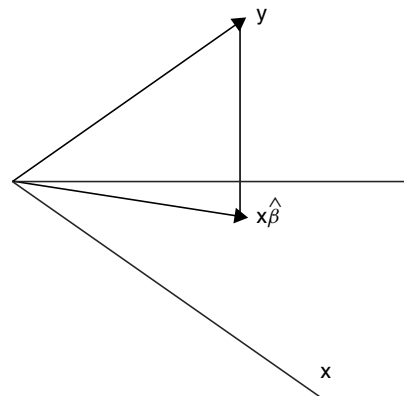


Figure 1 The projection of the vector $\mathbf{y}$ on the plane spanned by $\mathbf{X}$

2 Least Squares Estimation

matrix of $\hat{\beta}$. Now, it can be shown that, given $\mathbf{X}$, the covariance matrix of the estimator $\hat{\beta}$ is equal to

$$(\mathbf{X}'\mathbf{X})^{-1}\sigma^2,$$

where σ^2 is the variance of the noise. As an estimator of σ^2 , we take

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 = \frac{1}{n-p} \sum_{i=1}^n \hat{e}_i^2, \quad (5)$$

where the $\hat{e}_i$ are the residuals

$$\hat{e}_i = y_i - \mathbf{x}_{i,1}\hat{\beta}_1 - \cdots - \mathbf{x}_{i,p}\hat{\beta}_p. \quad (6)$$

The covariance matrix of $\hat{\beta}$ can, therefore, be estimated by

$$(\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2.$$

For example, the estimate of the variance of $\hat{\beta}_j$ is

$$\hat{\text{var}}(\hat{\beta}_j) = \tau_j^2 \hat{\sigma}^2,$$

where τ_j^2 is the j th element on the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$. A confidence interval for β_j is now obtained by taking the least squares estimator $\hat{\beta}_j \pm$ a margin:

$$\hat{\beta}_j \pm c\sqrt{\hat{\text{var}}(\hat{\beta}_j)}, \quad (7)$$

where c depends on the chosen confidence level. For a 95% confidence interval, the value $c = 1.96$ is a good approximation when n is large. For smaller values of n , one usually takes a more conservative c using the tables for the student distribution with $n - p$ degrees of freedom.

Numerical Example. Consider a regression with constant, linear and quadratic terms:

$$f_{\beta}(X) = \beta_1 + X\beta_2 + X^2\beta_3. \quad (8)$$

We take $n = 100$ and $x_i = i/n$, $i = 1, \dots, n$. The matrix $\mathbf{X}$ is now

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}. \quad (9)$$

This gives

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 100 & 50.5 & 33.8350 \\ 50.5 & 33.8350 & 25.5025 \\ 33.8350 & 25.5025 & 20.5033 \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.0937 & -0.3729 & 0.3092 \\ -0.3729 & 1.9571 & -1.8189 \\ 0.3092 & -1.8189 & 1.8009 \end{pmatrix}. \quad (10)$$

We simulated n independent standard normal random variables $e_1, \dots, e_n$, and calculated for $i = 1, \dots, n$,

$$y_i = 1 - 3x_i + e_i. \quad (11)$$

Thus, in this example, the parameters are

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}. \quad (12)$$

Moreover, $\sigma^2 = 1$. Because this is a simulation, these values are known.

To calculate the least squares estimator, we need the values of $\mathbf{X}'\mathbf{y}$, which, in this case, turn out to be

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} -64.2007 \\ -52.6743 \\ -42.2025 \end{pmatrix}. \quad (13)$$

The least squares estimate is thus

$$\hat{\beta} = \begin{pmatrix} 0.5778 \\ -2.3856 \\ -0.0446 \end{pmatrix}. \quad (14)$$

From the data, we also calculated the estimated variance of the noise, and found the value

$$\hat{\sigma}^2 = 0.883. \quad (15)$$

The data are represented in Figure 2. The dashed line is the true regression $f_{\beta}(x)$. The solid line is the estimated regression $f_{\hat{\beta}}(x)$.

The estimated regression is barely distinguishable from a straight line. Indeed, the value $\hat{\beta}_3 = -0.0446$ of the quadratic term is small. The estimated variance of $\hat{\beta}_3$ is

$$\hat{\text{var}}(\hat{\beta}_3) = 1.8009 \times 0.883 = 1.5902. \quad (16)$$

Using $c = 1.96$ in (7), we find the confidence interval

$$\beta_3 \in -0.0446 \pm 1.96\sqrt{1.5902} = [-2.5162, 2.470]. \quad (17)$$

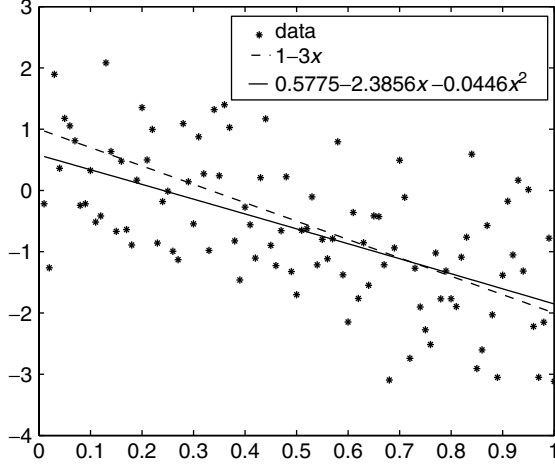


Figure 2 Observed data, true regression (dashed line), and least squares estimate (solid line)

Thus, β_3 is not significantly different from zero at the 5% level, and, hence, we do not reject the hypothesis $H_0: \beta_3 = 0$.

Below, we will consider general test statistics for testing hypotheses on β . In this particular case, the test statistic takes the form

$$T^2 = \frac{\hat{\beta}_3^2}{\text{var}(\hat{\beta}_3)} = 0.0012. \quad (18)$$

Using this test statistic is equivalent to the above method based on the confidence interval. Indeed, as $T^2 < (1.96)^2$, we do not reject the hypothesis $H_0: \beta_3 = 0$.

Under the hypothesis $H_0: \beta_3 = 0$, we use the least squares estimator

$$\begin{pmatrix} \hat{\beta}_{1,0} \\ \hat{\beta}_{2,0} \end{pmatrix} = (\mathbf{X}_0' \mathbf{X}_0)^{-1} \mathbf{X}_0' \mathbf{y} = \begin{pmatrix} 0.5854 \\ -2.4306 \end{pmatrix}. \quad (19)$$

Here,

$$\mathbf{X}_0 = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}. \quad (20)$$

It is important to note that setting β_3 to zero changes the values of the least squares estimates of β_1 and β_2 :

$$\begin{pmatrix} \hat{\beta}_{1,0} \\ \hat{\beta}_{2,0} \end{pmatrix} \neq \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}. \quad (21)$$

This is because $\hat{\beta}_3$ is correlated with $\hat{\beta}_1$ and $\hat{\beta}_2$. One may verify that the correlation matrix of $\hat{\beta}$ is

$$\begin{pmatrix} 1 & -0.8708 & 0.7529 \\ -0.8708 & 1 & -0.9689 \\ 0.7529 & -0.9689 & 1 \end{pmatrix}.$$

Testing Linear Hypotheses. The testing problem considered in the numerical example is a special case of testing a linear hypothesis $H_0: A\beta = 0$, where A is some $r \times p$ matrix. As another example of such a hypothesis, suppose we want to test whether two coefficients are equal, say $H_0: \beta_1 = \beta_2$. This means there is one restriction $r = 1$, and we can take A as the $1 \times p$ row vector

$$A = (1, -1, 0, \dots, 0). \quad (22)$$

In general, we assume that there are no linear dependencies in the r restrictions $A\beta = 0$. To test the linear hypothesis, we use the statistic

$$T^2 = \frac{\|\mathbf{X}\hat{\beta}_0 - \mathbf{X}\hat{\beta}\|^2/r}{\hat{\sigma}^2}, \quad (23)$$

where $\hat{\beta}_0$ is the least squares estimator under $H_0: A\beta = 0$. In the numerical example, this statistic takes the form given in (18). When the noise is normally distributed, critical values can be found in a table for the F distribution with r and $n - p$ degrees of freedom. For large n , approximate critical values are in the table of the χ^2 distribution with r degrees of freedom.

Some Extensions

Weighted Least Squares. In many cases, the variance σ_i^2 of the noise at measurement i depends on x_i . Observations where σ_i^2 is large are less accurate, and, hence, should play a smaller role in the estimation of β . The *weighted* least squares estimator is that value of b that minimizes the criterion

$$\sum_{i=1}^n \frac{(y_i - f_b(x_i))^2}{\sigma_i^2}.$$

overall possible b . In the linear case, this criterion is numerically of the same form, as we can make the change of variables $\tilde{y}_i = y_i/\sigma_i$ and $\tilde{\mathbf{x}}_{i,j} = \mathbf{x}_{i,j}/\sigma_i$.

4 Least Squares Estimation

The minimum χ^2 -estimator (see **Estimation**) is an example of a weighted least squares estimator in the context of density estimation.

Nonlinear Regression. When f_β is a nonlinear function of β , one usually needs iterative algorithms to find the least squares estimator. The variance can then be approximated as in the linear case, with $\dot{f}_\beta(x_i)$ taking the role of the rows of $\mathbf{X}$. Here, $\dot{f}_\beta(x_i) = \partial f_\beta(x_i)/\partial \beta$ is the row vector of derivatives of $f_\beta(x_i)$. For more details, see e.g. [4].

Nonparametric Regression. In nonparametric regression, one only assumes a certain amount of smoothness for f (e.g., as in [1]), or alternatively, certain qualitative assumptions such as monotonicity (see [3]). Many nonparametric least squares procedures have been developed and their numerical and theoretical behavior discussed in literature. Related developments include estimation methods for models

where the number of parameters p is about as large as the number of observations n . The *curse of dimensionality* in such models is handled by applying various complexity regularization techniques (see e.g., [2]).

References

- [1] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.
- [2] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, New York.
- [3] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.
- [4] Seber, G.A.F. & Wild, C.J. (2003). *Nonlinear Regression*, Wiley, New York.

SARA A. VAN DE GEER

Leverage Plot

PAT LOVIE

Volume 2, pp. 1045–1046

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Leverage Plot

A residual versus leverage plot, or leverage plot for short, is a **scatterplot** of studentized **residuals** against hat values; the residuals are usually placed on the vertical axis. Its particular merit as a tool for regression diagnostics is in flagging up observations that are potentially high leverage points, regression **outliers**, and/or influential values (see **Multiple Linear Regression**).

In general, influential observations are both regression outliers and high leverage points, indicated by large studentized residuals and large hat values, respectively, and thus will appear in the top and bottom right corners of the plot. Observations that are regression outliers but not particularly influential will show up in the left corners, while cases having high leverage alone will tend to lie around the middle of the right margin.

In Figure 1, the observation, labeled A, toward the upper left corner of the plot (large studentized residual but fairly small hat value) is identified as a regression outlier though it is probably not especially influential. On the other hand, B in the bottom right corner (large studentized residual and large hat value) is likely to be an influential observation. The observation C which lies midway down the right margin of the plot (large hat value) is a high leverage point.

However, any cases picked out for attention in this informal way should be assessed in relation to all the data, as well as against more formal criteria. Further illustrations of simple leverage plots based on real-life data can be found in [2].

The simple plot described above can be enhanced by contours corresponding to various combinations of residual and hat value generating a constant value (usually the rule of thumb cut-off point) of

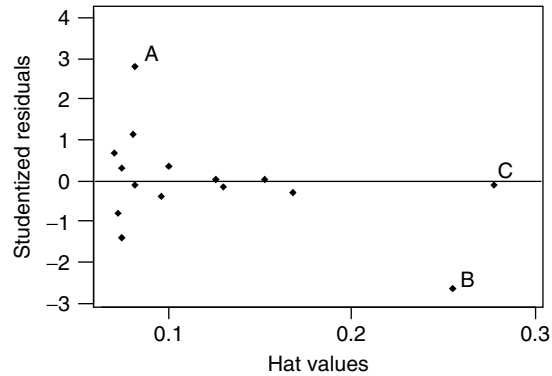


Figure 1 Leverage plot: Observation A shows up as a regression outlier, B as an influential observation, and C as a high leverage point

an influence measure such as DFITS. The contours allow the analyst to gauge the *relative* influence of observations (see, for example, [1]).

Another more sophisticated variant is based on a **bubble plot**. The studentized residuals and hat values are plotted in the usual way while the size of the circle used as the symbol is proportional to the value of the chosen influence measure for each observation.

References

- [1] Hoaglin, D.C. & Kempthorne, P.J. (1986). Comment on 'Influential observations, high leverage points and outliers in linear regression' by S. Chatterjee & A.S. Hadi, *Statistical Science* **1**, 408–412.
- [2] Lovie, P. (1991). Regression diagnostics: a rough guide to safer regression, in *New Developments in Statistics for Psychology and the Social Sciences*, Vol. 2, P. Lovie & A.D. Lovie, eds, BPS Books & Routledge, London.

PAT LOVIE

Liability Threshold Models

BEN NEALE

Volume 2, pp. 1046–1047

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Liability Threshold Models

Proposed by Pearson and Lee [5], the liability threshold model proposes that for dichotomous traits influenced by multiple factors of small effect, an underlying liability distribution exists. This population distribution of liability is assumed to be either normal or transformable to a normal distribution, with a mean of 0 and a variance of 1. On this liability distribution, there exists a threshold, t , above which all individuals exhibit the trait, and below which no individuals exhibit the trait (see Figure 1). An estimate of t can be estimated by determining the population frequency of the trait and inverting the normal distribution (also known as the *probit function*) with respect to the population frequency. Conceptually, this liability distribution reflects the biometrical model of quantitative genetics, with liability as the quantitative trait. Additionally, the model originates from the Pearson's formulation of the **tetrachoric correlation**.

This model can be extended to include multiple thresholds, when dealing with an ordered polychotomous variable (see Figure 1), such as categories including none, some, most, all or mild, moderate,

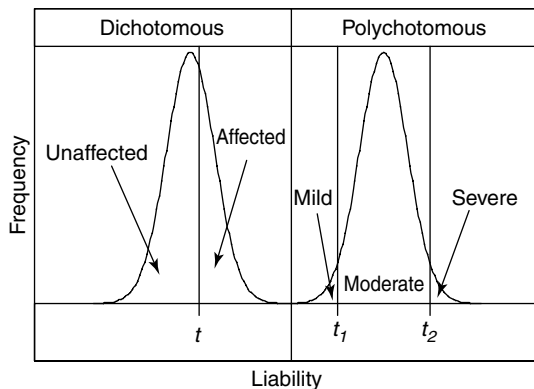


Figure 1 The two normal distributions represent the liability to develop the disorder. In the dichotomous case, t represents the threshold above which an individual will be come affected. In the polychotomous example, two thresholds exist, t_1 and t_2 . If an individual falls above t_1 and below t_2 , then he or she meets the criteria for mild affection status, whereas if an individual falls above t_2 , then he or she is severely affected

and severe diagnosis. The number of thresholds estimated by this model equals one less than the number of categories of the variable.

In practice, the liability threshold model is almost globally assumed for a genetic standpoint. In the context of **structural equation modelling** and the classical twin method (see **Twin Designs**), this model provides the backbone for the **maximum likelihood estimation** of ordinal data. Rather than estimate the means and variances of the traits, as is the case with continuous data, the thresholds for the liability distribution are estimated, while assuming that the distribution is normal, with a mean of 0 and a variance of 1. This approach is attractive when dealing with not only disease status, such as is the case with schizophrenia [2] or depression [1], but also when assessing measures such as neuroticism via the Eysenck Personality Questionnaire (EPQ) [4].

The same statistical procedures used to correct mean differences can be equally applied to thresholds. Conceptually, these procedures reflect mean shift in the liability distribution caused by a tractable factor, and act equally with respect to each threshold. For example, in the presence of an age effect, such that as age increases the threshold for affection decreases (e.g., Alzheimer's), an age regression on the single threshold can be applied.

One major issue with the liability threshold model for **2 x 2 tables** is that the model is fully specified. In the 2×2 table case, two thresholds and the correlation between the two assumed normal distributions are estimated. As the degrees of freedom for the model are defined by the number of elements in the table minus any constraints, in the case of the 2×2 table, we have 4 elements, but from a percentage perspective an element must be constrained because the proportion of the table sum to 1. Thus, we have 3 degrees of freedom, the same as the number of parameters, yielding a saturated model. So in the case of the 2×2 table, the liability threshold model represents a transformation of the data to a normal distribution rather than a model that can be tested. However, this issue is only relevant to the 2×2 case because with the increase in the number of categories of either of the variables comes an increase in the degrees of freedom, allowing for model testing, under the assumption of bivariate normal liability distribution [3]. In the case of univariate twin modelling, this

2 Liability Threshold Models

implies that the thresholds for twin 1 and twin 2 are equivalent for the single trait.

References

- [1] Kendler, K.S., Pedersen, N.L., Neale, M.C. & Mathe, A.A. (1995). A pilot Swedish twin study of affective illness including hospital- and population-ascertained subsamples: results of model fitting, *Behavior Genetics* **25**(3), 217–232.
- [2] McGuffin, P., Farmer, A.E., Gottesman, I.I., Murray, R.M. & Reveley, A.M. (1984). Twin concordance for operationally defined schizophrenia. Confirmation of familiarity and heritability, *Archives of General Psychiatry* **41**(6), 541–545.
- [3] Neale, M.C. & Maes, H.H.M. (2004). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, B.V. Dordrecht.
- [4] Neale, M.C., Rushton, J.P. & Fulker, D.W. (1986). Heritability of item responses on the Eysenck Personality Questionnaire, *Personality and Individual Differences* **7**, 771–779.
- [5] Pearson, K. & Lee, A. (1901). On the inheritance of characters not capable of exact quantitative measurement, *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences* **195**, 79–150.

BEN NEALE

Linear Model

JOSE CORTINA

Volume 2, pp. 1048–1049

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Linear Model

Although this term is used in a variety of contexts, its meaning is relatively simple. A linear model exists if the phenomenon in question can be accurately described with a ‘flat’ surface.

Consider a relationship between two variables A and C . If the surface that best describes this relationship is a straight line, then a linear model applies (Figure 1).

As can be seen in this **scatterplot**, the data points trend upward, and they do so in such a way that a straight line characterizes their ascent quite well.

Consider instead Figure 2.

Here, the linear model fits the data very poorly. This is reflected in the fact that the bivariate correlation between these two variables is zero. It is clear that there is a strong, but curvilinear, relationship between these two variables. This might be captured with the addition of an A -squared term to an equation that already contains A . This curvilinear model explains 56% of the variance in C , whereas the linear model (i.e., one that contains only A) explains none.

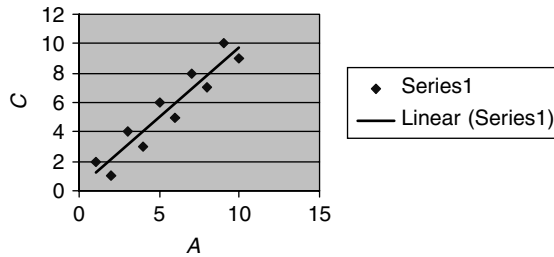


Figure 1 Scatterplot representing a linear bivariate relationship

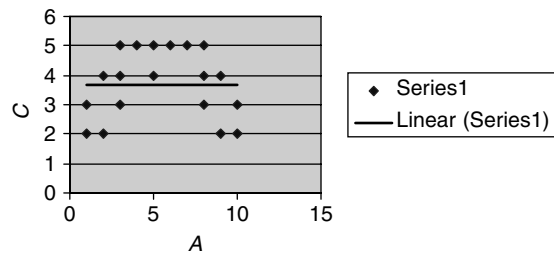


Figure 2 Scatterplot representing a curvilinear bivariate relationship

If we were to create a **three dimensional scatterplot** of the relationship among two predictors, A and B , and a criterion C , the data might conform to any number of shapes. A cylindrical form trending upward and outward from the origin would suggest positive relationships between the predictors and the criterion. Because we have multiple predictors, the best-fitting surface is a plane rather than a line. In the cylindrical case, the best-fitting plane would bisect the cylinder lengthwise. If the best-fitting surface is a flat plane, then the data conform to a linear model. If instead a curved plane fits best, then the best-fitting model is curvilinear. This would be the case if the cylinder trended upward initially, but then headed downward. This would also be the case if A and B interact in their effect on C . In this latter case, the scatterplot would likely be saddle-shaped rather than cylindrical.

The issue becomes more complicated if we consider that many curvilinear models are ‘intrinsically linear’. A model is intrinsically linear if it can be represented as

$$\begin{aligned} f(Y) = & g_0(b_0) + b_1 g_1(X_1, \dots, X_j) \\ & + b_2 g_2(X_1, \dots, X_j) + \dots \\ & + b_k g_k(X_1, \dots, X_j) + h(e). \end{aligned} \quad (1)$$

Here we have a dependent variable (DV) Y and a set of predictors (X s). The model that captures the relationship between the predictors and the criterion is said to be intrinsically linear if some function of Y equals some function of the intercept plus a weight (b_1) times some function of the predictors plus another weight times some other function of the predictors, and so on, plus some function of error. Consider the curvilinear model presented earlier. We might test this relationship with the equation

$$C' = b_0 + b_1(A) + b_2(A^2) \quad (2)$$

where b_0 is the intercept, b_1 is expected to be zero because there is no overall upward or downward trend, and b_2 is expected to be large and negative because the slope of the line representing the AC relationship gets smaller as A increases. Although this model contain a nonlinear term (A^2), it is said to be intrinsically linear because C is estimated by an intercept, a weight times a function of the predictors ($1 * A$), and another weight times another function of the predictors ($A * A$). Another way of stating

2 Linear Model

this is that, although the surface of best fit for the relationship between A and C is distinctly nonlinear, the surface of best fit for the relationship between A , A^2 , and C is a flat plane.

Estimators such as ordinary least squares (OLS), require intrinsic linearity (*see* **Least Squares Estimation**). It is primarily for this reason that OLS regression cannot be used with dichotomous dependent

variables. It can be shown that the S-shaped surface that usually fits relationship involving dichotomous DVs best cannot be represented merely as weights time functions of predictors. The fact forms the basis of **logistic regression**, typically based on **maximum likelihood estimation**.

JOSE CORTINA

Linear Models: Permutation Methods

BRIAN S. CADE

Volume 2, pp. 1049–1054

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Linear Models: Permutation Methods

Permutation tests (*see* **Permutation Based Inference**) for the **linear model** have applications in behavioral studies when traditional parametric assumptions about the error term in a linear model are not tenable. Improved validity of Type I error rates can be achieved with properly constructed permutation tests. Perhaps more importantly, increased statistical **power**, improved robustness to effects of **outliers**, and detection of alternative distributional differences can be achieved by coupling permutation inference with alternative linear model estimators. For example, it is well-known that estimates of the mean in the linear model are extremely sensitive to even a single outlying value of the dependent variable compared to estimates of the median [7, 19]. Traditionally, linear modeling focused on estimating changes in the center of distributions (means or medians). However, quantile regression allows distributional changes to be estimated in all or any selected part of a distribution or responses, providing a more complete statistical picture that has relevance to many biological questions [6].

Parameters from the linear model in either its location, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, or location scale, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\varepsilon}$, form can be tested with permutation arguments. Here, $\mathbf{y}$ is an $n \times 1$ vector of dependent responses, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, $\mathbf{X}$ is an $n \times p$ matrix of predictors (with commonly the first column consisting of 1's for an intercept term), $\boldsymbol{\Gamma}$ is a diagonal $n \times n$ matrix where the n diagonal elements are the n corresponding ordered elements of the $n \times 1$ vector $\mathbf{X}\boldsymbol{\gamma}$ ($\text{diag}(\mathbf{X}\boldsymbol{\gamma})$), $\boldsymbol{\gamma}$ is a $p \times 1$ vector of unknown scale parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors that are independent and identically distributed (iid) with density $f_{\boldsymbol{\varepsilon}}$, distribution $F_{\boldsymbol{\varepsilon}}$, and quantile $F_{\boldsymbol{\varepsilon}}^{-1}$ functions. Various parametric regression models are possible, depending on which parameter of the error distribution is restricted to equal zero; for example, setting the expected value $F_{\boldsymbol{\varepsilon}}(\boldsymbol{\mu}|\mathbf{X}) = 0$ yields the familiar mean regression, setting any quantile $F_{\boldsymbol{\varepsilon}}^{-1}(\tau|\mathbf{X}) = 0$ yields quantile ($0 \leq \tau \leq 1$) regression, and the special case of $F_{\boldsymbol{\varepsilon}}^{-1}(0.5|\mathbf{X}) = 0$ yields median (least absolute deviation) regression [14]. The location

model with homoscedastic error variance is just a special case of the linear-location scale model when $\boldsymbol{\gamma} = (1, 0, \dots, 0)'$.

Estimates ($\hat{\boldsymbol{\beta}}$) of the various parametric linear models are obtained by minimizing appropriate loss functions of the residuals, $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Minimizing the sum of squared **residuals** yields the **least squares estimates** of the mean model. Minimizing the sum of asymmetrically weighted (τ for + residuals and $1 - \tau$ for - and 0 residuals) absolute values of the residuals yields the quantile regression estimates, where least absolute deviation regression for the median ($\tau = 0.5$) model being just a special case [15]. Consistent estimates with reduced sampling variation can be obtained for linear location-scale models by implementing weighted versions of the estimators, where weights are the reciprocal of the scale parameters, $\mathbf{W} = \boldsymbol{\Gamma}^{-1}$. In applications, the $p \times 1$ vector of scale parameters $\boldsymbol{\gamma}$ would usually have to be estimated. Weighted regression estimates are obtained by multiplying $\mathbf{y}$ and $\mathbf{X}$ by $\mathbf{W}$ and minimizing the appropriate function of the residuals as before. Examples of various estimates are shown in Figure 1.

Test Statistic

A drop in dispersion, F -ratio-like, test statistic that is capable of testing hypotheses for individual or multiple coefficients can be evaluated by similar permutation arguments for any of the linear model estimators above [2, 5, 7, 19]. This pivotal test statistic takes the form $T = (S_{\text{reduced}} \div S_{\text{full}}) - 1$, where S_{reduced} is the sum minimized by the chosen estimator for the reduced parameter model specified by the null hypothesis ($\mathbf{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\xi}$) and S_{full} is the sum minimized for the full parameter model specified by the alternative hypothesis. This is equivalent to the usual F -ratio statistic for least squares regression but the degrees of freedom for reduced and full parameter models are deleted because they are not needed as they are invariant under the permutation arguments to follow. The reduced parameter model $\mathbf{y} - \mathbf{X}_2\boldsymbol{\xi} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ is constructed by partitioning $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is $n \times (p - q)$ and $\mathbf{X}_2$ is $n \times q$; and by partitioning $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ where $\boldsymbol{\beta}_1$ is a $(p - q) \times 1$ vector of unknown nuisance parameters under the null and $\boldsymbol{\beta}_2$ is the $q \times 1$ vector of parameters specified by the null hypothesis $\mathbf{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\xi}$

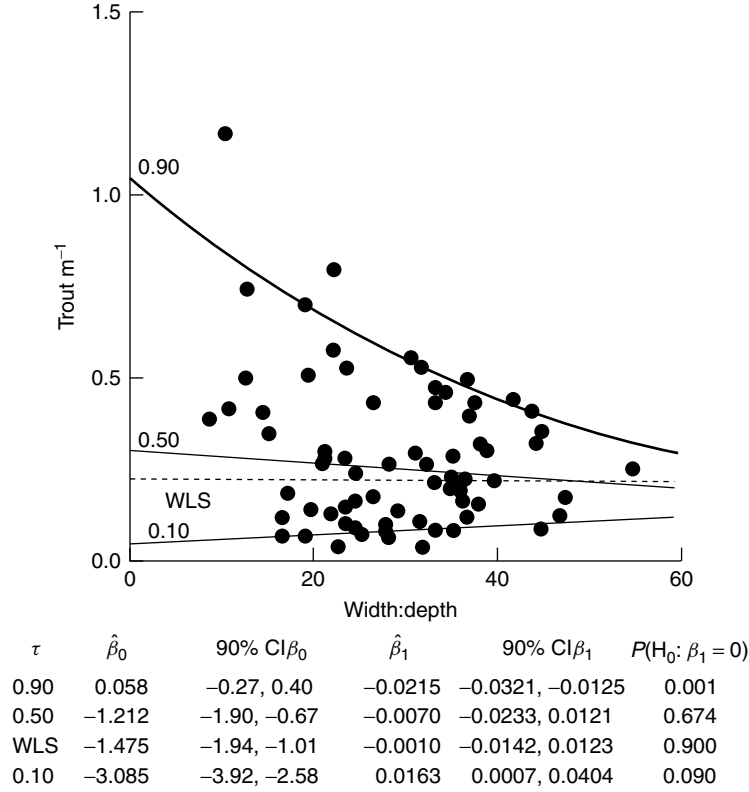


Figure 1 Lahontan cutthroat trout m^{-1} and width:depth ratios for small streams sampled 1993 to 1999 ($n = 71$) [10]; exponentiated estimates for 0.90, 0.50, and 0.10 regression quantiles and least squares (WLS) estimate for the weighted model $w(\ln y) = w(\beta_0 + \beta_1 X_1 + (\gamma_0 + \gamma_1 X_1)\epsilon)$, $w = (1.310 - 0.0017X_1)^{-1}$. Tabled values are estimates, 90% confidence intervals from inverting the permutation test, and estimated probabilities for $H_0: \beta_1 = 0$ based on the double permutation procedure for statistic T made with $m + 1 = 100000$ permutations

(frequently $\beta_2 = 0$). The unconstrained, full parameter model is $y = X_1\beta_1 + X_2\beta_2 + \epsilon$. To test hypotheses on weighted estimates in the linear location-scale model $y = X_1\beta_1 + X_2\beta_2 + \Gamma\epsilon$, the terms y , X_1 , and X_2 are replaced with their weighted counterparts Wy , WX_1 , and WX_2 , respectively, where W is the weights matrix, and the test statistic is constructed similarly. For homogeneous error models, $W = I$, where I is the $n \times n$ identity matrix.

When testing hypotheses on a single parameter, for example, $H_0: \beta_j = \xi_j$, the more general T statistic may be replaced with an equivalent t -ratio form $t = (\hat{\beta}_j - \xi_j) / \sqrt{S_{\text{full}}}$, where $\hat{\beta}_j$ is the full model estimate of β_j [1, 2, 18]. Use of test statistics that are not pivotal such as the actual parameter estimates is not recommended because they fail to maintain valid Type I error rates when there is multicollinearity (see

Collinearity) among the predictor variables in X [1, 2, 13].

Permutation Distribution of the Test Statistic

The test statistic for the observed data $T_{\text{obs}} = (S_{\text{reduced}} \div S_{\text{full}}) - 1$ is compared to a reference distribution of T formed by permutation arguments. The relevant exchangeable quantities under the null hypothesis to form a reference distribution are the errors ϵ from the null, reduced parameter model [1, 3, 17]. These can be permuted (shuffled) across the rows of X with equal probability. In general, the errors are unknown and unobservable. But, in the case of simultaneously testing all parameters other than the

intercept for the location model when $\mathbf{W} = \mathbf{I}$, $\mathbf{X}_1$ is an $n \times 1$ matrix of 1's, the residuals from the null, reduced parameter model $\mathbf{e}_{\text{red}} = \mathbf{y} - \mathbf{X}_1 \hat{\beta}_1$ or $\mathbf{y}$ differ from the errors $\boldsymbol{\varepsilon}$ by unknown constants that are invariant under permutation. Thus, in this case, a reference distribution for T can be constructed by either permuting $\mathbf{e}_{\text{red}}$ or $\mathbf{y}$ against the full model matrix $\mathbf{X}$ to yield probabilities under the null hypothesis (proportion of $T \geq T_{\text{obs}}$) that are exact, regardless of the unknown distribution of the errors [2, 5, 18]. The variables being tested in $\mathbf{X}_2$ may be continuous, indicators (1, 0, -1) for categorical groups, or both. Thus, exact probabilities for the null hypothesis are possible for the 1-way classification (ANOVA) (see **Analysis of Variance**) model for two or more treatment groups, for the single-slope parameter in a simple regression, and for all slope parameters simultaneously in a **multiple regression**. In practice, for reasonable n , a very large random sample of size m is taken from the $n!$ possible permutations so that probability under the null hypothesis is estimated by (the number of $T \geq T_{\text{obs}} + 1)/(m + 1)$. The error of estimating the P value by Monte Carlo resampling (see **Monte Carlo Simulation**) can be made arbitrarily small by using a large m , for example, $m + 1 \geq 10000$.

For null hypotheses on subsets of parameters from a linear model with multiple predictors, the reference distribution of T is approximated by permuting $\mathbf{e}_{\text{red}}$ against $\mathbf{X}$ [1, 2, 5, 7, 11, 21]. Permuting against the full model matrix $\mathbf{X}$ ensures that the correlation structure of the predictor variables is fixed, that is, the design is ancillary. As the residuals no longer differ from the errors $\boldsymbol{\varepsilon}$ by a constant, they are not exchangeable with equal probability and the resulting probability for the null hypothesis is no longer exact [1, 5, 9, 12]. Regardless, this permutation approach originally due to Freedman and Lane [11] was found to have perfect correlation asymptotically with the exact test for least squares regression (as if the errors $\boldsymbol{\varepsilon}$ were known) [3] and has performed well in simulations for least squares [2], least absolute deviation [7], and quantile regression [5]. Some authors have permuted $\mathbf{e}_{\text{full}} = \mathbf{y} - \mathbf{X} \hat{\beta}$ rather than $\mathbf{e}_{\text{red}}$ [1, 18, 21], but there is less theoretical justification for doing so, although it may yield similar results asymptotically for least squares regression estimates [2, 3]. There are alternative restricted permutation schemes that provide valid probabilities for linear models when the null hypothesis and hence $\mathbf{X}_2$ only

include indicator variables for categorical treatment groups [1, 4, 19, 20]. Permutation tests for random (see **Completely Randomized Design**) and mixed effects in multifactorial linear models (see **Linear Multilevel Models**) are discussed in [4].

There are several approaches to improving exchangeability of the residuals $\mathbf{e}_{\text{red}}$ under the null hypothesis to provide more valid Type I error rates. For the linear least squares regression estimator, linear **transformations** toward approximate exchangeability of the residuals from a model with $p - q$ parameters must reduce the dimension of $\mathbf{e}_{\text{red}}$ to $n - (p - q)$ or $n - (p - q) + 1$ and reduce $\mathbf{X}$ (or $\mathbf{WX}$) to conform, for example, by Gram-Schmidt orthogonalization [9]. This theory is not directly applicable to a nonlinear estimator like that used for quantile regression. But reducing the dimension of $\mathbf{e}_{\text{red}}$ by deleting $(p - q) - 1$ of the zero residuals and randomly deleting $(p - q) - 1$ rows of $\mathbf{X}$ to conform was found to improve Type I error rates for null hypotheses involving subsets of parameters for both linear location and location-scale quantile regression models with multiple independent variables [8]. This approach was motivated by [9] and the fact that quantile regression estimates for $p - q$ parameters must have at least $p - q$ zero residuals. An example of the Type I error rates for corrected and uncorrected residuals $\mathbf{e}_{\text{red}}$ for a 0.90 quantile regression model is shown in Figure 2. Another option for quantile regression is to use a τ -rank score procedure on $\mathbf{e}_{\text{red}}$, which transforms the $+$ residuals to τ , $-$ residuals to $\tau - 1$, and zero residuals to values in the interval $(\tau - 1, \tau)$, which are then used to compute a test statistic similar to the one above but based on a least squares (or weighted least squares) regression of the rank scores $\mathbf{r}$ on $\mathbf{X}$ [5]. This τ -rank score test can be evaluated by permuting $\mathbf{r}$ across the rows of $\mathbf{X}$ to yield a reference distribution to compute a probability under the null hypothesis. An example of the Type I error rates associated with this procedure also is shown in Figure 2.

An additional complication with permutation testing for the linear models occurs whenever the null model specified by the hypothesis does not include an intercept term so that the estimates are constrained through the origin. This includes testing a null hypothesis that includes the intercept term or when testing subsets of weighted parameter estimates for variables that are part of the weights function. Residuals from the estimates for the null, reduced parameter model $\mathbf{e}_{\text{red}}$ are no longer guaranteed to be

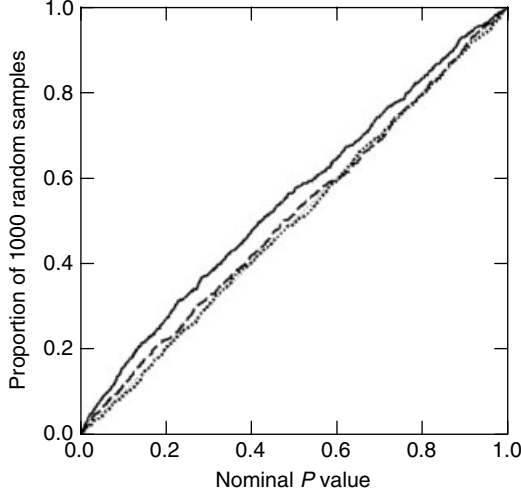


Figure 2 Cumulative distributions of 1000 estimated Type I errors for permutation tests of $H_0 : \beta_3 = \beta_5 = 0$ based on uncorrected permutation of raw residuals (solid line), deletion of $(p - q) - 1 = 3$ zero residuals (dashed line) prior to permutation, and by permuting τ -quantile rank scores (thick dotted line) for the 0.90 weighted quantile regression model $wy = w(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_3 X_4 + (1 + 0.05 X_1)\varepsilon)$; $X_1 = \text{Uniform}(0, 100)$, $X_2 = 4000 - 20X_1 + N(0, 300)$, $X_3 = 10 + 0.4X_1 + N(0, 16)$, X_4 is 0,1 indicator variable with half of n randomly assigned each, $\beta_0 = 36$, $\beta_1 = 0.10$, $\beta_2 = -0.005$, $\beta_4 = 2.0$, $\beta_3 = \beta_5 = 0$, ε is lognormal ($F_\varepsilon^{-1}(0.9) = 0$, $\sigma = 0.75$), $w = (1 + 0.05 X_1)^{-1}$, and $n = 90$. Each P value was estimated with $m + 1 = 10000$ permutations. Fine dotted 1:1 line is the expected cdf

centered on their appropriate distributional parameter, for example, $F_e(\bar{x}|\mathbf{X}_1) \neq 0$, although $F_e(\mu|\mathbf{X}_1) = 0$. Instead, the estimated distributional parameter associated with 0 for $\mathbf{e}_{\text{red}}$ has random binomial sampling variation that needs to be incorporated into the permutation scheme to provide valid Type I error rates. A double permutation scheme has been proposed for providing valid Type I errors for these hypotheses [5, 8, 16]. The first step uses a random binomial variable, $\tau^* \sim \text{binomial}(\tau, n)$, to determine the value on which the residuals $\mathbf{e}_{\text{red}}$ are centered, $\mathbf{e}_{\text{red}}^* = \mathbf{e}_{\text{red}} - F_e^{-1}(\tau^*|\mathbf{X}_1)$, and the second step permutes the randomly centered residuals $\mathbf{e}_{\text{red}}^*$ to the matrix $\mathbf{X}$. For least squares regression, τ is taken as 0.5. An example of Type I error rates for the double permutation compared to the uncorrected standard permutation test for the hypothesis $H_0 : \beta_0 = 0$ for a 0.5 quantile regression model is shown in Figure 3.

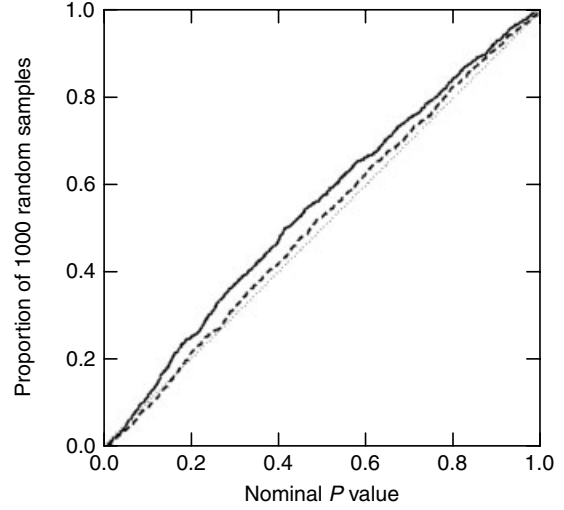


Figure 3 Cumulative distributions of 1000 estimated Type I errors for permutation tests of $H_0 : \beta_0 = 0$ based on uncorrected permutation of raw residuals (solid line) and double permutation of residuals (dashed line) for the 0.50 quantile regression model $y = \beta_0 + \beta_1 X_1 + \varepsilon$; $X_1 = \text{Uniform}(0, 100)$, $\beta_0 = 0$, $\beta_1 = 0.10$, ε is lognormal ($F_\varepsilon^{-1}(0.5) = 0$, $\sigma = 0.75$), and $n = 150$. Each P value was estimated with $m + 1 = 10000$ permutations. Fine dotted 1:1 line is the expected cdf

Example Application

In applications, we often make use of the fact that confidence intervals on parameters in a linear model may be constructed by inversion of permutation tests [5, 7, 18]. We obtain a $(1 - \alpha) \times 100\%$ confidence interval on a single parameter for a variable $\mathbf{x}_2 = \mathbf{X}_2$ by making the transformation $\mathbf{y} - \mathbf{x}_2 \xi$ on a sequence of values of ξ and collecting those values that have $P \geq \alpha$ for a test of the null hypothesis $H_0 : \beta_2 = \xi$.

The data in Figure 1 were from a study designed to evaluate changes in Lahontan cutthroat trout (*Oncorhynchus clarki henshawi*) densities as a function of stream channel morphology as it varies over the semidesert Lahontan basin of northern Nevada, USA [10]. The quantile regression analyses published in [10] used inferential procedures based on asymptotic distributional evaluations of the τ -quantile rank score statistic. Here, for a selected subset of quantiles, the 90% confidence intervals and hypothesis of zero slope were made with permutation tests based on the T statistic and permuting residuals $\mathbf{e}_{\text{red}}$. Because

the null model for the weighted estimates was implicitly forced through the origin, the double permutation scheme was required to provide valid Type I error rates. It is obvious in this location-scale model that restricting estimation and inferences to central distributional parameters, whether the mean or median, would have failed to detect changes in the trout densities at lower and higher portions of the distribution. Note that the permutation-based 90% confidence intervals for the weighted least squares estimate do not differ appreciably from the usual F distribution evaluation of the statistic nor do the intervals for the quantile estimates differ much from those based on the quantile rank score inversion [6].

Software

Computer routines for performing permutation tests on linear, least squares regression models are available in the Blossom software available from the U. S. Geological Survey (www.fort.usgs.gov/products/software/software.asp), in RT available from Western EcoSystems Technology, Inc. (www.west-inc.com/), in NPMANOVA from the web page of M. J. Anderson (www.stat.auckland.ac.nz/~mja/), and from the web page of P. Legendre (www.fas.umontreal.ca/BIOL/legendre/). The Blossom software package also features permutation tests for median and quantile regression. General permutation routines that can be customized for testing linear models are available in S-Plus and R.

References

- [1] Anderson, M.J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression, *Canadian Journal of Fisheries and Aquatic Science* **58**, 626–639.
- [2] Anderson, M.J. & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model, *Journal of Statistical Computation and Simulation* **62**, 271–303.
- [3] Anderson, M.J. & Robinson, J. (2001). Permutation tests for linear models, *Australian New Zealand Journal of Statistics* **43**, 75–88.
- [4] Anderson, M.J. & TerBraak, C.J.F. (2003). Permutation tests for multi-factorial analysis of variance, *Journal of Statistical Computation and Simulation* **73**, 85–113.
- [5] Cade, B.S. (2003). *Quantile regression models of animal habitat relationships*, Ph.D. dissertation, Colorado State University, Fort Collins.
- [6] Cade, B.S. & Noon, B.R. (2003). A gentle introduction to quantile regression for ecologists, *Frontiers in Ecology and the Environment* **1**, 412–420.
- [7] Cade, B.S. & Richards, J.D. (1996). Permutation tests for least absolute deviation regression, *Biometrics* **52**, 886–902.
- [8] Cade, B.S. & Richards, J.D. (In review). A drop in dispersion permutation test for regression quantile estimates. *Journal of Agricultural, Biological, and Environmental Statistics*.
- [9] Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests, *Journal of Nonparametric Statistics* **15**, 171–185.
- [10] Dunham, J.B., Cade, B.S. & Terrell, J.W. (2002). Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles, *Transactions of the American Fisheries Society* **131**, 86–98.
- [11] Freedman, D. & Lane, D. (1983). A nonstochastic interpretation of reported significance levels, *Journal of Business and Economic Statistics* **1**, 292–298.
- [12] Good, P. (2002). Extensions of the concept of exchangeability and their applications, *Journal of Modern Applied Statistical Methods* **1**, 243–247.
- [13] Kennedy, P.E. & Cade, B.S. (1996). Randomization tests for multiple regression, *Communications in Statistics – Simulation and Computation* **25**, 923–936.
- [14] Koenker, R. & Bassett, G. (1978). Regression quantiles, *Econometrica* **46**, 33–50.
- [15] Koenker, R. & Hallock, K.F. (2001). Quantile regression, *Journal of Economic Perspectives* **15**(4), 143–156.
- [16] Legendre, P. & Desdevises, Y. (2002). *Independent Contrasts and Regression Through the Origin*, (In review, available at www.fas.umontreal.ca/BIOL/legendre/reprints/index.html).
- [17] LePage, R. & Podgórski, K. (1993). Resampling permutations in regression with exchangeable errors, *Computing Science and Statistics* **24**, 546–553.
- [18] Manly, B.F.J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, New York.
- [19] Mielke Jr, P.W. & Berry, K.J. (2001). *Permutation Methods: A Distance Function Approach*, Springer-Verlag, New York.
- [20] Pesarin, F. (2001). *Multivariate Permutation Tests: With Applications in Biostatistics*, John Wiley & Sons, New York.
- [21] ter Braak, C.J.F. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA, in *Bootstrapping and Related Techniques*, K.H. Jöreskog, G. Rothe & W. Sendler, eds, Springer-Verlag, New York.

BRIAN S. CADE

Linear Multilevel Models

JAN DE LEEUW

Volume 2, pp. 1054–1061

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Linear Multilevel Models

Hierarchical Data

Data are often hierarchical (*see* **Hierarchical Models**). By this we mean that data contain information about observation units at various levels, where the lower-level units are nested within the higher-level units. Some examples may clarify this. In repeated measurement or growth curve data (*see* **Repeated Measures Analysis of Variance; Growth Curve Modeling**), we have several observations in time on each of a number of different individuals. The time points are the lowest level and individuals are the higher level. In school effectiveness studies we have observations on students (the lowest or first level), on the schools in which these students are enrolled (the second level), and maybe even on school districts these schools are in (a third level).

Once we have data on various levels, we have to decide at which level to analyze. We can aggregate student variables to the school level or disaggregate school variables to the student level. In the first case, we lose potentially large amounts of useful information, because information about individual students disappears from the analysis. In the second case, we artificially create dependencies in the data, because students in the same school by definition get the same score on a disaggregated school variable.

Another alternative is to do a separate analysis for each higher-level unit. For example, we do a student-level regression analysis for each school separately. This, however, tends to introduce a very large number of parameters. It also ignores the fact that it makes sense to assume the different analyses will be related, because all schools are functioning within the same education system.

Multilevel models combine information about variables at different levels in a single model, without aggregating or disaggregating. It provides more data reduction than a separate analysis for each higher-level unit, and it models the dependency between lower-level units in a natural way. Multilevel models originated in school effectiveness research, and the main textbooks discussing this class of techniques still have a strong emphasis on educational applications [5, 13]. But hierarchical data occur in many disciplines, so there are now applications in

the health sciences, in biology, in sociology, and in economics.

Linear Multilevel Model

A linear multilevel model, in the two-level case, is a regression model specified in two stages, corresponding with the two levels. We start with separate linear regression models for each higher-level unit, specified as

$$\underline{y}_j = X_j \underline{\beta}_j + \underline{\epsilon}_j. \quad (1a)$$

Higher level units (schools) are indexed by j , and there are m of them. Unit j contains n_j observations (students), and thus the outcomes $\underline{y}_j$ and the error terms $\underline{\epsilon}_j$ are vectors with n_j elements. The predictors for unit j are collected in an $n_j \times p$ matrix X_j .

In our model specification random variables, and random vectors, are underlined. This shows clearly how our model differs from classical separate linear regression models, in which the regression coefficients β_j are nonrandom. This means, in the standard frequentist interpretation, that if we were to replicate our experiment then in the classical case all replications have the same regression coefficients, while in our model the random regression coefficients would vary because they would be independent realizations of the same random vector $\underline{\beta}_j$. The differences are even more pronounced, because we also use a second-level regression model, which has the first-level regression coefficients as outcomes. This uses a second set of q regressors, at the second level. The submodel is

$$\underline{\beta}_j = Z_j \gamma + \underline{\delta}_j, \quad (1b)$$

where the Z_j are now $p \times q$ and γ is a fixed set of q regression coefficients that all second level units have in common.

In our regression model, we have not underlined the predictors in X_j and Z_j , which means we think of them as fixed values. They are either fixed by design, which is quite uncommon in social and behavioral sciences, or they are fixed by the somewhat artificial device of conditioning on the values of the predictors. In the last case, the predictors are really random variables, but we are only interested in what happens if these variables are set to their observed values.

2 Linear Multilevel Models

We can combine the specifications in (1a) and (1b) to obtain the linear mixed model

$$\underline{y}_j = X_j Z_j \gamma + X_j \underline{\delta}_j + \underline{\epsilon}_j. \quad (2)$$

This model has both fixed regression coefficients γ and random regression coefficients $\underline{\delta}_j$. In most applications, we suppose that both regressions have an intercept, which means that all X_j and all Z_j have a column with elements equal to one.

For the error terms $\underline{\delta}_j$ and $\underline{\epsilon}_j$ in the two parts of the regression model, we make the usual strong assumptions. Both have expectation zero, they are uncorrelated with each other within the same second-level unit, and they are uncorrelated between different second-level units. We also assume that first-level disturbances are homoscedastic, and that both errors have the same variance-covariance matrix in all second-level units. Thus, $V(\underline{\epsilon}_j) = \sigma^2 I$ and we write $V(\underline{\delta}_j) = \Omega$. This implies that

$$E(\underline{y}_j) = X_j Z_j \gamma, \quad (3a)$$

$$V(\underline{y}_j) = X_j \Omega X_j' + \sigma^2 I. \quad (3b)$$

Thus, we can also understand mixed linear models as heteroscedastic regression models with a specific interactive structure for the expectations and a special **factor analysis** structure for the covariance matrix of the disturbances. Observations in the same two-level unit are correlated, and thus we have correlations between students in the same school and between observations within the same individual at different time-points. We also see the correlation is related to the similarity in first-level predictor values of units i and k . Students with similar predictor values in X_j will have a higher correlation.

To explain more precisely how the $n_j \times q$ design matrices $U_j = X_j Z_j$ for the fixed effects usually look, we assume that Z_j has the form

$$Z_j = \begin{bmatrix} h_j' & 0 & \cdots & 0 \\ 0 & h_j' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_j' \end{bmatrix}, \quad (4)$$

where h_j is a vector with r predictor values for second-level unit j . Thus, $q = pr$, and $\underline{\beta}_{js} = h_j' \gamma_s + \underline{\delta}_{js}$. With this choice of Z_j , which is the usual one in multilevel models, the matrix U_j is of the form

$$U_j = [x_{j1} h_j' | \cdots | x_{jp} h_j'], \quad (5)$$

that is, each column of U is the product of a first-level predictor and a second-level predictor. All $p \times r$ cross-level interactions get their own column in U . If the X_j have their first column equal to one, and the h_j have their first element equal to one, then it follows that the columns of X_j themselves and the (disaggregated) columns of H are among the pq interactions.

In the balanced case of the multilevel model, all second-level units j have the same $n \times p$ matrix of predictors X . This happens, for example, in growth curve models, in which X contains the same fixed functions (orthogonal polynomials, for instance) of time. In the balanced case, we can collect our various observations and parameters in matrices, and write $\underline{Y} = \underline{B} X' + \underline{E}$ and $\underline{B} = Z \Gamma + \underline{\Delta}$ or $\underline{Y} = Z \Gamma X' + \underline{\Delta} X' + \underline{E}$. Here the outcomes $\underline{Y}$ are in a matrix of order $m \times n$, individuals by time points, and the fixed parameters are in a $q \times p$ matrix Γ . This shows, following Stenio et al. [17], how multilevel ideas can be used to generalize the basic **growth curve model** of Pothoff and Roy [12].

Parameter Constraints

If p and q , the number of predictors at both levels, are at all large, then obviously their product pq will be very large. Thus, we will have a linear model with a very large number of regression coefficients, and in addition to the usual residual variance parameter σ^2 we will also have to estimate the $1/2p(p+1)$ parameters in Ω . The problem of having too many parameters for fast and stable estimation is compounded by the fact that the interactions in U will generally be highly correlated, and that consequently the regression problem is ill-conditioned. This is illustrated forcefully by the relatively small examples in Kreft and De Leeuw [7].

The common procedure in multilevel analysis to deal with the large number of parameters is the same as in other forms of regression analysis. Free parameters are set equal to zero, or, to put it differently, we use variable selection procedures. Setting regression coefficients (values of γ) equal to zero is straightforward, because it simply means that cross-level interactions are eliminated from the model. Nevertheless, the usual variable selection problem applies, if we have pq variables to include

or exclude, we can make 2^{pq} possible model choices and for large pq there is no optimal way to make such a choice. It is argued forcefully in [7] that either multilevel modeling should be limited to situations with a small numbers of variables or it should only be applied in areas in which there is sufficient scientific theory on the basis of which to choose predictors.

Another aspect of variable selection is that we can set some of the random coefficients in $\underline{\delta}_j$ to zero. Thus, the corresponding predictor in X only has a fixed effect, not a random effect. This means that particular row and column of Ω corresponding with that predictor are set to zero. It is frequently useful to use this strategy in a rather extreme way and set the random parts of all regression coefficients, except the intercept, equal to zero. This leads to random intercept models, which have far fewer parameters and are much better conditioned. They are treated in detail in Longford [10].

If we set parts on Ω to zero, we must be careful. In Kreft et al. [8] it is shown that requiring Ω to be diagonal, for instance, destroys the invariance of the results under centering of the variables. Thus, in a model of this form, we need meaningful zero points for the variables, and meaningful zero points are quite rare in social and behavioral applications (see **Centering in Linear Multilevel Models**).

Generalizations

The linear multilevel model can be, and has been, generalized in many different directions. It is based on many highly restrictive assumptions, and by relaxing some or all of these assumptions we get various generalizations.

First, we can relax the interaction structure of $U_j = X_j Z_j$ and look at the multilevel model for general $n_j \times q$ design matrices U_j . Thus, we consider more general models in which some of the predictors have fixed coefficients and some of the predictors have random coefficients. We can write such models, in the two-level case, simply as

$$\underline{y}_j = U_j \beta_j + X_j \underline{\delta}_j + \underline{\epsilon}_j. \quad (6)$$

It is possible, in fact, that there is overlap in the two sets of predictors U_j and X_j , which means that regressions coefficients have both a fixed part and a

random part. Second, we can relax the homoscedasticity assumptions $V(\underline{\epsilon}_j) = \sigma^2 I$. We can introduce σ_j^2 , so that the error variance is different for different second-level units. Or we can allow for more general parametric error structures $V(\underline{\epsilon}_j) = \Sigma_j(\theta)$, for example, by allowing auto-correlation between errors at different time points within the same individual (see **Heteroscedasticity and Complex Variation**).

Third, it is comparatively straightforward to generalize the model to more than two levels. The notation can become somewhat tedious, but when all the necessary substitutions have been made we still have a linear mixed model with nested random effects, and the estimation and data analysis proceed in the same way as in the two-level model.

Fourth, the device of modeling parameter vectors as random is strongly reminiscent of the Bayesian approach to statistics (see **Bayesian Statistics**). The main difference is that in our approach to multilevel analysis we still have the fixed parameters γ , σ^2 and Ω that must be estimated. In a fully Bayesian approach one would replace these fixed parameters by random variables with some prior distribution and one can then compute the posterior distribution of the parameters vectors, which are now all random effects. The Bayesian approach to multilevel modeling (or hierarchical linear modeling) has been explored in many recent publications, especially since the powerful **Markov Chain Monte Carlo** tools became available.

Fifth, we can drop the assumption of linearity and consider nonlinear multilevel models or generalized linear multilevel models (see **Nonlinear Mixed Effects Models; Generalized Linear Mixed Models**). Both are discussed in detail in the basic treatises of Raudenbush and Bryk [13] and Goldstein [5], but discussing them here would take us too far astray. The same is true for models with multivariate outcomes, in which the elements of the vector y_j are themselves vectors, or even matrices. A recent application of multilevel models in this context is analysis of fMRI data [2].

And finally, we can move multilevel analysis from the regression context to the more general framework of latent variable modeling. This leads to multilevel factor analysis and to various multilevel structural equation models. A very complete treatment of current research in that field is in Skrondal and Rabe-Hesketh [16].

Estimation

There is a voluminous literature on estimating multi-level models, or, more generally, mixed linear models [15]. Most methods are based on assuming normality of the random effects and then using **maximum likelihood estimation**. The likelihood function depends on the regression coefficients for the fixed variables and the variances and covariances of the random effects. It is easily minimized, for instance, by alternating minimization over γ for fixed σ^2 and Ω , and then minimization over σ^2 and Ω for fixed γ , until convergence. This is sometimes known as or Iterative Generalized Least Squares IGLS, [4]. It is also possible to treat the random coefficients as missing data and apply the EM algorithm [13] (see **Maximum Likelihood Estimation**), or to apply Newton's method or Fisher Scoring to optimize the likelihood [3, 9] (see **Optimization Methods**).

There is more than one likelihood function we can use. In the early work, the likelihood of the observations was used; later on this was largely replaced by using the likelihood of the least squares residuals. The likelihood of the observations is a function of σ^2 , Ω and γ . The disadvantage of the FIML estimates obtained by maximizing this full information likelihood function is that variance components tend to be biased, in the same way, and for the same reason, why the maximum likelihood of the sample variance is biased. In the case of the sample variance, we correct for the bias of the estimate by maximizing the likelihood of the deviations of the sample mean. In the same way, we can study the likelihood of a set of linear combinations of the observations, where the coefficients of the linear combinations are chosen orthogonal to the X_j . This means that γ disappears from the residual or reduced likelihood, which is now only a function of the variance and covariance components. The resulting REML estimates, originally due to Patterson and Thomson [11], can be computed with small variations of the more classical maximum likelihood algorithms (IGLS, EM, Scoring), because the two types of likelihood functions are closely related.

Of course REML does not give an estimate of the fixed regression coefficients, because the residual likelihood does not depend on γ . This problem is resolved by estimating γ by generalized least squares, using the REML estimates of the variance components. Neither REML nor FIML gives estimates of

the random regression coefficients or random effects. Random variables are not fixed parameters, and consequently they cannot be estimated in the classical sense. What we can estimate is the conditional expectation of the random effects given the data. These conditional expectations can be estimated by plug-in estimates using the REML or FIML estimates of the fixed parameters. They are also known as the *best linear unbiased predictors*, the *empirical Bayes estimates*, or the BLUP's [14] (see **Random Effects in Multivariate Linear Models: Prediction**).

There is a large number of software packages designed specifically for linear multilevel models, although most of them by now also incorporate the generalizations we have discussed in the previous section. The two most popular special purpose packages are HLM, used in Raudenbush and Bryk [13], and MLWin, used in Goldstein [5]. Many of the standard statistical packages, such as SAS, SPSS, Stata, and R now also have multilevel extensions written in their interpreted matrix languages (see **Software for Statistical Analyses**).

School Effectiveness Example

We use school examination data previously analyzed with multilevel methods by Goldstein et al. [6]. Data are collected on 4059 students in 65 schools in Inner London. For each student, we have a normalized exam score (`normexam`) as the outcome variable. Student-level predictors are gender (coded as a dummy `genderM`) and standardized London Reading Test score (`standlrt`). The single school-level predictors we use is school gender (mixed, boys, or girls school, abbreviated as `schgend`). This is a categorical variable, which we code using a boyschool-dummy `schgendboys` and a girlschool-dummy `schgendgirls`.

Our first model is a simple random intercept model, with a single variance component. Only the intercept is random, all other regression coefficients are fixed. The model is

$$\begin{aligned} \text{normexam}_{ij} &= \alpha_j + \text{standlrt}_{ij}\beta_1 \\ &\quad + \text{gender}_{ij}\beta_2 + \epsilon_{ij}, \\ \alpha_j &= \text{schgendboys}_j\gamma_1 \\ &\quad + \text{schgendgirls}_j\gamma_2 + \delta_j \quad (7) \end{aligned}$$

Table 1 Estimates school effectiveness example

source	Random Model	Fixed Model
intercept	−0.00 (0.056)	−0.03 (0.025)
standlrt	0.56 (0.012)	0.56 (0.017)
genderM	−0.17 (0.034)	−0.17 (0.034)
schgendboys	0.18 (0.113)	0.18 (0.043)
schgendgirls	0.16 (0.089)	0.17 (0.033)
ω^2	0.086	—
σ^2	0.563	0.635
ρ	0.132	—

We compare this with the model without a random intercept δ_j .

REML estimation of both models gives the following table of estimates, with standard errors in parentheses. In Table 1 we use ω^2 for the variance of the random intercept, the single element of the matrix Ω in this case.

This is a small example, but it illustrates some basic points. The estimated intraclass correlation ρ is only 0.132 in this case, but the fact that it is nonzero has important consequences. We see that if the random coefficient model holds then the standard errors of the regression coefficient from the fixed model are far too small. In fact, in the fixed model the school variables `schgendboys` and `schgendgirls` are highly significant, while they are not even significant at the 5% level in the random model. We also see that the estimate of σ^2 is higher in the fixed model, which is not surprising because the random model allows for an additional parameter to model the variation. Another important point is that the actual values of the regression coefficients in the fixed and random model are very close. Again, this is not that surprising, because after all in REML the fixed coefficients are estimated with least squares methods as well.

Growth Curve Example

We illustrate repeated measure examples with a small dataset taken from the classical paper by Pothoff and Roy [12]. Distances between pituitary gland and pterygomaxillary fissure were measured using x-rays in $n = 27$ children (16 males and 11 females) at $m = 4$ time points, at ages 8, 10, 12, and 14. Data can be collected in a $n \times m$ matrix Y . We also use a $m \times p$

matrix X of the first $p = 2$ orthogonal polynomials on the m time points.

The first class of models we consider is $\underline{Y} = B\underline{X}' + \underline{E}$ with B a $n \times p$ matrix of regression coefficients, one for each subject, and with $\underline{E}$ the $n \times m$ matrix of disturbances. We suppose the rows of $\underline{E}$ are independent, identically distributed centered normal vectors, with dispersion Σ . Observe that the model here tells us the growth curves are straight lines, not that the deviations from the average growth curves are on a straight line (see **Growth Curve Modeling**).

Within this class of models we can specify various submodels. The most common one supposes that $\Sigma = \sigma^2 I$. Using the orthogonality of the polynomials in X , we find that in this case the regression coefficients are estimated simply by $\hat{B} = YX$. But many other specifications are possible. We can, on the one hand, require Σ to be a scalar, diagonal, or free matrix. And we can, on the other hand, require the regression coefficients to be all the same, the same for all boys and the same for all girls, or free (all different). These are all fixed regression models. The minimum deviances (minus two times the maximized likelihood) are shown in the first three rows of Table 2. In some combinations there are too many parameters. As in other linear models this means the likelihood is unbounded above and the maximum likelihood estimate does not exist [1].

We show the results for the simplest case, with the regression coefficients ‘free’ and the dispersion matrix ‘scalar’. The estimated growth curves are in Figure 1. Boys are solid lines, girls are dashed. The estimated σ^2 is 0.85.

We also give the results for the ‘gender’ regression coefficients and the ‘free’ dispersion matrix. The two regression lines are in Figure 2. The regression line for boys is both higher and steeper than the one for girls. There is much less room in this model to incorporate the variation in the data using the regression coefficients, and thus we expect the

Table 2 Mixed model fit

	B equal	B gender	B free
Σ scalar	307(3)	280(5)	91(55)
Σ diagonal	305(6)	279(8)	−∞(58)
Σ free	233(12)	221(14)	−∞(64)
random	240(6)	229(8)	−∞(58)

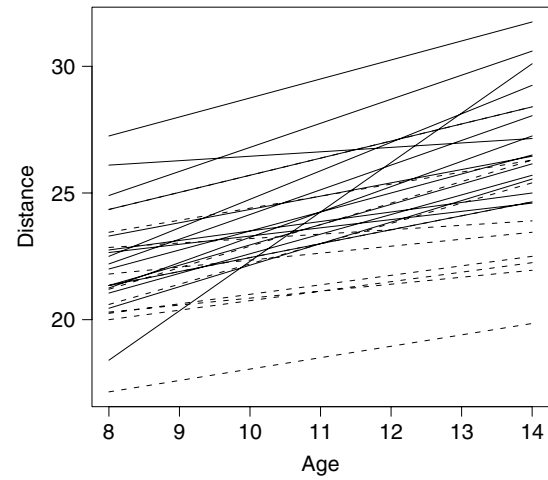


Figure 1 Growth curves for the free/scalar model

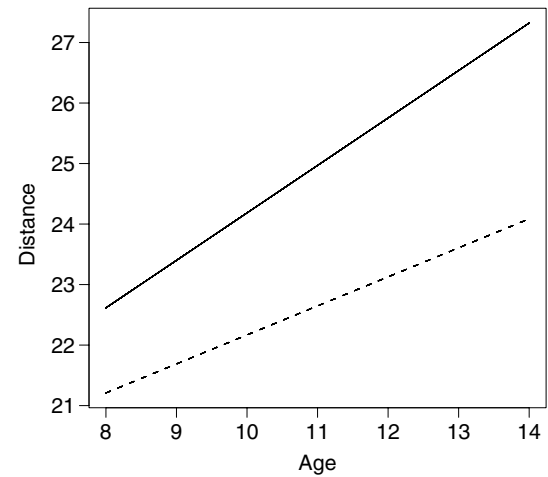


Figure 2 Growth curves for the gender/free model

Table 3 Σ from gender/free model				
	8	10	12	14
Correlations	1.00			
	0.54	1.00		
	0.65	0.56	1.00	
	0.52	0.72	0.73	1.00
Variances	5.12	3.93	5.98	4.62

estimate of the residual variance to be larger. In Table 3 we give the variances and correlations from

the estimated Σ . The estimated correlations between the errors are clearly substantial.

The general problem with fixed effects models in this context is clear from both the figures and the tables. To make models realistic we need a lot of parameters, but if there are many parameters we cannot expect the estimates to be very good. In fact in some cases we have unbounded likelihoods and the estimates we look for do not even exist. Also, it is difficult to make sense of so many parameters at the same time, as Figure 1 shows.

Next consider random coefficient models of the form $\underline{Y} = \underline{B}X' + \underline{E}$, where the rows of $\underline{B}$ are uncorrelated with each other and with all of $\underline{E}$. By writing $\underline{B} = B + \underline{\Delta}$ with $B = E(\underline{B})$ we see that we have a mixed linear model of the form $\underline{Y} = BX' + \underline{\Delta}X' + \underline{E}$. Use Ω for the dispersion of the rows of $\underline{\Delta}$. It seems that we have made our problems actually worse by introducing more parameters. But allowing random variation in the regression coefficients makes the restrictive models for the fixed part more sensible. We fit the ‘equal’ and ‘gender’ versions for the regression coefficients B , together with the ‘scalar’ version of Σ , leaving Ω ‘free’.

Deviances for the random coefficient model are shown in the last row of Table 2. We see a good fit, with a relatively small number of parameters. To get growth curves for the individuals we compute the BLUP, or conditional expectation, $E(\underline{B}|Y)$, which

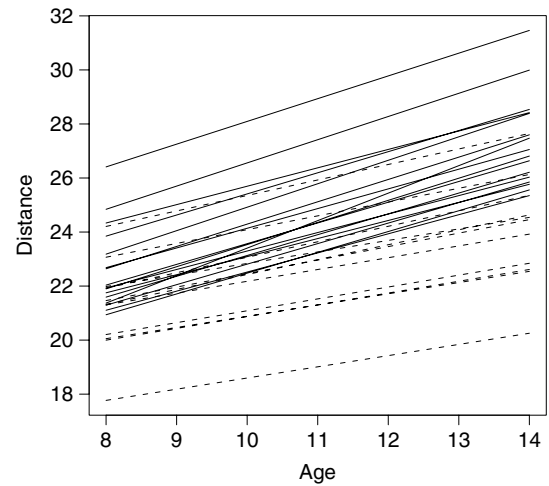


Figure 3 Growth curves for the mixed gender model

turns out to be

$$\begin{aligned} E(\underline{B}|Y) &= \tilde{B}[I - \Omega(\Omega + \sigma^2 I)^{-1}] \\ &\quad + \hat{B}\Omega(\Omega + \sigma^2 I)^{-1}, \end{aligned} \quad (8)$$

where $\tilde{B}$ is the mixed model estimate and $\hat{B} = YX$ is the least squares estimate portrayed in Figure 1. Using the ‘gender’ restriction on the regression coefficients the conditional expectations are plotted in Figure 3.

We see they provide a compromise solution, that shrinks the ordinary least squares estimates in the direction of the ‘gender’ mixed model estimates. We more clearly see the variation of the growth curves for the two genders around the mean gender curve. The estimated σ^2 for this model is 1.72.

References

- [1] Anderson, T.W. (1984). Estimating linear statistical relationships, *The Annals of Statistics* **12**(1), 1–45.
- [2] Beckmann, C.F., Jenkinson, M. & Smith, S.M. (2003). General multilevel linear modeling for group analysis in FMRI, *NeuroImage* **20**, 1052–1063.
- [3] De Leeuw, J. & Kreft, I.G.G. (1986). Random coefficient models for multilevel analysis, *Journal of Educational Statistics* **11**, 57–86.
- [4] Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares, *Biometrika* **73**, 43–56.
- [5] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold Publishing.
- [6] Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Panand, H., Nuttall, D. & Thomas, S. (1993). A multilevel analysis of school examination results, *Oxford Review of Education* **19**, 425–433.
- [7] Kreft, G.G. & de Leeuw, J. (1998). *Introduction to Multilevel Modelling*, Sage Publications, Thousand Oaks.
- [8] Kreft, G.G., De Leeuw, J. & Aiken, L.S. (1995). The effects of different forms of centering in hierarchical linear models, *Multivariate Behavioral Research* **30**, 1–21.
- [9] Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects, *Biometrika* **74**, 817–827.
- [10] Longford, N.T. (1993). *Random Coefficient Models*, Number 11 in Oxford Statistical Science Series, Oxford University Press, Oxford.
- [11] Patterson, H.D. & Thomson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**, 545–554.
- [12] Pothoff, R.F. & Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika* **51**, 313–326.
- [13] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods*, 2nd Edition, Sage Publications, Newbury Park.
- [14] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion), *Statistical Science* **6**, 15–51.
- [15] Searle, S.R., Casella, G. & McCulloch, C.E. (1992). *Variance Components*, Wiley, New York.
- [16] Skrondal, A. & Rabe-Hesketh, S. (2004). Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models, *Interdisciplinary Statistics*, Chapman & Hall.
- [17] Strenio, J.L.F., Weisberg, H.I. & Bryk, A.S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates, *Biometrics* **39**, 71–86.

JAN DE LEEUW

Linear Statistical Models for Causation: A Critical Review

D.A. FREEDMAN

Volume 2, pp. 1061–1073

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Linear Statistical Models for Causation: A Critical Review

Introduction

Regression models are often used to infer causation from association. For instance, Yule [79] showed – or tried to show – that welfare was a cause of poverty. Path models and structural equation models are later refinements of the technique. Besides Yule, examples to be discussed here include Blau and Duncan [12] on stratification, as well as Gibson [28] on the causes of McCarthyism. Strong assumptions are required to infer causation from association by modeling. The assumptions are of two kinds: (a) causal, and (b) statistical. These assumptions will be formulated explicitly, with the help of response schedules in hypothetical experiments. In particular, parameters and error distributions must be stable under intervention. That will be hard to demonstrate in observational settings. Statistical conditions (like independence) are also problematic, and latent variables create further complexities. Causal modeling with path diagrams will be the primary topic. The issues are not simple, so examining them from several perspectives may be helpful. The article ends with a review of the literature and a summary.

Regression Models in Social Science

Legendre [49] and Gauss [27] developed regression to fit data on orbits of astronomical objects. The relevant variables were known from Newtonian mechanics, and so were the functional forms of the equations connecting them. Measurement could be done with great precision, and much was known about the nature of errors in the measurements and in the equations. Furthermore, there was ample opportunity for comparing predictions to reality. By the turn of the century, investigators were using regression on social science data where such conditions did not hold, even to a rough approximation. **Yule** [79] was a pioneer. At the time, paupers in England were supported either inside grim Victorian institutions called *poor-houses* or outside, according to decisions made by

local authorities. Did policy choices affect the number of paupers? To study this question, Yule proposed a regression equation,

$$\Delta \text{Paup} = a + b \times \Delta \text{Out} + c \times \Delta \text{Old} + d \times \Delta \text{Pop} + \text{error}. \quad (1)$$

In this equation,

Δ is percentage change over time,

Paup is the number of Paupers

Out is the out-relief ratio N/D ,

N = number on welfare outside the poor-house,

D = number inside,

Old is the population over 65,

Pop is the population.

Data are from the English Censuses of 1871, 1881, and 1891. There are two Δ 's, one each for 1871–1881 and 1881–1891.

Relief policy was determined separately in each ‘union’, a small geographical area like a parish. At the time, there were about 600 unions, and Yule divides them into four kinds: rural, mixed, urban, metropolitan. There are $4 \times 2 = 8$ equations, one for each type of union and time period. Yule fits each equation to data by least squares. That is, he determines a , b , c , and d by minimizing the sum of squared errors,

$$\sum (\Delta \text{Paup} - a - b \times \Delta \text{Out} - c \times \Delta \text{Old} - d \times \Delta \text{Pop})^2.$$

The sum is taken over all unions of a given type in a given time period – which assumes, in essence, that coefficients are constant within each combination of geography and time. For example, consider the metropolitan unions. Fitting the equation to the data for 1871–1881, Yule gets

$$\Delta \text{Paup} = 13.19 + 0.755 \Delta \text{Out} - 0.022 \Delta \text{Old} - 0.322 \Delta \text{Pop} + \text{error}. \quad (2)$$

For 1881–1891, his equation is

$$\Delta \text{Paup} = 1.36 + 0.324 \Delta \text{Out} + 1.37 \Delta \text{Old} - 0.369 \Delta \text{Pop} + \text{error}. \quad (3)$$

2 Linear Statistical Models for Causation

Table 1 Pauperism, out-relief ratio, proportion of old, population. Ratio of 1881 data to 1871 data, times 100. Metropolitan Unions, England. Yule (79, Table XIX)

	Paup	Out	Old	Pop
Kensington	27	5	104	136
Paddington	47	12	115	111
Fulham	31	21	85	174
Chelsea	64	21	81	124
St. George's	46	18	113	96
Westminster	52	27	105	91
Marylebone	81	36	100	97
St. John, Hampstead	61	39	103	141
St. Pancras	61	35	101	107
Islington	59	35	101	132
Hackney	33	22	91	150
St. Giles'	76	30	103	85
Strand	64	27	97	81
Holborn	79	33	95	93
City	79	64	113	68
Shoreditch	52	21	108	100
Bethnal Green	46	19	102	106
Whitechapel	35	6	93	93
St. George's East	37	6	98	98
Stepney	34	10	87	101
Mile End	43	15	102	113
Poplar	37	20	102	135
St. Saviour's	52	22	100	111
St. Olave's	57	32	102	110
Lambeth	57	38	99	122
Wandsworth	23	18	91	168
Camberwell	30	14	83	168
Greenwich	55	37	94	131
Lewisham	41	24	100	142
Woolwich	76	20	119	110
Croydon	38	29	101	142
West Ham	38	49	86	203

The coefficient of ΔOut being relatively large and positive, Yule concludes that outrelief causes poverty.

Table 1 has the ratio of 1881 data to 1871 data for Pauperism, Out-relief ratio, Proportion of Old, and Population. If we subtract 100 from each entry, column 1 gives ΔPaup in equation (2). Columns 2, 3, 4 give the other variables. For Kensington (the first union in the table),

$$\Delta\text{Out} = 5 - 100 = -95, \Delta\text{Old} = 104 - 100 = 4,$$

$$\Delta\text{Pop} = 136 - 100 = 36.$$

The predicted value for ΔPaup from (2) is therefore

$$13.19 + 0.755 \times (-95) - 0.022 \times 4 \\ - 0.322 \times 36 = -70.$$

The actual value for ΔPaup is -73 , so the error is -3 . Other lines in the table are handled in a similar way. As noted above, coefficients were chosen to minimize the sum of the squared errors.

Quetelet [67] wanted to uncover 'social physics' – the laws of human behavior – by using statistical technique:

'In giving my work the title of Social Physics, I have had no other aim than to collect, in a uniform order, the phenomena affecting man, nearly as physical science brings together the phenomena appertaining to the material world. . . . in a given state of society, resting under the influence of certain causes, regular effects are produced, which oscillate, as it were, around a fixed mean point, without undergoing any sensible alterations.' . . .

'This study. . . has too many attractions – it is connected on too many sides with every branch of science, and all the most interesting questions in philosophy – to be long without zealous observers, who will endeavor to carry it further and further, and bring it more and more to the appearance of a science.'

Yule is using regression to infer the social physics of poverty. But this is not so easily to be done. Confounding is one issue. According to Pigou (a leading welfare economist of Yule's era), parishes with more efficient administrations were building poor-houses and reducing poverty. Efficiency of administration is then a confounder, influencing both the presumed cause and its effect. Economics may be another confounder. Yule occasionally tries to control for this, using the rate of population change as a proxy for economic growth. Generally, however, he pays little attention to economics. The explanation: 'A good deal of time and labour was spent in making trial of this idea, but the results proved unsatisfactory, and finally the measure was abandoned altogether. [p. 253]'

The form of Yule's equation is somewhat arbitrary, and the coefficients are not consistent over time and space. This is not necessarily fatal. However, unless the coefficients have some existence apart from the data, how can they predict the results of interventions that would change the data? The distinction between parameters and estimates runs throughout statistical theory; the discussion of response schedules, below, may sharpen the point.

There are other interpretive problems. At best, Yule has established association. Conditional on the covariates, there is a positive association between

ΔPaup and ΔOut . Is this association causal? If so, which way do the causal arrows point? For instance, a parish may choose not to build poor-houses in response to a short-term increase in the number of paupers. Then pauperism is the cause and outrelief the effect. Likewise, the number of paupers in one area may well be affected by relief policy in neighboring areas. Such issues are not resolved by the data analysis. Instead, answers are assumed *a priori*. Although he was busily parceling out changes in pauperism – so much is due to changes in out-relief ratios, so much to changes in other variables, so much to random effects – Yule was aware of the difficulties. With one deft footnote (number 25), he withdrew all causal claims: ‘Strictly speaking, for “due to” read “associated with”.’

Yule’s approach is strikingly modern, except there is no causal diagram with stars indicating statistical significance. Figure 1 brings him up to date. The arrow from ΔOut to ΔPaup indicates that ΔOut is included in the regression equation that explains ΔPaup . Three asterisks mark a high degree of statistical significance. The idea is that a statistically significant coefficient must differ from zero. Thus, ΔOut has a causal influence on ΔPaup . By contrast, a coefficient that lacks statistical significance is thought to be zero. If so, ΔOld would not exert a causal influence on ΔPaup .

The reasoning is seldom made explicit, and difficulties are frequently overlooked. Statistical assumptions are needed to determine significance from the data. Even if significance can be determined and the null hypothesis rejected or accepted, there is a deeper problem. To make causal inferences, it must be assumed that equations are stable under proposed interventions. Verifying such assumptions – without making the interventions – is problematic. On the other hand, if the coefficients and error terms change when variables are manipulated, the equation has

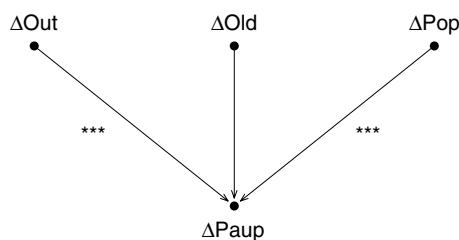


Figure 1 Yule’s model. Metropolitan unions, 1871–1881

only a limited utility for predicting the results of interventions.

Social Stratification

Blau and Duncan [12] are thinking about the stratification process in the United States. According to Marxists of the time, the United States is a highly stratified society. Status is determined by family background, and transmitted through the school system. Blau and Duncan present cross-tabs (in their Chapter 2) to show that the system is far from deterministic, although family background variables do influence status. The United States has a permeable social structure, with many opportunities to succeed or fail. Blau and Duncan go on to develop the path model shown in Figure 2, in order to answer questions like these:

‘how and to what degree do the circumstances of birth condition subsequent status? how does status attained (whether by ascription or achievement) at one stage of the life cycle affect the prospects for a subsequent stage?’

The five variables in the diagram are father’s education and occupation, son’s education, son’s first job, and son’s occupation. Data come from a special supplement to the March 1962 Current Population

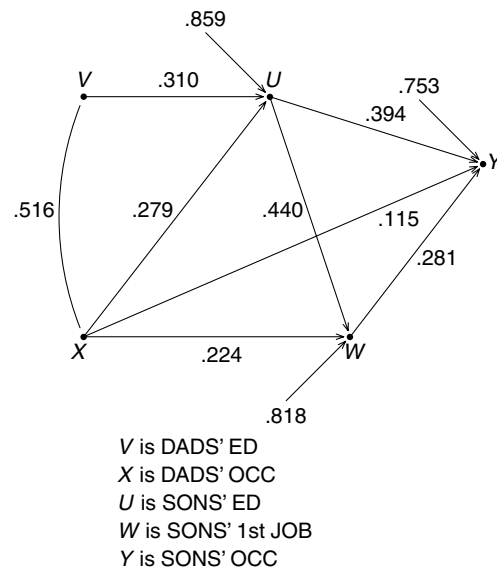


Figure 2 Path model. Stratification, US, 1962

4 Linear Statistical Models for Causation

Survey. The respondents are the sons (age 20–64), who answer questions about current jobs, first jobs, and parents. There are 20 000 respondents. Education is measured on a scale from 0 to 8, where 0 means no schooling, 1 means 1–4 years of schooling, and so forth; 8 means some postgraduate education. Occupation is measured on Duncan’s prestige scale from 0 to 96. The scale takes into account income, education, and raters’ opinions of job prestige. Hucksters are at the bottom of the ladder, with clergy in the middle, and judges at the top.

How is Figure 2 to be read? The diagram unpacks to three regression equations:

$$U = aV + bX + \delta, \quad (4)$$

$$W = cU + dX + \epsilon, \quad (5)$$

$$Y = eU_i + fX + gW + \eta. \quad (6)$$

Parameters are estimated by least squares. Before regressions are run, variables are standardized to have mean 0 and variance 1. That is why no intercepts are needed, and why estimates can be computed from the correlations in Table 2.

In Figure 2, the arrow from V to U indicates a causal link, and V is entered on the right-hand side in the regression equation (4) that explains U . The path coefficient .310 next to the arrow is the estimated coefficient $\hat{a}$ of V . The number .859 on the ‘free arrow’ that points into U is the estimated standard deviation of the error term δ in (4). The other arrows are interpreted in a similar way. The curved line joining V and X indicates association rather than causation: V and X influence each other or are influenced by some common causes, not further analyzed in the diagram. The number on the curved line is just the correlation between V and X (Table 2). There are three equations because three variables in the diagram (U , W , Y) have arrows pointing into them.

The large standard deviations in Figure 2 show the permeability of the social structure. (Since variables are standardized, it is a little theorem that the standard deviations cannot exceed 1.) Even if father’s education and occupation are given, as well as respondent’s education and first job, the variation in status of current job is still large. As social physics, however, the diagram leaves something to be desired. Why linearity? Why are the coefficients the same for everybody? What about variables like intelligence or motivation? And where are the mothers?

The choice of variables and arrows is up to the analyst, as are the directions in which the arrows point. Of course, some choices may fit the data less well, and some may be illogical. If the graph is ‘complete’ – every pair of nodes joined by an arrow – the direction of the arrows is not constrained by the data [22 pp. 138, 142]. Ordering the variables in time may reduce the number of options.

If we are trying to find laws of nature that are stable under intervention, standardizing may be a bad idea, because estimated parameters would depend on irrelevant details of the study design (see below). Generally, the intervention idea gets muddier with standardization. Are means and standard deviations held constant even though individual values are manipulated? On the other hand, standardizing might be sensible if units are meaningful only in comparative terms (e.g., prestige points). Standardizing may also be helpful if the meaning of units changes over time (e.g., years of education), while correlations are stable. With descriptive statistics for one data set, it is really a matter of taste: do you like pounds, kilograms, or standard units? Moreover, all variables are on the same scale after standardization, which makes it easier to compare regression coefficients.

Table 2 Correlation matrix for variables in Blau and Duncan’s path model

		Y Sons’occ	W Sons’1 st job	U Sons’ed	X Dads’occ	V Dads’ed
Y	Sons’occ	1.000	.541	.596	.405	.322
W	Sons’1 st job	.541	1.000	.538	.417	.332
U	Sons’ed	.596	.538	1.000	.438	.453
X	Dads’occ	.405	.417	.438	1.000	.516
V	Dads’ed	.322	.332	.453	.516	1.000

Hooke's Law

According to Hooke's law, stretch is proportional to weight. If weight x is hung on a spring, the length of the spring is $a + bx + \epsilon$, provided x is not too large. (Near the elastic limit of the spring, the physics will be more complicated.) In this equation, a and b are physical constants that depend on the spring not the weights. The parameter a is the length of the spring with no load. The parameter b is the length added to the spring by each additional unit of weight. The ϵ is random measurement error, with the usual assumptions. Experimental verification is a classroom staple.

If we were to standardize, the crucial slope parameter would depend on the weights and the accuracy of the measurements. Let v be the variance of the weights used in the experiment, let σ^2 be the variance of ϵ , and let s^2 be the mean square of the deviations from the fitted regression line. The standardized regression coefficient is

$$\sqrt{\frac{\hat{b}^2 v}{\hat{b}^2 v + s^2}} \approx \sqrt{\frac{b^2 v}{b^2 v + \sigma^2}}, \quad (7)$$

as can be verified by examining the sample covariance matrix. Therefore, the standardized coefficient depends on v and σ^2 , which are features of our measurement procedure not the spring.

Hooke's law is an example where regression is a very useful tool. But the parameter to estimate is b , the unstandardized regression coefficient. It is the unstandardized coefficient that says how the spring will respond when the load is manipulated. If a regression coefficient is stable under interventions, standardizing it is probably not a good idea, because stability gets lost in the shuffle. That is what (7) shows. Also see [4], ([11], p. 451).

Political Repression During the McCarthy Era

Gibson [28] tries to determine the causes of McCarthyism in the United States. Was repression due to the masses or the elites? He argues that elite intolerance is the root cause, the chief piece of evidence being a path model (Figure 3, redrawn from the paper). The dependent variable is a measure of repressive legislation in each state. The

independent variables are mean tolerance scores for each state, derived from the Stouffer survey of masses and elites. The 'masses' are just respondents in a probability sample of the population. 'Elites' include school board presidents, commanders of the American Legion, bar association presidents, labor union leaders. Data on masses were available for 36 states; on elites, for 26 states. The two straight arrows in Figure 3 represent causal links: mass and elite tolerance affect repression. The curved double-headed arrow in Figure 3 represents an association between mass and elite tolerance scores. Each one can influence the other, or both can have some common cause. The association is not analyzed in the diagram.

Gibson computes correlations from the available data, then estimates a standardized regression equation,

$$\begin{aligned} \text{Repression} = & \beta_1 \text{Mass tolerance} \\ & + \beta_2 \text{Elite tolerance} + \delta. \end{aligned} \quad (8)$$

He says, 'Generally, it seems that elites, not masses, were responsible for the repression of the era. . . . The beta for mass opinion is $-.06$; for elite opinion, it is $-.35$ (significant beyond .01)'.

The paper asks an interesting question, and the data analysis has some charm too. However, as social physics, the path model is not convincing. What hypothetical intervention is contemplated? If none, how are regressions going to uncover causal relationships? Why are relationships among the variables supposed to be linear? Signs apart, for example, why does a unit increase in tolerance have the same effect

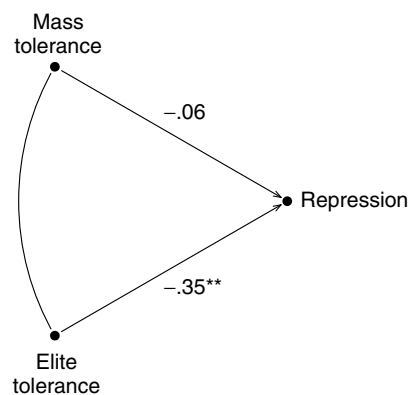


Figure 3 Path model. The causes of McCarthyism

on repression as a unit decrease? Are there other variables in the system? Why are the states statistically independent? Such questions are not addressed in the paper.

McCarthy became a force in national politics around 1950. The turning point came in 1954, with public humiliation in the Army-McCarthy hearings. Censure by the Senate followed in 1957. Gibson scores repressive legislation over the period 1945–1965, long before McCarthy mattered, and long after. The Stouffer survey was done in 1954, when the McCarthy era was ending. The timetable is puzzling.

Even if such issues are set aside, and we grant the statistical model, the difference in path coefficients fails to achieve significance. Gibson finds that $\hat{\beta}_1$ is significant and $\hat{\beta}_2$ is insignificant, but that does not impose much of a constraint on $\hat{\beta}_1 - \hat{\beta}_2$. (The standard error for this difference can be computed from data generously provided in the paper.) Since $\beta_1 = \beta_2$ is a viable hypothesis, the data are not strong enough to distinguish masses from elites.

Inferring Causation by Regression

Path models are often thought to be rigorous statistical engines for inferring causation from association. Statistical techniques can be rigorous, given their assumptions. But the assumptions are usually imposed on the data by the analyst. This is not a rigorous process, and it is rarely made explicit. The assumptions have a causal component as well as a statistical component. It will be easier to proceed in terms of a specific example. In Figure 4, a hypothesized causal relationship between Y and Z is confounded by X . The free arrows leading into Y and Z are omitted.

The diagram describes two hypothetical experiments, and an observational study where the data are collected. The two experiments help to define the assumptions. Furthermore, the usual statistical analysis can be understood as an effort to determine what would happen under those assumptions *if* the experiments were done. Other interpretations of the analysis are not easily to be found. The experiments will now be described.

1. *First hypothetical experiment.* Treatment is applied to a subject, at level x . A response Y is observed,

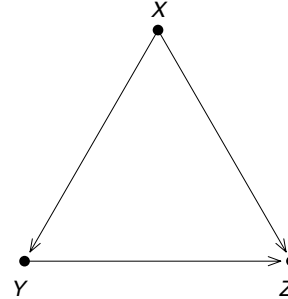


Figure 4 Path model. The relationship between Y and Z is confounded by X . Free arrows leading into Y and Z are not shown

corresponding to the level of treatment. There are two parameters, a and b , that describe the response. With no treatment, the response level for each subject will be a , up to random error. All subjects are assumed to have the same value for a . Each additional unit of treatment adds b to the response. Again, b is the same for all subjects, at all levels of x , by assumption. Thus, if treatment is applied at level x , the response Y is assumed to be

$$a + bx + \text{random error.} \quad (9)$$

For Hooke's law, x is weight and Y is length of a spring under load x . For evaluation of job training programs, x might be hours spent in training and Y might be income during a follow-up period.

2. *Second hypothetical experiment.* In the second experiment, there are two treatments and a response variable Z . There are two treatments because there are two arrows leading into Z ; the treatments are labeled X and Y (Figure 4). Both treatments may be applied to a subject. There are three parameters, c , d , and e . With no treatment, the response level for each subject is taken to be c , up to random error. Each additional unit of treatment #1 adds d to the response. Likewise, each additional unit of treatment #2 adds e to the response. The constancy of parameters across subjects and levels of treatment is an assumption. If the treatments are applied at levels x and y , the response Z is assumed to be

$$c + dx + ey + \text{random error.} \quad (10)$$

Three parameters are needed because it takes three parameters to specify the linear relationship (10),

namely, an intercept and two slopes. Random errors in (9) and (10) are assumed to be independent from subject to subject, with a distribution that is constant across subjects; expectations are zero and variances are finite. The errors in (9) are assumed to be independent of the errors in (10).

The observational study. When using the path model in Figure 4 to analyze data from an observational study, we assume that levels for the variable X are independent of the random errors in the two hypothetical experiments (‘exogeneity’). In effect, we pretend that Nature randomized subjects to levels of X for us, which obviates the need for experimental manipulation. The exogeneity of X has a graphical representation: arrows come out of X , but no arrows lead into X .

We take the descriptions of the two experiments, including the assumptions about the response schedules and the random errors, as background information. In particular, we take it that Nature generates Y as if by substituting X into (9). Nature proceeds to generate Z as if by substituting X and Y – the same Y that has just been generated from X – into (10). In short, (9) and (10) are assumed to be the causal mechanisms that generate the observational data, namely, X , Y , and Z for each subject. The system is ‘recursive’, in the sense that output from (9) is used as input to (10) but there is no feedback from (9) to (8).

Under these assumptions, the parameters a , b can be estimated by regression of Y on X . Likewise, c , d , e can be estimated by regression of Z on X and Y . Moreover, these regression estimates have legitimate causal interpretations. This is because causation is built into the background assumptions, via the response schedules (9) and (10). If causation were not assumed, causation would not be demonstrated by running the regressions.

One point of running the regressions is usually to separate out direct and indirect effects of X on Z . The direct effect is d in (10). If X is increased by one unit with Y held fast, then Z is expected to go up by d units. But this is shorthand for the assumed mechanism in the second experiment. Without the thought experiments described by (9) and (10), how can Y be held constant when X is manipulated? At a more basic level, how would manipulation get into the picture?

Another path-analytic objective is to determine the effect e of Y on Z . If Y is increased by one unit with

X held fast, then Z is expected to go up by e units. (If $e = 0$, then manipulating Y would not affect Z , and Y does not cause Z after all.) Again, the interpretation depends on the thought experiments. Otherwise, how could Y be manipulated and X held fast?

To state the model more carefully, we would index the subjects by a subscript i in the range from 1 to n , the number of subjects. In this notation, X_i is the value of X for subject i . Similarly, Y_i and Z_i are the values of Y and Z for subject i . The level of treatment #1 is denoted by x , and $Y_{i,x}$ is the response for variable Y if treatment at level x is applied to subject i . Similarly, $Z_{i,x,y}$ is the response for variable Z if treatment #1 at level x and treatment #2 at level y are applied to subject i . The response schedules are to be interpreted causally:

- $Y_{i,x}$ is what Y_i would be if X_i were set to x by intervention.
- $Z_{i,x,y}$ is what Z_i would be if X_i were set to x and Y_i were set to y by intervention.

Counterfactual statements are even licensed about the past: $Y_{i,x}$ is what Y_i would have been, if X_i had been set to x . Similar comments apply to $Z_{i,x,y}$.

The diagram unpacks into two equations, which are more precise versions of (9) and (10), with a subscript i for subjects. Greek letters are used for the random error terms.

$$Y_{i,x} = a + bx + \delta_i. \quad (11)$$

$$Z_{i,x,y} = c + dx + ey + \epsilon_i. \quad (12)$$

The parameters a , b , c , d , e and the error terms δ_i , ϵ_i are not observed. The parameters are assumed to be the same for all subjects.

Additional assumptions, which define the statistical component of the model, are imposed on the error terms:

1. δ_i and ϵ_i are independent of each other within each subject i .
2. δ_i and ϵ_i are independent across subjects.
3. The distribution of δ_i is constant across subjects; so is the distribution of ϵ_i . (However, δ_i and ϵ_i need not have the same distribution.)
4. δ_i and ϵ_i have expectation zero and finite variance.
5. The δ ’s and ϵ ’s are independent of the X ’s.

The last is ‘exogeneity’.

8 Linear Statistical Models for Causation

According to the model, Nature determines the response Y_i for subject i by substituting X_i into (10):

$$Y_i = Y_{i,X_i} = a + bX_i + \delta_i. \quad (13)$$

Here, X_i is the value of X for subject i , chosen for us by Nature, as if by randomization. The rest of the response schedule – the $Y_{i,x}$ for other x – is not observed, and therefore stays in the realm of counterfactual hypotheticals. After all, even in an experiment, subject i would be assigned to one level of treatment, foreclosing the possibility of observing the response at other levels.

Similarly, we observe $Z_{i,x,y}$ only for $x = X_i$ and $y = Y_i$. The response for subject i is determined by Nature, as if by substituting X_i and Y_i into (12):

$$Z_i = Z_{i,X_i,Y_i} = c + dX_i + eY_i + \epsilon_i. \quad (14)$$

The rest of the response schedule, $Z_{i,x,y}$ for other x and y , remains unobserved. Economists call the unobserved $Y_{i,x}$ and $Z_{i,x,y}$ ‘potential outcomes’. The model specifies unobservable response schedules, not just regression equations. Notice too that a subject’s responses are determined by levels of treatment for that subject only. Treatments applied to subject j are not relevant to subject i . The response schedules (11) and (12) represent the causal assumptions behind the path diagram.

The conditional expectation of Y given $X = x$ is the average of Y for subjects with $X = x$. The formalism connects two very different ideas of conditional expectation: (a) finding subjects with $X = x$, versus (b) an intervention that sets X to x . The first is something you can actually do with observational data. The second would require manipulation. The model is a compact way of stating the assumptions that are needed to go from observational data to causal inferences.

In econometrics and cognate fields, ‘structural’ equations describe causal relationships. The model gives a clearer meaning to this idea, and to the idea of ‘stability under intervention’. The parameters in Figure 4, for instance, are defined through the response schedules (9) and (10), separately from the data. These parameters are constant across subjects and levels of treatment (by assumption, of course). Parameters are the same in a regime of passive observation and in a regime of active manipulation. Similar assumptions of stability are imposed on the error distributions. In summary, regression equations are

structural, with parameters that are stable under intervention, when the equations derive from response schedules like (11) and (12).

Path models do not infer causation from association. Instead, path models *assume* causation through response schedules, and – using additional statistical assumptions – estimate causal effects from observational data. The statistical assumptions (independence, expectation zero, constant variance) justify estimation by ordinary least squares. With large samples, confidence intervals and significance tests would follow. With small samples, the errors would have to follow a normal distribution in order to justify t Tests.

The box model in Figure 5 illustrates the statistical assumptions. Independent errors with constant distributions are represented as draws made at random with replacement from a box of potential errors [26]. Since the box remains the same from one draw to another, the probability distribution of one draw is the same as the distribution of any other. The distribution is constant. Furthermore, the outcome of one draw cannot affect the distribution of another. That is independence. Verifying the causal assumptions (11) and (12), which are about potential outcomes, is a daunting task. The statistical assumptions present difficulties of their own. Assessing the degree to which the modeling assumptions hold is therefore problematic. The difficulties noted earlier – in Yule on poverty, Blau and Duncan on stratification, Gibson on McCarthyism – are systemic.

Embedded in the formalism is the conditional distribution of Y , if we were to intervene and set the value of X . This conditional distribution is a counterfactual, at least when the study is observational. The conditional distribution answers the question, what would have happened if we had intervened and set X to x , rather than letting Nature take its course?

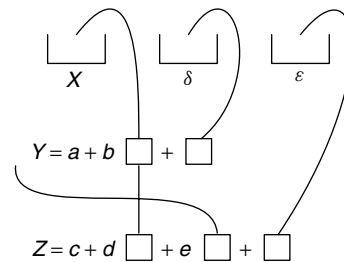


Figure 5 The path diagram as a box model

The idea is best suited to experiments or hypothetical experiments.

There are also nonmanipulationist ideas of causation: the moon causes the tides, earthquakes cause property values to go down, time heals all wounds. Time is not manipulable; neither are earthquakes or the moon. Investigators may hope that regression equations are like laws of motion in classical physics. (If position and momentum are given, you can determine the future of the system and discover what would happen with different initial conditions.) Some other formalism may be needed to make this nonmanipulationist account more precise.

Latent Variables

There is yet another layer of complexity when the variables in the path model remain ‘latent’ – unobserved. It is usually supposed that the manifest variables are related to the latent variables by a series of regression-like equations (‘measurement models’). There are numerous assumptions about error terms, especially when likelihood techniques are used. In effect, latent variables are reconstructed by some version of factor analysis and the path model is fitted to the results. The scale of the latent variables is not usually identifiable, so variables are standardized to have mean 0 and variance 1. Some algorithms will infer the path diagram as well as the latents from the data, but there are additional assumptions that come into play. Anderson [7] provides a rigorous discussion of statistical inference for models with latent variables, given the requisite statistical assumptions. He does not address the connection between the models and the phenomena. Kline [46] is a well-known text. Ullman and Bentler [78] survey recent developments.

A possible conflict in terminology should be mentioned. In psychometrics and cognate fields, ‘structural equation modeling’ (typically, path modeling with latent variables) is sometimes used for causal inference and sometimes to get parsimonious descriptions of covariance matrices. For causal inference, questions of stability are central. If no causal inferences are made, stability under intervention is hardly relevant; nor are underlying equations ‘structural’ in the econometric sense described earlier. The statistical assumptions (independence, distributions of error terms constant across subjects, parametric models for error distributions) would remain on the table.

Literature Review

There is by now an extended critical literature on statistical models, starting perhaps with the exchange between Keynes [44, 45] and Tinbergen [77]. Other familiar citations in the economics literature include Liu [52], Lucas [53], and Sims [71]. Manski [54] returns to the under-identification problem that was posed so sharply by Liu and Sims. In brief, *a priori* exclusion of variables from causal equations can seldom be justified, so there will typically be more parameters than data. Manski suggests methods for bounding quantities that cannot be estimated. Sims’ idea was to use simple, low-dimensional models for policy analysis, instead of complex-high dimensional ones. Leamer [48] discusses the issues created by specification searches, as does Hendry [35]. Heckman [33] traces the development of econometric thought from Haavelmo and Frisch onwards, stressing the role of ‘structural’ or ‘invariant’ parameters, and ‘potential outcomes’. Lucas too was concerned about parameters that changed under intervention. Engle, Hendry, and Richard [17] distinguish several kinds of exogeneity, with different implications for causal inference. Recently, some econometricians have turned to natural experiments for the evaluation of causal theories. These investigators stress the value of careful data collection and data analysis. Angrist and Krueger [8] have a useful survey.

One of the drivers for modeling in economics and other fields is rational choice theory. Therefore, any discussion of empirical foundations must take into account a remarkable series of papers, initiated by Kahneman and Tversky [41], that explores the limits of rational choice theory. These papers are collected in Kahneman, Slovic, and Tversky [40], and in Kahneman and Tversky [43]. The heuristics and biases program has attracted its own critics [29]. That critique is interesting and has some merit. But in the end, the experimental evidence demonstrates severe limits to the power of rational choice theory [42]. If people are trying to maximize expected utility, they generally do not do it very well. Errors are large and repetitive, go in predictable directions, and fall into recognizable categories. Rather than making decisions by optimization – or bounded rationality, or satisficing – people seem to use plausible heuristics that can be identified. If so, rational choice theory is generally not a good basis for justifying empirical models of behavior. Drawing in part on the work

of Kahneman and Tversky, Sen [69] gives a far-reaching critique of rational choice theory. This theory has its place, but also leads to ‘serious descriptive and predictive problems’.

Almost from the beginning, there were critiques of modeling in other social sciences too [64]. Bernert [10] reviews the historical development of causal ideas in sociology. Recently, modeling issues have been much canvassed in sociology. Abbott [2] finds that variables like income and education are too abstract to have much explanatory power, with a broader examination of causal modeling in Abbott [3]. He finds that ‘an unthinking causalism today pervades our journals’; he recommends more emphasis on descriptive work and on middle-range theories. Berk [9] is skeptical about the possibility of inferring causation by modeling, absent a strong theoretical base. Clogg and Haritou [14] review difficulties with regression, noting that you can too easily include endogenous variables as regressors.

Goldthorpe [30, 31, 32] describes several ideas of causation and corresponding methods of statistical proof, with different strengths and weaknesses. Although skeptical of regression, he finds rational choice theory to be promising. He favors use of descriptive statistics to determine social regularities, and statistical models that reflect generative processes. In his view, the manipulationist account of causation is generally inadequate for the social sciences. Hedström and Swedberg [34] present a lively collection of essays by sociologists who are quite skeptical about regression models; rational choice theory also takes its share of criticism. There is an influential book by Lieberman [50], with a follow-up by Lieberman and Lynn [51]. Ní Bhrolcháin [60] has some particularly forceful examples to illustrate the limits of modeling. Sobel [72] reviews the literature on social stratification, concluding that ‘the usual modeling strategies are in need of serious change’. Also see Sobel [73].

Meehl [57] reports the views of an empirical psychologist. Also see Meehl [56], with data showing the advantage of using regression to make predictions, rather than experts. Meehl and Waller [58] discuss the choice between two similar path models, viewed as reasonable approximations to some underlying causal structure, but do not reach the critical question – how to assess the adequacy of the approximation. Steiger [75] has a critical review of structural equation models. Larzalere and Kuhn [47] offer a more

general discussion of difficulties with causal inference by purely statistical methods. Abelson [1] has an interesting viewpoint on the use of statistics in psychology.

There is a well-known book on the logic of causal inference, by Cook and Campbell [15]. Also see Shadish, Cook, and Campbell [70], which has among other things a useful discussion of manipulationist versus nonmanipulationist ideas of causation. In political science, Duncan [16] is far more skeptical about modeling than Blau and Duncan [12]. Achen [5, 6] provides a spirited and reasoned defense of the models. Brady and Collier [13] compare regression methods with case studies; invariance is discussed under the rubric of causal homogeneity.

Recently, strong claims have been made for non-linear methods that elicit the model from the data and control for unobserved confounders [63, 74]. However, the track record is not encouraging [22, 24, 25, 39]. Cites from other perspectives include [55, 61, 62], as well as [18, 19, 20, 21, 23].

The statistical model for causation was proposed by Neyman [59]. It has been rediscovered many times since: see, for instance, [36, Section 9.4]. The setup is often called ‘Rubin’s model’, but that simply mistakes the history. See the comments by Dabrowska and Speed on their translation of Neyman [59], with a response by Rubin; compare to Rubin [68] and Holland [37]. Holland [37, 38] explains the setup with a super-population model to account for the randomness, rather than individualized error terms. Error terms are often described as the overall effects of factors omitted from the equation. But this description introduces difficulties of its own, as shown by Pratt and Schlaifer [65, 66]. Stone [76] presents a super-population model with some observed covariates and some unobserved. Formal extensions to observational studies – in effect, assuming these studies are experiments after suitable controls have been introduced – are discussed by Holland and Rubin among others.

Conclusion

Causal inferences can be drawn from nonexperimental data. However, no mechanical rules can be laid down for the activity. Since Hume, that is almost a truism. Instead, causal inference seems to require an enormous investment of skill, intelligence, and hard work. Many convergent lines of evidence must be

developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested. Before anything else, the right question needs to be framed. Naturally, there is a desire to substitute intellectual capital for labor. That is why investigators try to base causal inference on statistical models. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. The models themselves demand critical scrutiny. Mathematical equations are used to adjust for confounding and other sources of bias. These equations may appear formidably precise, but they typically derive from many somewhat arbitrary choices. Which variables to enter in the regression? What functional form to use? What assumptions to make about parameters and error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.

Acknowledgments

Richard Berk, Persi Diaconis, Michael Finkelstein, Paul Humphreys, Roger Purves, and Philip Stark made useful comments. This paper draws on Freedman [19, 20, 22–24]. Figure 1 appeared in Freedman [21, 24]; figure 2 is redrawn from Blau and Duncan [12]; figure 3, from Gibson [28], also see Freedman [20].

References

- [1] Abelson, R. (1995). *Statistics as Principled Argument*, Lawrence Erlbaum Associates, Hillsdale.
- [2] Abbott, A. (1997). Of time and space: the contemporary relevance of the Chicago school, *Social Forces* **75**, 1149–1182.
- [3] Abbott, A. (1998). The Causal Devolution, *Sociological Methods and Research* **27**, 148–181.
- [4] Achen, C. (1977). Measuring representation: perils of the correlation coefficient, *American Journal of Political Science* **21**, 805–815.
- [5] Achen, C. (1982). *Interpreting and Using Regression*, Sage Publications.
- [6] Achen, C. (1986). *The Statistical Analysis of Quasi-Experiments*, University of California Press, Berkeley.
- [7] Anderson, T.W. (1984). Estimating linear statistical relationships, *Annals of Statistics* **12**, 1–45.
- [8] Angrist, J.D. & Krueger, A.K. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments, *Journal of Business and Economic Statistics* **19**, 2–16.
- [9] Berk, R.A. (2003). *Regression Analysis: A Constructive Critique*, Sage Publications, Newbury Park.
- [10] Bernert, C. (1983). The career of causal analysis in American sociology, *British Journal of Sociology* **34**, 230–254.
- [11] Blalock, H.M. (1989). The real and unrealized contributions of quantitative sociology, *American Sociological Review* **54**, 447–460.
- [12] Blau, P.M. & Duncan O.D. (1967). *The American Occupational Structure*, Wiley. Reissued by the Free Press, 1978. Data collection described on page 13, coding of education on pages 165–66, coding of status on pages 115–27, correlations and path diagram on pages 169–170.
- [13] Brady, H. & Collier, D., eds (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Rowman & Littlefield Publishers.
- [14] Clogg, C.C. & Haritou, A. (1997). The regression method of causal inference and a dilemma confronting this method, in *Causality in Crisis*, V. McKim & S. Turner, eds, University of Notre Dame Press, pp. 83–112.
- [15] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Houghton Mifflin, Boston.
- [16] Duncan, O.D. (1984). *Notes on Social Measurement*, Russell Sage, New York.
- [17] Engle, R.F., Hendry, D.F. & Richard, J.F. (1983). Exogeneity, *Econometrica* **51**, 277–304.
- [18] Freedman, D.A. (1985). Statistics and the scientific method, in *Cohort Analysis in Social Research: Beyond the Identification Problem*, W.M. Mason & S.E. Fienberg, eds, Springer-Verlag, New York, pp. 343–390, (with discussion).
- [19] Freedman, D.A. (1987). As others see US: a case study in path analysis, *Journal of Educational Statistics* **12**, 101–223, (with discussion), Reprinted in *The Role of Models in Nonexperimental Social Science*, J. Shaffer, ed., AERA/ASA, Washington, 1992, pp. 3–125.
- [20] Freedman, D.A. (1991). Statistical models and shoe leather, in *Sociological Methodology 1991*, Chap. 10, Peter Marsden, ed., American Sociological Association, Washington, (with discussion).
- [21] Freedman, D.A. (1995). Some issues in the foundation of statistics, *Foundations of Science* **1**, 19–83, (with discussion), Reprinted in *Some Issues in the Foundation of Statistics*, B.C. van Fraassen, ed., (1997). Kluwer, Dordrecht, pp. 19–83 (with discussion).
- [22] Freedman, D.A. (1997). From association to causation via regression, in *Causality in Crisis? V. McKim & S. Turner*, eds, University of Notre Dame Press, South Bend, pp. 113–182, with discussion. Reprinted in *Advances in Applied Mathematics* **18**, 59–110.

- [23] Freedman, D.A. (1999). From association to causation: some remarks on the history of statistics, *Statistical Science* **14**, 243–258, Reprinted in *Journal de la Société Française de Statistique* **140**, (1999). 5–32, and in *Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century*, J. Panaretos, ed., (2003). Lawrence Erlbaum, pp. 45–71.
- [24] Freedman, D.A. (2004). On specifying graphical models for causation, and the identification problem, *Evaluation Review* **26**, 267–293.
- [25] Freedman, D.A. & Humphreys, P. (1999). Are there algorithms that discover causal structure? *Synthese* **121**, 29–54.
- [26] Freedman, D.A., Pisani, R. & Purves, R.A. (1998). *Statistics*, 3rd Edition, W. W. Norton, New York.
- [27] Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium*, Perthes et Besser, Hamburg, Reprinted in 1963 by Dover, New York.
- [28] Gibson, J.L. (1988). Political intolerance and political repression during the McCarthy red scare, *APSR* **82**, 511–529.
- [29] Gigerenzer, G. (1996). On narrow norms and vague heuristics, *Psychological Review* **103**, 592–596.
- [30] Goldthorpe, J.H. (1998). *Causation, Statistics and Sociology*, Twenty-ninth Geary Lecture, Nuffield College, Oxford. Published by the Economic and Social Research Institute, Dublin.
- [31] Goldthorpe, J.H. (2000). *On Sociology: Numbers, Narratives, and Integration of Research and Theory*, Oxford University Press.
- [32] Goldthorpe, J.H. (2001). Causation, Statistics, and Sociology, *European Sociological Review* **17**, 1–20.
- [33] Heckman, J.J. (2000). Causal parameters and policy analysis in economics: a twentieth century retrospective, *The Quarterly Journal of Economics* **CVX**, 45–97.
- [34] Hedström, P. & Swedberg, R., eds (1998). *Social Mechanisms*, Cambridge University Press.
- [35] Hendry, D.F. (1993). *Econometrics—Alchemy or Science?* Blackwell, Oxford.
- [36] Hodges Jr, J.L. & Lehmann, E. (1964). *Basic Concepts of Probability and Statistics*, Holden-Day, San Francisco.
- [37] Holland, P. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **8**, 945–960.
- [38] Holland, P. (1988). Causal inference, path analysis, and recursive structural equation models, in *Sociological Methodology 1988*, Chap. 13, C. Clogg, ed., American Sociological Association, Washington.
- [39] Humphreys, P. & Freedman, D.A. (1996). The grand leap, *British Journal for the Philosophy of Science* **47**, 113–123.
- [40] Kahneman, D., Slovic, P., and Tversky, A., eds (1982). *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press.
- [41] Kahneman, D. & Tversky, A. (1974). Judgment under uncertainty: heuristics and bias, *Science* **185**, 1124–1131.
- [42] Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions, *Psychological Review* **103**, 582–591.
- [43] Kahneman, D. & Tversky, A., eds (2000). *Choices, Values, and Frames*, Cambridge University Press.
- [44] Keynes, J.M. (1939). Professor Tinbergen's method, *The Economic Journal* **49**, 558–570.
- [45] Keynes, J.M. (1940). Comment on Tinbergen's response, *The Economic Journal* **50**, 154–156.
- [46] Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, Guilford Press, New York.
- [47] Larzalore, R.E. & Kuhn, B.R. (2004). The intervention selection bias: an underrecognized confound in intervention research, *Psychological Bulletin* **130**, 289–303.
- [48] Leamer, E. (1978). *Specification Searches*, John Wiley, New York.
- [49] Legendre, A.M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier, Paris. Reprinted in 1959 by Dover, New York.
- [50] Lieberman, S. (1985). *Making it Count*, University of California Press, Berkeley.
- [51] Lieberman, S. & Lynn, F.B. (2002). Barking up the wrong branch: alternative to the current model of sociological science, *Annual Review of Sociology* **28**, 1–19.
- [52] Liu, T.C. (1960). Under-identification, structural estimation, and forecasting, *Econometrica* **28**, 855–865.
- [53] Lucas Jr, R.E. (1976). Econometric policy evaluation: a critique, in K. Brunner & A. Meltzer, eds, *The Phillips Curve and Labor Markets*, Vol. 1 of the Carnegie-Rochester Conferences on Public Policy, supplementary series to the Journal of Monetary Economics, North-Holland, Amsterdam, pp. 19–64. (With discussion.)
- [54] Manski, C.F. (1995). *Identification Problems in the Social Sciences*, Harvard University Press.
- [55] McKim, V. & Turner, S., eds (1997). *Causality in Crisis? Proceedings of the Notre Dame Conference on Causality*, University of Notre Dame Press.
- [56] Meehl, P.E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, University of Minnesota Press, Minneapolis.
- [57] Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology, *Journal of Consulting and Clinical Psychology* **46**, 806–834.
- [58] Meehl, P.E. & Waller, N.G. (2002). The path analysis controversy: a new statistical approach to strong appraisal of verisimilitude, *Psychological Methods* **7**, 283–337. (with discussion).
- [59] Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes, *Roczniki Nauk Rolniczki* **10**, 1–51, in Polish. English translation by Dabrowska, D. & Speed, T. (1990). *Statistical Science* **5**, 463–480, with discussion.
- [60] Ní Bhrolcháin, M. (2001). Divorce effects and causality in the social sciences, *European Sociological Review* **17**, 33–57.
- [61] Oakes, M.W. (1986). *Statistical Inference*, Epidemiology Resources, Chestnut Hill.
- [62] Pearl, J. (1995). Causal diagrams for empirical research, *Biometrika* **82**, 669–710, with discussion.

-
- [63] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press.
 - [64] Platt, J. (1996). *A History of Sociological Research Methods in America*, Cambridge University Press.
 - [65] Pratt, J. & Schlaifer, R. (1984). On the nature and discovery of structure, *Journal of the American Statistical Association* **79**, 9–21.
 - [66] Pratt, J. & Schlaifer, R. (1988). On the interpretation and observation of laws, *Journal of Econometrics* **39**, 23–52.
 - [67] Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, Ou essai de physique sociale*, Bachelier, Paris.
 - [68] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.
 - [69] Sen, A.K. (2002). *Rationality and Freedom*, Harvard University Press.
 - [70] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
 - [71] Sims, C.A. (1980). Macroeconomics and reality, *Econometrica* **48**, 1–47.
 - [72] Sobel, M.E. (1998). Causal inference in statistical models of the process of socioeconomic achievement—a case study, *Sociological Methods & Research* **27**, 318–348.
 - [73] Sobel, M.E. (2000). Causal inference in the social sciences, *Journal of the American Statistical Association* **95**, 647–651.
 - [74] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, 2nd Edition 2000, *Springer Lecture Notes in Statistics*, Vol. 81, Springer-Verlag, MIT Press, New York.
 - [75] Steiger, J.H. (2001). Driving fast in reverse, *Journal of the American Statistical Association* **96**, 331–338.
 - [76] Stone, R. (1993). The assumptions on which causal inferences rest, *Journal of the Royal Statistical Society Series B* **55**, 455–466.
 - [77] Tinbergen, J. (1940). Reply to Keynes, *The Economic Journal* **50**, 141–154.
 - [78] Ullman, J.B. and Bentler, P.M. (2003). Structural equation modeling, in I.B. Weiner, J.A. Schinka & W.F. Velicer, eds, *Handbook of Psychology. Volume 2: Research Methods in Psychology*, Wiley, Hoboken, pp. 607–634.
 - [79] Yule, G.U. (1899). An investigation into the causes of changes in Pauperism in England, chiefly during the last two intercensal decades, *Journal of the Royal Statistical Society* **62**, 249–295.

D.A. FREEDMAN

Linkage Analysis

P. SHAM

Volume 2, pp. 1073–1075

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Linkage Analysis

Genetic linkage refers to the nonindependent segregation of the alleles at genetic loci close to one another on the same chromosome. Mendel's law of segregation states that an individual with heterozygous genotype (Aa) has equal probability of transmitting either allele (A or a) to an offspring. The same is true of any other locus with alleles B and b. Under Mendel's second law, that of independent assortment, the probabilities of transmitting the four possible combinations of alleles (AB, Ab, aB, ab) are all equal, namely *one-quarter*. This law is, however, only true for pairs of loci that are on separate chromosomes. For two loci that are on the same chromosome (known technically as syntenic), the probabilities of the four gametic classes (AB, Ab, aB, ab) are not equal, with an excess of the same allelic combinations as those that were transmitted to the individual from his or her parents. In other words, if the individual received the allelic combination AB from one parent and ab from the other, then he or she will transmit these same combinations with greater probabilities than the others (i.e., Ab and aB). The former allelic combinations are known as parental types and the latter, recombinants. The strength of genetic linkage between two loci is measured by the recombination fraction, defined as the probability that a recombinant of the two loci is transmitted to an offspring. Recombination fraction ranges from 0 (complete linkage) to 0.5 (complete absence of linkage).

Recombinant gametes of two syntenic loci are generated by the crossing-over of homologous chromosomes at certain semirandom locations during meiosis. The smaller the distance between two syntenic loci, the less likely that they will be separated by crossing-over, and therefore the smaller the recombination fraction. A recombination of 0.01 corresponds approximately to a genetic map distance of 1 centiMorgan (cM). The crossing-over rate varies between males and females, and for different chromosomal regions, but on average a genetic distance of 1 cM corresponds approximately to a physical distance of one million DNA base pairs. The total genetic length of the human genome is approximately 3500 cM.

Mapping Disease Genes by Linkage Analysis

For many decades, linkage analysis was restricted to Mendelian phenotypes such as the ABO blood groups and HLA antigens. Recent developments in molecular genetics have enabled nonfunctional polymorphisms to be detected and measured. Standard sets of such genetic markers, evenly spaced throughout the entire genome, have been developed for systematic linkage analysis to localize genetic variants that increase the risk of disease. This is a particularly attractive method of mapping the genes for diseases since no knowledge of the pathophysiology is required. For this reason, the use of linkage analysis to map disease genes is also called *positional cloning*.

Linkage Analysis for Mendelian Diseases

Linkage analysis in humans presents interesting statistical challenges. For Mendelian diseases, the challenges are those of variable pedigree structure and size, and the common occurrence of missing data. The standard method of analysis involves calculating the likelihood with respect to the recombination fraction between disease and marker loci, or the map position of the disease locus in relation to a set of marker loci, while the disease model is assumed known (e.g., dominant or recessive). Traditionally, the strength of evidence for linkage is summarized as a lod score, defined as the common (i.e., base10) logarithm of the ratio of the likelihood given a certain recombination fraction between marker and disease loci to that under no linkage. The likelihood calculations were traditionally accomplished by use of the Elston–Stewart Algorithm, implemented in linkage analysis programs such as LINKAGE and VITESSE. A lod score of 3 or more is conventionally regarded as significant evidence of linkage. For Mendelian disorders, 98% of reports of linkage that meet this criterion have been subsequently confirmed. Linkage analysis has successfully localized and identified the genes for hundreds of Mendelian disorders.

In multipoint linkage analysis, the likelihood (*see Maximum Likelihood Estimation*) is evaluated for different chromosomal locations of the disease locus relative to a fixed set of genetic markers, and lod scores are defined as the common (i.e., base10)

2 Linkage Analysis

logarithm of the ratio of the likelihood given certain positions of the disease locus on the map to that at an unlinked location. Multipoint calculations present serious computational difficulties for the Elston–Stewart Algorithm, but for modest size, pedigrees are feasible using the Lander–Green, implemented for example in GENEHUNTER, ALLEGRO, and MERLIN.

In both two-point and multipoint lod score analyses, it is standard to assume a generalized single locus model, in which the parameters, namely, the frequencies of the disease and normal alleles (A and a), and the probabilities of disease given the three genotypes (AA , Aa , and aa) are specified by the user. For example, for a rare dominant condition, the allele frequency of the disease allele (A) is set to be a low value, while the probabilities of disease given the AA , Aa , aa genotypes are set at close to 1, 1, and 0, respectively. Similarly, for a recessive condition the allele frequency of the disease allele (A) is set to be a moderate value, the probabilities of disease given the AA , Aa , aa genotypes are set at close to 1, 0, and 0, respectively. The need to assume a generalized single locus model and to specify the parameters of such a model has led to the terms model-based or parametric linkage analysis.

Locus heterogeneity in linkage analysis refers to the situation in which the mode of inheritance, but not the actual disease locus, is the same across different pedigrees. In other words, there are multiple disease loci that are indistinguishable from each other both in terms of manifestations at the individual level and in the pattern of familial transmission. Under these circumstances, the power to detect linkage is much diminished, even with lod scores modified to take account of locus heterogeneity (hlod), especially for samples consisting of small pedigrees. Because of this, linkage analysis has the greatest chance of success when carried out on large pedigrees with multiple affected members.

Linkage Analysis for Complex Traits

For common diseases that do not show a simple Mendelian pattern of inheritance and are therefore likely to be the result of multiple genetic and environmental factors, linkage analysis is a difficult task. For such diseases, we typically would have an idea

of the overall importance of genetic factors (i.e., **heritability**) but no detailed knowledge of genetic architecture in terms of the number of vulnerability genes or the magnitude of their effects. There are two major approaches to the linkage analysis of such complex diseases. The first is to adopt a lod score approach, but modified to allow for a number of more or less realistic models for the genotype–phenotype relationship, and to adjust the largest lod score over these models for multiple testing. The second approach is ‘model-free’ in the sense that a disease model does not have to be specified for the analysis. Instead, the analysis proceeds by defining some measure of allele sharing between individuals in a pedigree, and relating the extent of allele sharing to phenotypic similarity.

One popular version of model-free linkage analysis is the affected sib-pair (ASP) method, which is based on the detection of excessive allele sharing at a marker locus for a sample of sibling pairs where both members are affected by the disorder. The usual definition of allele sharing in model-free linkage analysis is ‘identity-by-descent’, which refers to alleles that are descended from (and are therefore replicates of) a single ancestral allele in a recent common ancestor. Algorithms for estimating the extent of local IBD from marker genotype data have been developed and implemented in programs such as ERPA, GENEHUNTER, and MERLIN. Generalizations of the ASP method to other family structures with multiple affected individuals have resulted in nonparametric linkage (NPL) statistics and their improved likelihood-based versions developed by Kong and Cox.

Methods of linkage analysis have been developed also for quantitative traits (e.g., blood pressure, body mass index). A particularly simple method is based on a regression of phenotypic similarity on allele sharing, developed by Haseman and Elston. A more sophisticated approach is based on a variance components model, in which a component of variance is specified to have covariance between relatives that is proportional to the extent of allele sharing between the relatives. However, the Haseman–Elston approach has been modified so that it can be applied to general pedigrees selected to contain individuals with extreme trait values.

Regardless of the statistical method used for the linkage analysis of complex traits, there are two major inherent limitations of the approach. The

first is that the sample sizes required to detect a locus with a small effect size are very large, potentially many thousands of families. The second is low resolving power, in that the region that shows linkage is typically very broad, covering a region with potentially hundreds of genes. For these

reasons, linkage is usually combined with association strategy in the search for the genetic determinants of multigenic diseases.

P. SHAM

Logistic Regression

RONALD CHRISTENSEN

Volume 2, pp. 1076–1082

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Logistic Regression

Logistic regression is a method for predicting the outcomes of ‘either-or’ trials. Either-or trials occur frequently in research. A person responds appropriately to a drug or does not; the dose of the drug may affect the outcome. A person may support a political party or not; the response may be related to their income. A person may have a heart attack in a 10-year period; the response may be related to age, weight, blood pressure, or cholesterol. We discuss basic ideas of modeling in the section titled ‘Basic Ideas’ and basic ideas of data analysis in the section titled ‘Fundamental Data Analysis’. Subsequent sections examine model testing, variable selection, outliers and influential observations, methods for testing lack of fit, exact conditional inference, random effects, and Bayesian analysis.

Basic Ideas

A binary response is one with two outcomes. Denote these as $y = 1$ indicating ‘success’ and $y = 0$ indicating ‘failure’. Let p denote the probability of getting the response $y = 1$, so

$$\Pr[y = 1] = p, \quad \Pr[y = 0] = 1 - p. \quad (1)$$

Let predictor variables such as age, weight, blood pressure, and cholesterol be denoted $x_1, x_2, \dots, x_{k-1}$. A logistic regression model is a method for relating the probabilities to the predictor variables. Specifically, logistic regression is a **linear model** for the logarithm of the odds of success. The odds of success are $p/(1 - p)$ and a linear model for the log odds is

$$\log \left[\frac{p}{1 - p} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1}. \quad (2)$$

Here the β_j s are unknown regression coefficients. For simplicity of notation, let the log odds be

$$\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1}. \quad (3)$$

It follows that

$$p = \frac{e^\eta}{1 + e^\eta}, \quad 1 - p = \frac{1}{1 + e^\eta}. \quad (4)$$

The transformation that changes η into p is known as the ‘logistic’ transformation, hence the name logistic regression. The transformation $\log[p/(1 - p)] = \eta$ is known as the ‘logit’ transformation. Logit models are the same thing as logistic models. In fact, it is impossible to perform logistic regression without using both the logistic transformation and the logit transformation. Historically, ‘logistic regression’ described models for continuous predictor variables $x_1, x_2, \dots, x_{k-1}$ such as age and weight, while ‘logit models’ were used for categorical predictor variables such as sex, race, and alternative medical treatments. Such distinctions are rarely made anymore.

Typical data consist of independent observations on, say, n individuals. The data are a collection $(y_i, x_{i1}, \dots, x_{i,k-1})$ for $i = 1, \dots, n$ with y_i being 1 or 0 for the i th individual and x_{ij} being the value of the j th predictor variable on the i th individual. For example, Christensen [3] uses an example from the Los Angeles Heart Study (LAHS) that involves $n = 200$ men: y_i is 1 if an individual had a coronary incident in the previous 10 years; $k - 1 = 6$ with x_{i1} = age, x_{i2} = systolic blood pressure, x_{i3} = diastolic blood pressure, x_{i4} = cholesterol, x_{i5} = height, and x_{i6} = weight for a particular individual. The specific logistic regression model is

$$\log \left[\frac{p_i}{1 - p_i} \right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}. \quad (5)$$

Note that the model involves $k = 7$ unknown β_j parameters.

There are two ways of analyzing logistic regression models: frequentist methods and Bayesian methods. Historically, logistic regression was developed after standard regression (see **Multiple Linear Regression**) and has been taught to people who already know standard regression. The methods of analysis are similar to those for standard regression.

Fundamental Data Analysis

Standard methods for frequentist analysis depend on the assumption that the sample size n is large relative to the number of β_j parameters in the model, that is, k . The usual results of the analysis are estimates of the β_j s, say $\hat{\beta}_j$ s, and standard errors for the $\hat{\beta}_j$ s, say

2 Logistic Regression

$SE(\hat{\beta}_j)$ s. In addition, a likelihood ratio test statistic, also known as a *deviance* (D), and *degrees of freedom* (df) for the deviance are given. The degrees of freedom for the deviance are $n - k$. (Some computer programs give alternative versions of the deviance that are less useful. This will be discussed later but can be identified by the fact that the degrees of freedom are less than $n - k$.)

The β_j s are estimated by choosing values that maximize the *likelihood function*,

$$L(\beta_0, \dots, \beta_{k-1}) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (6)$$

wherein it is understood that the p_i s depend on the β_j s through the logistic regression model. Such estimates are called **maximum likelihood estimates**. The deviance can be taken as

$$D = -2 \log[L(\hat{\beta}_0, \dots, \hat{\beta}_{k-1})]. \quad (7)$$

For example, in the LAHS data

Variable	$\hat{\beta}_j$	$SE(\hat{\beta}_j)$
Intercept	-4.5173	7.451
x_1	0.04590	0.02344
x_2	0.00686	0.02013
x_3	-0.00694	0.03821
x_4	0.00631	0.00362
x_5	-0.07400	0.1058
x_6	0.02014	0.00984
$D = 134.9,$		$df = 193$

Given this information, a number of results can be obtained. For example, β_6 is the regression coefficient for weight. A 95% **confidence interval** for β_6 has endpoints

$$\hat{\beta}_6 \pm 1.96 SE(\hat{\beta}_6) \quad (8)$$

and is (0.00085, 0.03943). The value 1.96 is the 97.5% point of a standard normal distribution. To test $H_0: \beta_6 = 0$, one looks at

$$\frac{\hat{\beta}_6 - 0}{SE(\hat{\beta}_6)} = \frac{0.02014 - 0}{0.00984} = 2.05 \quad (9)$$

The P value for this test is .040, which is the probability that a standard normal random variable

is greater than 2.05 or less than -2.05. To test $H_0: \beta_6 = 0.01$, one looks at

$$\frac{\hat{\beta}_6 - 0.01}{SE(\hat{\beta}_6)} = \frac{0.02014 - 0.01}{0.00984} = 1.03 \quad (10)$$

and obtains a P value by comparing 1.03 to a standard normal distribution. The use of the standard normal distribution is an approximation based on having n much larger than k .

Perhaps the most useful things to estimate in logistic regression are probabilities. The estimated log odds are

$$\log \left[\frac{\hat{p}}{1 - \hat{p}} \right] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6. \quad (11)$$

For a 60-year-old man with blood pressure of 140 over 90, a cholesterol reading of 200, who is 69 inches tall and weighs 200 pounds, the estimated log odds of a coronary incident are

$$\begin{aligned} \log \left[\frac{\hat{p}}{1 - \hat{p}} \right] &= -4.5173 + .04590(60) \\ &+ .00686(140) - .00694(90) + .00631(200) \\ &- 0.07400(69) + 0.02014(200) = -1.2435. \end{aligned} \quad (12)$$

The probability of a coronary incident is estimated as

$$\hat{p} = \frac{e^{-1.2435}}{1 + e^{-1.2435}} = .224. \quad (13)$$

To see what the model says about the effect of age (x_1) and cholesterol (x_4), one might plot the estimated probability of a coronary incident as a function of age for people with blood pressures, height, and weight of $x_2 = 140$, $x_3 = 90$, $x_5 = 69$, $x_6 = 200$, and, say, both cholesterol $x_4 = 200$ and $x_4 = 300$. Unfortunately, while confidence intervals for this p can be computed without much difficulty, they are not readily available from many computer programs.

Testing Models

The deviance D and its degrees of freedom are useful for comparing alternative logistic regression models.

The actual number reported in the example, $D = 134.9$ with $193df$, is of little use by itself without

another model to compare it to. The value $D = 134.9$ is found by comparing the 7 variable model to a model with $n = 200$ parameters while we only have $n = 200$ observations. Clearly, 200 observations is not a large sample relative to a model with 200 parameters, hence large sample theory does not apply and we have no way to evaluate whether $D = 134.9$ is an appropriate number.

What we can do is compare the 6 predictor variable model (full model) with $k = 7$ to a smaller model (reduced model) involving, say, only age and weight,

$$\log \left[\frac{p_i}{1 - p_i} \right] = \beta_0 + \beta_1 x_{i1} + \beta_6 x_{i6} \quad (14)$$

which has $k = 3$. Fitting this model gives

Variable	$\hat{\beta}_j$	SE($\hat{\beta}_j$)
Intercept	-7.513	1.706
x_1	0.06358	0.01963
x_6	0.01600	0.00794
$D = 138.8,$		$df = 197.$

To test the adequacy of the reduced model compared to the full model compute the difference in the deviances, $D = 138.8 - 134.9 = 3.9$ and compare that to a chi-squared distribution with degrees of freedom determined by the difference in the deviance degrees of freedom, $197 - 193 = 4$. The probability that a $\chi^2(4)$ distribution is greater than 3.9 is .42, which is the P value for the test.

There is considerable flexibility in the model testing approach. Suppose we suspect that it is not the blood pressure readings that are important but rather the difference between the blood pressure readings, $(x_2 - x_3)$. We can construct a model that has $\beta_3 = -\beta_2$. If we incorporate this hypothesis into the full model, we get

$$\begin{aligned} \log \left[\frac{p_i}{1 - p_i} \right] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + (-\beta_2) x_{i3} + \beta_4 x_{i4} \\ &\quad + \beta_5 x_{i5} + \beta_6 x_{i6} \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i2} - x_{i3}) + \beta_4 x_{i4} \\ &\quad + \beta_5 x_{i5} + \beta_6 x_{i6} \end{aligned} \quad (15)$$

which gives $D = 134.9$ on $df = 194$. This model is a special case of the full model, so a test of the

models has

$$D = 134.9 - 134.9 = 0, \quad (16)$$

with $df = 194 - 193 = 1$. The deviance difference is essentially 0, so the data are consistent with the reduced model.

This is a good time to discuss the alternate deviance computation. Logistic regression does not need to be performed on binary data. It can be performed on binomial data. Binomial data consist of the number of successes in a predetermined number of trials. The data would be $(N_i, y_i, x_{i1}, \dots, x_{i,k-1})$ for $i = 1, \dots, n$ where N_i is the number of trials in the i th case, y_i is the number of successes, and the predictor variables are as before. The logistic model is identical since it depends only on p_i , the probability of success in each trial, and the predictor variables. The likelihood and deviance require minor modifications.

For example, in our $k = 3$ model with age and weight, some computer programs will pool together all the people that have the same age and weight when computing the deviance. If there are 5 people with the same age-weight combination, they examine the number of coronary incidents out of this group of 5. Instead of having $n = 200$ cases, these 5 cases are treated as one case and n is reduced to $n - 4 = 196$. There can be several combinations of people with the same age and weight, thus reducing the effective number of cases even further. Call this new number of effective cases n' . The degrees of freedom for this deviance will be $n' - k$, which is different from the $n - k$ one gets using the original deviance computation. If the deviance degrees of freedom are something other than $n - k$, the computer program is using the alternative deviance computation.

Our original deviance compares the $k = 3$ model based on intercept, age, and weight to a model with $n = 200$ parameters and large sample theory does not apply. The alternative deviance compares the $k = 3$ model to a model with n' parameters, but in most cases, large sample theory will still not apply. (It should apply better. The whole point of the alternative computation is to make it apply better. But with binary data and continuous predictors, it rarely applies well enough to be useful.)

The problem with the alternative computation is that pooling cases together eliminates our ability to

use the deviance for model comparisons. In comparing our $k = 3$ model with our $k = 7$ model, it is likely that, with $n = 200$ men, some of them would have the same age and weight. However, it is very unlikely that these men would also have the same heights, blood pressures, and cholesterols. Thus, different people get pooled together in different models, and this is enough to invalidate model comparisons based on these deviances.

Variable Selection

Variable selection methods from standard regression such as forward selection, backwards elimination, and stepwise selection of variables carry over to logistic regression with almost no change. Extending the ideas of best subset selection from standard regression to logistic regression is more difficult.

One approach to best subset selection in logistic regression is to compare all possible models to the model

$$\log \left[\frac{p_i}{1 - p_i} \right] = \beta_0 \quad (17)$$

by means of score tests. The use of score tests makes this computationally feasible but trying to find a good model by comparing various models to a poor model is a bad idea. Typically, model (17) will fit the data poorly because it does not involve any of the predictor variables. A better idea is to compare the various candidate models to the model that contains all of the predictor variables. In our example, that involves comparing the $2^6 = 64$ possible models (every variable can be in or out of a possible model) to the full 6 variable ($k = 7$) model. Approximate methods for doing this are relatively easy (see [11] or [3, Section 4.4]), but are not commonly implemented in computer programs.

As in standard regression, if the predictor variables are highly correlated, it is reasonable to deal with the *collinearity* by performing *principal components regression* (see **Principal Component Analysis**). The principal components depend only on the predictor variables, not on y . In particular, the principal components do not depend on whether y is binary (as is the case here) or is a measurement (as in standard regression).

Outliers and Influential Observations

In standard regression, one examines potential **outliers** in the dependent variable y and unusual combinations of the predictor variables $(x_1, \dots, x_{k-1})$. Measures of influence combine information on the unusualness of y and $(x_1, \dots, x_{k-1})$ into a single number.

In binary logistic regression, there is really no such thing as an outlier in the y s. In binary logistic regression, y is either 0 or 1. Only values other than 0 or 1 could be considered outliers. Unusual values of y relate to what our model tells us about y . Young, fit men have heart attacks. They do not have many heart attacks, but some of them do have heart attacks. These are not outliers, they are to be expected. If we see a lot of young, fit men having heart attacks and our model does not explain it, we have a problem with the model (lack of fit), rather than outliers. Nonetheless, having a young fit man with a heart attack in our data would have a large influence on the overall nature of the fitted model. It is interesting to identify cases that have a large influence on different aspects of the fitted model.

Many of the influence measures from standard regression have analogues for logistic regression. Unusual combinations of predictor variables can be identified using a modification of the leverage. Analogues to Cook's distance (see **Multiple Linear Regression**) measure the influence (see **Influential Observations**) of an observation on estimated regression coefficients and on changes in the fitted log odds ($\hat{\eta}_i$ s). Pregibon [12] first developed diagnostic measures for logistic regression. Johnson [8] pointed out that influence measures should depend on the aspects of the model that are important in the application.

Testing Lack of Fit

Lack of fit occurs when the model being used is inadequate to explain the data. The basic problem in testing lack of fit is to dream up a model that is more general than the one currently being used but that has a reasonable chance of fitting the data. Testing the current (reduced) model against the more general (full) model provides a test for lack of fit. This is a modeling idea, so

the relevant issues are similar to those for standard regression.

One method of obtaining a more general model is to use a basis expansion. The idea of a basis expansion is that for some unknown continuous function of the predictor variables, say, $f(x_1, \dots, x_{k-1})$, the model $\log[p/(1-p)] = f(x_1, \dots, x_{k-1})$ should be appropriate. In turn, $f(x_1, \dots, x_{k-1})$ can be approximated by linear combinations of known functions belonging to some collection of basis functions. The basis functions are then used as additional predictor variables in the logistic regression so that this more general model can approximate a wide variety of possible models. The most common method for doing this is simply to fit additional polynomial terms. In our example, we could incorporate additional terms for age squared (x_1^2), age cubed (x_1^3), height squared times weight cubed ($x_5^2 x_6^3$), etc. Alternatives to fitting additional polynomial terms would be to add trigonometric terms such as $\sin(2x_1)$ or $\cos(3x_4)$ to the model. (Typically, the predictor variables should be appropriately standardized before applying trigonometric functions.) Other options are to fit wavelets or even splines (see **Scatterplot Smoothers**). See [4, Chapter 7] for an additional discussion of these methods.

A problem with this approach is that the more general models quickly become unwieldy. For example, with the LAHS data, an absolute minimum for a basis expansion would be to add all the terms x_{ik}^2 , $k = 1, \dots, 6$ and all pairs $x_{ik}x_{ik'}$ for $k \neq k'$. Including all of the original variables in the model, this gives us a new model with $1 + 6 + 6 + 15 = 28$ parameters for only 200 observations. A model that includes all possible second-order polynomial terms involves $3^6 = 729$ parameters.

An alternative to using basis expansions is to partition the predictor variable data into subsets. For example, if the subsets constitute sets of predictor variables that are nearly identical, one could fit the original model but add a separate intercept effect for each near replicate group. Alternatively, one could simply fit the entire model on different subsets and see whether the model changes from subset to subset. If it changes, it suggests lack of fit in the original model. Christensen [5, Section 6.6] discusses these ideas for standard regression.

A commonly used partitioning method involves creating subsets of the data that have similar $\hat{p}_i$ values. This imposes a partition on the predictor

variables. However, allowing the subsets to depend on the binary data (through the fitted values $\hat{p}_i$) causes problems in terms of finding an appropriate reference distribution for the difference in deviances test statistic. The usual χ^2 distribution does not apply for partitions selected using the binary data, see [7].

Landwehr, Pregibon, and Shoemaker [9] have discussed graphical methods for detecting lack of fit.

Exact Conditional Analysis

An alternative to the methods of analysis discussed above are methods based on exact conditional tests and their related confidence regions. These methods are mathematically correct, have the advantage of not depending on large sample approximations for their validity, and are similar in spirit to *Fisher's exact test* (see **Exact Methods for Categorical Data**) for **2 x 2 contingency tables**. Unfortunately, they are computationally intensive and require specialized software. From a technical perspective, they involve treating certain random quantities as fixed in order to perform the computations. Whether it is appropriate to fix (i.e., condition on) these quantities is an unresolved philosophical issue. See [1, Section 5] or, more recently, [10] for a discussion of this approach.

Random Effects

Mixed models are models in which some of the β_j coefficients are unobservable random variables rather than being unknown fixed coefficients. These random effects are useful in a variety of contexts. Suppose we are examining whether people are getting adequate pain relief. Further suppose we have some predictor variables such as age, sex, and income, say, x_1, x_2, x_3 . The responses y are now 1 if pain relief is adequate and 0 if not, however, the data involve looking at individuals on multiple occasions. Suppose we have n individuals and each individual is evaluated for pain relief T times. The overall data are y_{ij} , $i = 1, \dots, n$; $j = 1, \dots, T$. The responses on one individual should be more closely related than responses on different individuals and that can be incorporated into the model by using a random effect. For example, we might have a model in which, given an effect for individual i , say, β_{i0} , the y_{ij} s are

independent with probability determined by

$$\log \left[\frac{p_{ij}}{1 - p_{ij}} \right] = \beta_{i0} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}. \quad (18)$$

However, $\beta_{10}, \dots, \beta_{n0}$ are independent observations from a $N(\beta_0, \sigma^2)$ distribution. With this model, observations on different people are independent but observations on a given person i are dependent because they all depend on the outcome of the common random variable β_{i0} . Notice that the variables x_{ijs} are allowed to vary over the T times, even though in our description of the problem variables like age, sex, and income are unlikely to change substantially over the course of a study.

Such random effects models are examples of **generalized linear mixed models**. These are much more difficult to analyze than standard logistic regression models unless you perform a Bayesian analysis, see the section titled ‘Bayesian Analysis’.

Bayesian Analysis

Bayesian analysis (see **Bayesian Statistics**) involves using the data to update the analyst’s beliefs about the problem. It requires the analyst to provide a probability distribution that describes his knowledge/uncertainty about the problem prior to data collection. It then uses the likelihood function to update those views into a ‘posterior’ probability distribution.

Perhaps the biggest criticism of the Bayesian approach is that it is difficult and perhaps even inappropriate for the analyst to provide a prior probability distribution. It is difficult because it typically involves giving a prior distribution for the k regression parameters. The regression parameters are rather esoteric quantities and it is difficult to quantify knowledge about them. Tsutakawa and Lin [13] and Bedrick, Christensen, and Johnson [2] argued that it is more reasonable to specify prior distributions for the probabilities of success at various combinations of the predictor variables and to use these to induce a prior probability distribution on the regression coefficients. The idea that it may be inappropriate to specify a prior distribution is based on the fact that different analysts will have different information and thus can arrive at different results. Many practitioners of Bayesian analysis would argue that with sufficient

data, different analysts will substantially agree in their results and with insufficient data it is appropriate that they should disagree.

One advantage of Bayesian analysis is that it does not rely on large sample approximations. It is based on exact distributional results, although in practice it relies on computers to approximate the exact distributions. It requires specialized software, but such software is freely available. Moreover, the Bayesian approach deals with more complicated models, such as random effects models (see **Random Effects in Multivariate Linear Models: Prediction**), with no theoretical and minimal computational difficulty. See [6] for a discussion of Bayesian methods.

References

- [1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**, 131–153.
- [2] Bedrick, E.J., Christensen, R. & Johnson, W. (1997). Bayesian binomial regression, *The American Statistician* **51**, 211–218.
- [3] Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, 2nd Edition, Springer-Verlag, New York.
- [4] Christensen, R. (2001). *Advanced Linear Modeling*, 2nd Edition, Springer-Verlag, New York.
- [5] Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, 3rd Edition, Springer-Verlag, New York.
- [6] Congdon, P. (2003). *Applied Bayesian Modelling*, John Wiley and Sons, New York.
- [7] Hosmer, D.W., Hosmer, T., le Cessie, S. & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine* **16**, 965–980.
- [8] Johnson, W. (1985). Influence measures for logistic regression: another point of view, *Biometrika* **72**, 59–65.
- [9] Landwehr, J.M., Pregibon, D. & Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models, *Journal of the American Statistical Association* **79**, 61–71.
- [10] Mehta, C.R., Patel, N.R. & Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression, *Journal of the American Statistical Association* **95**, 99–108.
- [11] Nordberg, L. (1982). On variable selection in generalized linear and related regression models, *Communications in Statistics, A* **11**, 2427–2449.
- [12] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**, 705–724.
- [13] Tsutakawa, R.K. & Lin, H.Y. (1986). Bayesian estimation of item response curves, *Psychometrika* **51**, 251–267.

Further Reading

Hosmer, D.W. & Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley and Sons, New York.
McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman and Hall, London.

Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistic regression model, *Biometrika* **67**, 250–251.

RONALD CHRISTENSEN

Log-linear Models

JEROEN K. VERMUNT

Volume 2, pp. 1082–1093

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Log-linear Models

Introduction

Log-linear analysis has become a widely used method for the analysis of multivariate frequency tables obtained by cross-classifying sets of nominal, ordinal, or discrete interval level variables. (see **Scales of Measurement**) Examples of textbooks discussing categorical data analysis by means of log-linear models are [2], [4], [14], [15], [16], and [27].

We start by introducing the standard hierarchical log-linear modelling framework. Then, attention is paid to more advanced types of log-linear models that make it possible to impose interesting restrictions on the model parameters, for example, restrictions for ordinal variables. Subsequently, we present ‘regression analytic’, ‘path-analytic’, and ‘factor-analytic’ variants of log-linear analysis. The last section discusses parameter estimation by **maximum likelihood**, testing, and software for log-linear analysis.

Hierarchical Log-linear Models

Saturated Models

Suppose we have a frequency table formed by three categorical variables which are denoted by A , B , and C , with indices a , b , and c . The number of categories of A , B , and C is denoted by A^* , B^* , and C^* , respectively. Let m_{abc} be the expected frequency for the cell belonging to category a of A , b of B , and c of C . The saturated log-linear model for the three-way table ABC is given by

$$\log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} + \lambda_{abc}^{ABC}. \quad (1)$$

It should be noted that the log transformation of m_{abc} is tractable because it restricts the expected frequencies to remain within the admissible range. The consequence of specifying a linear model for the log of m_{abc} is that a multiplicative model is obtained for m_{abc} , that is,

$$m_{abc} = \exp(\lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} + \lambda_{abc}^{ABC})$$

$$= \tau \tau_a^A \tau_b^B \tau_c^C \tau_{ab}^{AB} \tau_{ac}^{AC} \tau_{bc}^{BC} \tau_{abc}^{ABC}. \quad (2)$$

From (1) and (2), it can be seen that the saturated model contains all interactions terms among A , B , and C . That is, no *a priori* restrictions are imposed on the data. However, (1) and (2) contain too many parameters to be identifiable. Given the values for the expected frequencies m_{abc} , there is not a unique solution for the λ and τ parameters. Therefore, constraints must be imposed on the log-linear parameters to make them identifiable. One option is to use **analysis of variance** (ANOVA)-like constraints, namely,

$$\begin{aligned} \sum_a \lambda_a^A &= \sum_b \lambda_b^B = \sum_c \lambda_c^C = 0, \\ \sum_a \lambda_{ab}^{AB} &= \sum_b \lambda_{ab}^{AB} = \sum_a \lambda_{ac}^{AC} = \sum_c \lambda_{ac}^{AC} \\ &= \sum_b \lambda_{bc}^{BC} = \sum_c \lambda_{bc}^{BC} = 0, \\ \sum_a \lambda_{abc}^{ABC} &= \sum_b \lambda_{abc}^{ABC} = \sum_c \lambda_{abc}^{ABC} = 0. \end{aligned}$$

This parameterization, in which every set of parameters sums to zero over each of its subscripts, is called *effect coding*. In effect coding, the λ term denotes the grand mean of $\log m_{abc}$. The one-variable parameters λ_a^A , λ_b^B , and λ_c^C indicate the relative number of cases at the various levels of A , B , and C as deviations from the mean. More precisely, they describe the partial skewness of a variable, that is, the skewness within the combined categories of the other variables. The two-variable interaction terms λ_{ab}^{AB} , λ_{ac}^{AC} , and λ_{bc}^{BC} indicate the strength of the partial association between A and B , A and C , and B and C , respectively. The partial association can be interpreted as the mean association between two variables within the levels of the third variable. And finally, the three-factor interaction parameters λ_{abc}^{ABC} indicate how much the conditional two-variable interactions differ from one another within the categories of the third variable.

Another method to identify the log-linear parameters involves fixing the parameters to zero for one category of A , B , and C , respectively. This parameterization, which is called *dummy coding*, is often used in regression models with nominal regressors.

2 Log-linear Models

Although the expected frequencies under both parameterizations are equal, the interpretation of the parameters is rather different. When effect coding is used, the parameters must be interpreted in terms of deviations from the mean, while under dummy coding, they must be interpreted in terms of deviations from the reference category.

Nonsaturated Models

As mentioned above, in a saturated log-linear model, all possible interaction terms are present. In other words, no *a priori* restrictions are imposed on the model parameters apart from the identifying restrictions. However, in most applications, the aim is to specify and test more parsimonious models, that is, models in which some *a priori* restrictions are imposed on the parameters. Log-linear models in which the parameters are restricted in some way are called *nonsaturated models*. There are different kinds of restrictions that can be imposed on the log-linear parameters. One particular type of restriction leads to the family of hierarchical log-linear models. These are models in which the log-linear parameters are fixed to zero in such a way that when a particular interaction term is fixed to zero, all higher-order interaction terms containing all its indices as a subset must also be fixed to zero. For example, if the partial association between A and B (λ_{ab}^{AB}) is assumed not to be present, the three-variable interaction λ_{abc}^{ABC} must be fixed to zero as well. Applying this latter restriction to (1) results in the following nonsaturated hierarchical log-linear model:

$$\log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}. \quad (3)$$

Another example of a nonsaturated hierarchical log-linear model is the (trivariate) independence model

$$\log m_{abc} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C.$$

Hierarchical log-linear models are the most popular log-linear models because, in most applications, it is not meaningful to include higher-order interaction terms without including the lower-order interaction terms concerned. Another reason is that it is relatively easy to estimate the parameters of hierarchical log-linear models because of the existence of simple minimal sufficient statistics (see **Estimation**).

Other Types of Log-linear Models

General Log-linear Model

So far, attention has been paid to only one special type of log-linear models, the hierarchical log-linear models. As demonstrated, hierarchical log-linear models are based on one particular type of restriction on the log-linear parameters. But, when the goal is to construct models which are as parsimonious as possible, the use of hierarchical log-linear models is not always appropriate. To be able to impose other kinds of linear restrictions on the parameters, it is necessary to use more general kinds of log-linear models.

As shown by McCullagh and Nelder [23], log-linear models can also be defined in a much more general way by viewing them as a special case of the **generalized linear modelling (GLM)** family. In its most general form, a log-linear model can be defined as

$$\log m_i = \sum_j \lambda_j x_{ij}, \quad (4)$$

where m_i denotes a cell entry, λ_j a log-linear parameter, and x_{ij} an element of the design matrix. The design matrix provides us with a very flexible tool for specifying log-linear models with various restrictions on the parameters. For detailed discussions on the use of design matrices in log-linear analysis, see, for example, [10], [14], [15], and [25].

Let us first suppose we want to specify the design matrix for an *hierarchical* log-linear model of the form $\{AB, BC\}$. Assume that A^* , B^* , and C^* , the number of categories of A , B , and C , are equal to 3, 3, and 4, respectively. Because in that case model $\{AB, BC\}$ has 18 independent parameters to be estimated, the design matrix will consist of 18 columns: 1 column for the main effect λ , 7 ($[A^* - 1] + [B^* - 1] + [C^* - 1]$) columns for the one-variable terms λ_a^A , λ_b^B , and λ_c^C , and 10 ($[A^* - 1] * [B^* - 1] + [B^* - 1] * [C^* - 1]$) columns for the two-variable terms λ_{ab}^{AB} and λ_{bc}^{BC} . The exact values of the cells of the design matrix, the x_{ij} , depend on the restrictions which are imposed to identify the parameters. Suppose, for instance, that column j refers to the one-variable term λ_a^A and that the highest level of A , A^* , is used as the (arbitrary) omitted category. In effect coding, the element of the design matrix corresponding to the i th cell, x_{ij} , will equal 1 if $A = a$, -1 if $A = A^*$, and otherwise 0. On

the other hand, in dummy coding, x_{ij} would be 1 if $A = a$, and otherwise 0. The columns of the design matrix referring to the two-variable interaction terms can be obtained by multiplying the columns for the one-variable terms for the variables concerned (see [10] and [14]).

The design matrix can also be used to specify all kinds of *nonhierarchical* and *nonstandard* models. Actually, by means of the design matrix, three kinds of linear restrictions can be imposed on the log-linear parameters: a parameter can be fixed to zero, specified to be equal to another parameter, and specified to be in a fixed ratio to another parameter.

The first kind of restriction, *fixing to zero*, is accomplished by deleting the column of the design matrix referring to the effect concerned. Note that, in contrast to hierarchical log-linear models, parameters can be fixed to be equal to zero without the necessity of deleting the higher-order effects containing the same indices as a subset.

Equating parameters is likewise very simple. Equality restrictions are imposed by adding up the columns of the design matrix which belong to the effects which are assumed to be equal. Suppose, for instance, that we want to specify a model with a symmetric association between the variables A and B ,¹ each having three categories. This implies that

$$\lambda_{ab}^{AB} = \lambda_{ba}^{AB}.$$

The design matrix for the unrestricted effect λ_{ab}^{AB} contains four columns, one for each of the parameters λ_{11}^{AB} , λ_{12}^{AB} , λ_{21}^{AB} , λ_{22}^{AB} . In terms of these four parameters, the symmetric association between A and B implies that λ_{12}^{AB} is assumed to be equal to λ_{21}^{AB} . This can be accomplished by summing the columns of the design matrix referring to these two effects.

As already mentioned above, parameters can also be restricted to be in a *fixed ratio* to each other. This is especially useful when the variables concerned can be assumed to be measured on an ordinal or interval level scale, with known scores for the different categories. Suppose, for instance, that we wish to restrict the one-variable effect of variable A to be linear. Assume that the categories scores of A , denoted by a , are equidistant, that is, that they take on the values 1, 2, and 3. Retaining the effect coding scheme, a linear effect of A is obtained by

$$\lambda_a^A = (a - \bar{a})\lambda^A.$$

Here, $\bar{a}$ denotes the mean of the category scores of A , which in this case is 2. Moreover, λ^A denotes the single parameter describing the one-variable term for A . It can be seen that the distance between the λ_a^A parameters of adjacent categories of A is λ^A . In terms of the design matrix, such a specification implies that instead of including $A^* - 1$ columns for the one-variable term for A , one column with scores $(a - \bar{a})$ has to be included.

These kinds of linear constraints can also be imposed on the bivariate association parameters of a log-linear model. The best known examples are linear-by-linear interaction terms and row- or column-effect models (see [5], [7], [13], and [15]). When specifying a linear-by-linear interaction term, it is assumed that the scores of the categories of both variables are known. Assuming equidistant scores for the categories of the variables A and B and retaining the effect coding scheme, the linear-by-linear interaction between A and B is given by

$$\lambda_{ab}^{AB} = (a - \bar{a})(b - \bar{b})\lambda^{AB}. \quad (5)$$

Using this specification, which is sometimes also called *uniform association*, the (partial) association between A and B is described by a single parameter instead of using $(A^* - 1)(B^* - 1)$ independent λ_{ab}^{AB} parameters. As a result, the design matrix contains only one column for the interaction between A and B consisting of the scores $(a - \bar{a})(b - \bar{b})$.

A row association structure is obtained by assuming the column variable to be linear. When A is the row variable, it is defined as

$$\lambda_{ab}^{AB} = (b - \bar{b})\lambda_a^{AB}.$$

Note that for every value of A , there is a λ_a^{AB} parameter. Actually, there are $(A^* - 1)$ independent row parameters. Therefore, the design matrix will contain $(A^* - 1)$ columns which are based on the scores $(b - \bar{b})$. The column association model is, in fact, identical to the row association model, only the roles of the column and row variable change.

Log-rate Model

The general log-linear model discussed in the previous section can be extended to include an additional

4 Log-linear Models

component, viz., a cell weight ([14] and [19]). The log-linear model with cell weights is given by

$$\log\left(\frac{m_i}{z_i}\right) = \sum_j \lambda_j x_{ij}$$

which can also be written as

$$\begin{aligned} \log m_i &= \log z_i + \sum_j \lambda_j x_{ij}, \\ m_i &= z_i \exp\left(\sum_j \lambda_j x_{ij}\right), \end{aligned}$$

where the z_i are the fixed *a priori* cell weights. Sometimes the vector with elements $\log z_i$ is also called the *offset matrix*.

The specification of a z_i for every cell of the contingency table has several applications. One of its possible uses is in the specification of Poisson regression models that take into account the population size or the length of the observation period. This leads to what is called a *log-rate model*, a model for rates instead of frequency counts ([6] and [14]). A rate is a number of events divided by the size of the population exposed to the risk of having the event.

The weight vector can also be used for taking into account sampling or nonresponse weights, in which case the z_i are equated to the inverse of the sampling weights ([1] and [6]). Another use is the inclusion of *fixed effects* in a log-linear model. This can be accomplished by adding the values of the λ parameters which attain fixed values to the corresponding $\log z_i$'s. The last application I will mention is in the analysis of tables with structural zeros, sometimes also called *incomplete tables* [15]. This simply involves setting the $z_i = 0$ for the structurally zero cells.

Log-multiplicative Model

The log-linear model is one of the GLMs, that is, it is a linear model for the logs of the cell counts in a frequency table. However, extensions of the standard log-linear model have been proposed that imply the inclusion of nonlinear terms, the best known example being the log-multiplicative row-column (RC) association models developed by Goodman [13] and Clogg [5] (see [7]). These RC

association models differ from the association models discussed in section 'General Log-linear Model' in that the row and column scores are not *a priori* fixed, but are treated as unknown parameters which have to be estimated as well. More precisely, a linear-by-linear association is assumed between two variables, given the unknown column and row scores.

Suppose we have a model for a three-way frequency table ABC containing log-multiplicative terms for the relationships between A and B and B and C . This gives the following log-multiplicative model:

$$\begin{aligned} \log m_{abc} &= \lambda + \lambda_a^A + \lambda_b^B + \lambda_c^C + \mu_a^{AB} \phi^{AB} \mu_b^{AB} \\ &\quad + \mu_b^{BC} \phi^{BC} \mu_c^{BC}. \end{aligned} \quad (6)$$

The ϕ parameters describe the strength of the association between the variables concerned. The μ 's are the unknown scores for the categories of the variables concerned. As in standard log-linear models, identifying restrictions have to be imposed on the parameters μ . One possible set of identifying restrictions on the log-multiplicative parameters which was also used by Goodman [13] is:

$$\begin{aligned} \sum_a \mu_a^{AB} &= \sum_b \mu_b^{AB} = \sum_b \mu_b^{BC} = \sum_c \mu_c^{BC} = 0 \\ \sum_a (\mu_a^{AB})^2 &= \sum_b (\mu_b^{AB})^2 = \sum_b (\mu_b^{BC})^2 \\ &= \sum_c (\mu_c^{BC})^2 = 1. \end{aligned}$$

This gives row and column scores with a mean of zero and a sum of squares of one.

On the basis of the model described in (6), both more restricted models and less restricted models can be obtained. One possible restriction is to assume the row and column scores within a particular partial association to be equal, for instance, μ_a^{AB} equal to μ_b^{AB} for all a equal to b . Of course, this presupposes that the number of rows equals the number of columns. Such a restriction is often used in the analysis of mobility tables [22]. It is also possible to assume that the scores for a particular variable are equal for different partial associations [5], for example, $\mu_b^{AB} = \mu_b^{BC}$. Less restricted models may allow for different μ and/or ϕ parameters within the levels of some other variable [5], for example,

different values of μ_a^{AB} , μ_b^{AB} , or ϕ^{AB} within levels of C . To test whether the strength of the association between the variables father's occupation and son's occupation changes linearly with time, Luijkx [22] specified models in which the ϕ parameters are a linear function of time.

As mentioned above, the RC association models assume a linear-by-linear interaction in which the row and column scores are unknown. Xie [31] demonstrated that the basic principle behind Goodman's RC association models, that is, linearly restricting log-linear parameters with unknown scores for the linear terms, can be applied to any kind of log-linear parameter. He proposed a general class of log-multiplicative models in which higher-order interaction terms can be specified in a parsimonious way.

Regression-, Path-, and Factor-analytic Models

Log-linear Regression Analysis: the Logit Model

In the log-linear models discussed so far, the relationships between the categorical variables are modeled without making a priori assumptions about their 'causal' ordering: no distinction is made between dependent and independent variables. However, one is often interested in predicting the value of a categorical response variable by means of explanatory variables. The logit model is such a 'regression analytic' model for a categorical dependent variable.

Suppose we have a response variable denoted by C and two categorical explanatory variables denoted by A and B . Moreover, assume that both A and B influence C , but that their effect is equal within levels of the other variable. In other words, it is assumed that there is no interaction between A and B with respect to their effect on C . This gives the following logistic model for the conditional probability of C given A and B , $\pi_{c|ab}$:

$$\pi_{c|ab} = \frac{\exp(\lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC})}{\sum_c \exp(\lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC})}. \quad (7)$$

When the response variable C is dichotomous, the logit can also be written as

$$\log\left(\frac{\pi_{1|ab}}{1 - \pi_{1|ab}}\right) = \log\left(\frac{\pi_{1|ab}}{\pi_{2|ab}}\right)$$

$$\begin{aligned} &= (\lambda_1^C - \lambda_2^C) + (\lambda_{a1}^{AC} - \lambda_{a2}^{AC}) \\ &\quad + (\lambda_{b1}^{BC} - \lambda_{b2}^{BC}) \\ &= \beta + \beta_a^A + \beta_b^B. \end{aligned}$$

It should be noted that the logistic form of the model guarantees that the probabilities remain in the admissible interval between 0 and 1.

It has been shown that a logit model is equivalent to a log-linear model which not only includes the same λ terms, but also the effects corresponding to the marginal distribution of the independent variables ([1], [11], [14]). For example, the logit model described in (7) is equivalent to the following log-linear model

$$\log m_{abc} = \alpha_{ab}^{AB} + \lambda_c^C + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}, \quad (8)$$

where

$$\alpha_{ab}^{AB} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}.$$

In other words, it equals log-linear model $\{AB, AC, BC\}$ for the frequency table with expected counts m_{abc} . With polytomous response variables, the log-linear or logit model of the form given in (8) is sometimes referred to as a multinomial response model. As shown by Haberman [15], in its most general form, the multinomial response model may be written as

$$\log m_{ik} = \alpha_k + \sum_j \lambda_j x_{ijk}, \quad (9)$$

where k is used as the index for the joint distribution of the independent variables and i as an index for the response variable.

Log-linear Path Analysis

After presenting a 'regression analytic' extension, we will now discuss a 'path-analytic' extension of log-linear analysis introduced by Goodman [12]. As is shown below, his 'modified **path analysis** approach' that makes it possible to take into account information on the causal and/or time ordering between the variables involves specifying a series of logit models.

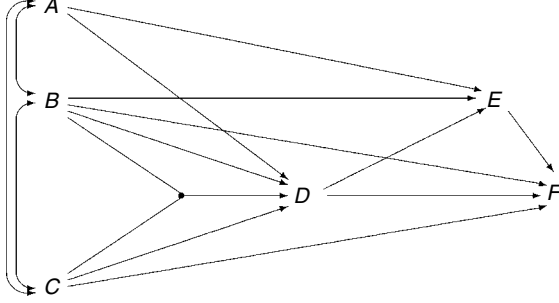


Figure 1 Modified path model

Suppose we want to investigate the causal relationships between six categorical variables denoted by A , B , C , D , E , and F . Figure 1 shows the assumed causal ordering and relationships between these variables, where a pointed arrow indicates that variables are directly related to each other, and a ‘knot’ that there is a higher-order interaction. The variables A , B , and C are exogenous variables. This means that neither their mutual causal order nor their mutual relationships are specified. The other variables are endogenous variables, where E is assumed to be posterior to D , and F is assumed to be posterior to E . From Figure 1, it can be seen that D is assumed to depend on A and on the interaction of B and C . Moreover, E is assumed to depend on A , B , and D , and F on B , C , D , and E .

Let $\pi_{def|abc}$ denote the probability that $D = d$, $E = e$, and $F = f$, given $A = a$, $B = b$, and $C = c$. The information on the causal ordering of the endogenous variables is used to decompose this probability into a product of marginal conditional probabilities ([12] and [30]). In this case, $\pi_{def|abc}$ can also be written as

$$\pi_{def|abc} = \pi_{d|abc} \pi_{e|abcd} \pi_{f|abcde}. \quad (10)$$

This is a straightforward way to indicate that the value on a particular variable can only depend on the preceding variables and not on the posterior ones. For instance, E is assumed to depend only on the preceding variables A , B , C , and D , but not on the posterior variable F . Therefore, the probability that $E = e$ depends only on the values of A , B , C , and D , and not on the value of F .

Decomposing the joint probability $\pi_{def|abc}$ into a set of marginal conditional probabilities is only the first step in describing the causal relationships

between the variables under study. In fact, the model given in (10) is still a saturated model in which it is assumed that a particular dependent variable depends on all its posterior variables, including all the higher-order interaction terms. A more parsimonious specification is obtained by using a log-linear or logit parameterization for the conditional probabilities appearing in (10) [12]. While only simple hierarchical log-linear models will be here used, the results presented apply to other kinds of log-linear models as well, including the log-multiplicative models discussed in section ‘Log-multiplicative Model’.

A system of logit models consistent with the path model depicted in Figure 1 leads to the following parameterization of the conditional probabilities appearing in (10):

$$\begin{aligned} \pi_{d|abc} &= \frac{\exp(\lambda_d^D + \lambda_{ad}^{AD} + \lambda_{bd}^{BD} + \lambda_{cd}^{CD} + \lambda_{bcd}^{BCD})}{\sum_d \exp(\lambda_d^D + \lambda_{ad}^{AD} + \lambda_{bd}^{BD} + \lambda_{cd}^{CD} + \lambda_{bcd}^{BCD})}, \\ \pi_{e|abcd} &= \frac{\exp(\lambda_e^E + \lambda_{ae}^{AE} + \lambda_{be}^{BE} + \lambda_{de}^{DE})}{\sum_e \exp(\lambda_e^E + \lambda_{ae}^{AE} + \lambda_{be}^{BE} + \lambda_{de}^{DE})}, \\ \pi_{f|abcde} &= \frac{\exp(\lambda_f^F + \lambda_{bf}^{BF} + \lambda_{cf}^{CF} + \lambda_{df}^{DF} + \lambda_{ef}^{EF})}{\sum_f \exp(\lambda_f^F + \lambda_{bf}^{BF} + \lambda_{cf}^{CF} + \lambda_{df}^{DF} + \lambda_{ef}^{EF})}. \end{aligned}$$

As can be seen, variable D depends on A , B , and C , and there is a three-variable interaction between B , C , and D ; E depends on A , B , and D , but there are no higher-order interactions between E and the independent variables; and F depends on B , C , D , and E .

Log-linear Factor Analysis: the Latent Class Model

As many concepts in the social sciences are difficult or impossible to measure directly, several directly observable variables, or indicators, are often used as indirect measures of the concept to be measured. The values of the indicators are assumed to be determined only by the unobservable value of the underlying variable of interest and by measurement error. In latent structure models, this principle is implemented statistically by assuming probabilistic

relationships between latent and manifest variables and by the assumption of local independence. Local independence means that the indicators are assumed to be independent of each other given a particular value of the unobserved or latent variable; in other words, they are only correlated because of their common cause (*see Conditional Independence*).

Latent structure models can be classified according to the measurement level of the latent variable(s) and the measurement level of the manifest variables. When both the latent and observed variables are categorical, one obtains a model called latent class model. As shown by Haberman [15], the latent class model can be defined as a log-linear model with one or more unobserved variables, yielding a ‘factor-analytic’ variant of the log-linear model.

Suppose there is, as depicted in Figure 2, a latent class model with one latent variable W with index w and 4 indicators A , B , C , and D with indices a , b , c , and d . Moreover, let W^* denote the number of latent classes. This latent class model is equivalent to the hierarchical log-linear model $\{WA, WB, WC, WD\}$; that is,

$$\log m_{abcd} = \lambda + \lambda_w^W + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + \lambda_{wa}^{WA} + \lambda_{wb}^{WB} + \lambda_{wc}^{WC} + \lambda_{wd}^{WD}. \quad (11)$$

In addition to the overall mean and the one-variable terms, it contains only the two-variable associations between the latent variable W and the manifest variables. As none of the interactions between the manifest variables are included, it can be seen that they are assumed to be conditionally independent of each other given W .

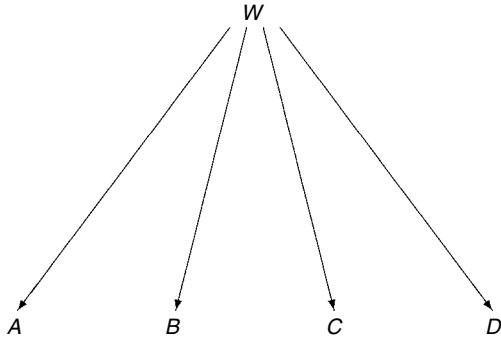


Figure 2 Latent class model

In its classical parameterization proposed by Lazarsfeld [21], the latent class model is defined as

$$\pi_{wabcd} = \pi_w \pi_{a|w} \pi_{b|w} \pi_{c|w} \pi_{d|w}. \quad (12)$$

It can be seen that again the observed variables A , B , C , and D are postulated to be mutually independent given a particular score on the latent variable W . Note that this is in fact a log-linear path model in which one variable is unobserved. The relation between the conditional probabilities appearing in (12) and the log-linear parameters appearing in (11) is

$$\pi_{a|w} = \frac{\exp(\lambda_a^A + \lambda_{wa}^{WA})}{\sum_a \exp(\lambda_a^A + \lambda_{wa}^{WA})}. \quad (13)$$

Estimation, Testing, and Software

Maximum Likelihood Estimation

Maximum likelihood (ML) estimates for the expected frequencies of a specific log-linear model are most easily derived assuming a Poisson sampling scheme (*see Catalogue of Probability Density Functions*), but the same estimates are obtained with a multinomial or product-multinomial sampling scheme. Denoting an observed frequency in a three-way table by n_{abc} , the relevant part of the Poisson log-likelihood function is

$$\log \mathcal{L} = \sum_{abc} (n_{abc} \log m_{abc} - m_{abc}), \quad (14)$$

where the expected frequencies m_{abc} are a function of the unknown λ parameters.

Suppose we want to find ML estimates for the parameters of the hierarchical log-linear model described in (3). Substituting (3) into (14) and collapsing the cells containing the same λ parameter, yields the following log-likelihood function:

$$\begin{aligned} \log \mathcal{L} = & n_{+++} \lambda + \sum_a n_{a++} \lambda_a^A + \sum_b n_{+b+} \lambda_b^B \\ & + \sum_c n_{++c} \lambda_c^C + \sum_{ab} n_{ab+} \lambda_{ab}^{AB} \\ & + \sum_{bc} n_{+bc} \lambda_{bc}^{BC} - \sum_{abc} \exp(u + \lambda_a^A + \lambda_b^B \\ & + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{bc}^{BC}), \end{aligned} \quad (15)$$

where a $+$ is used as a subscript to denote that the observed frequencies have to be collapsed over the dimension concerned. It can now be seen that the observed marginals n_{+++} , n_{a++} , n_{+b+} , n_{++c} , n_{ab+} , and n_{+bc} contain all the information needed to estimate the unknown parameters. Because knowledge of the bivariate marginals AB and BC implies knowledge of n_{+++} , n_{a++} , n_{+b+} , and n_{++c} , n_{ab+} and n_{+bc} are called *the minimal sufficient statistics*, the minimal information needed for estimating the log-linear parameters of the model of interest.

In hierarchical log-linear models, the minimal sufficient statistics are always the marginals corresponding to the interaction terms of the highest order. For this reason, hierarchical log-linear models are mostly denoted by their minimal sufficient statistics. The model given in (3) may then be denoted as $\{AB, BC\}$, the independence model as $\{A, B, C\}$, and the saturated model as $\{ABC\}$.

When no closed form expression exists for $\hat{m}_{abc}$, ML estimates for the expected cell counts can be found by means of the iterative proportional fitting algorithm (IPF) [8]. Let $\hat{m}_{abc}^{(v)}$ denote the estimated expected frequencies after the v th IPF iteration. Before starting the first iteration, arbitrary starting values are needed for the log-linear parameters that are in the model. In most computer programs based on the IPF algorithm, the iterations are started with all the λ parameters equal to zero, in other words, with all estimated expected frequencies $\hat{m}_{abc}^{(0)}$ equal to 1. For the model in (3), every IPF iteration consists of the following two steps:

$$\hat{m}_{abc}^{(v)'} = \hat{m}_{abc}^{(v-1)} \frac{n_{ab+}}{\hat{m}_{ab+}^{(v-1)}},$$

$$\hat{m}_{abc}^{(v)} = \hat{m}_{abc}^{(v)'} \frac{n_{+bc}}{\hat{m}_{+bc}^{(v)'}},$$

where the $\hat{m}_{abc}^{(v)'}$ and $\hat{m}_{abc}^{(v)}$ denote the improved estimated expected frequencies after imposing the ML related restrictions. The log-linear parameters are easily computed from the estimated expected frequencies.

Finding ML estimates for the parameters of other types of log-linear models is a bit more complicated than for the hierarchical log-linear model because the sufficient statistics are no longer equal to particular observed marginals. Most programs solve this problem using a *Newton-Raphson algorithm*. An alternative to

the Newton-Raphson algorithm is the *unidimensional Newton algorithm*. It differs from the multidimensional Newton algorithm in that it adjusts only one parameter at a time instead of adjusting them all simultaneously. In that sense, it resembles IPF. Goodman [13] proposed using the unidimensional Newton algorithm for the estimation of log-multiplicative models.

For ML estimation of latent class models, one can make use of an IPF-like algorithm called the *Expectation-Maximization* (EM) algorithm, a Newton-Raphson algorithm, or a combination of these (see **Maximum Likelihood Estimation; Optimization Methods**).

Model Selection

The goodness of fit of a postulated log-linear model can be assessed by comparing the observed frequencies, n , with the estimated expected frequencies, $\hat{m}$. For this purpose, usually two chi-square statistics are used: the likelihood-ratio statistic and the Pearson statistic. For a three-way table, the Pearson chi-square statistic equals

$$X^2 = \sum_{abc} \frac{(n_{abc} - \hat{m}_{abc})^2}{\hat{m}_{abc}},$$

and the likelihood-ratio chi-square statistic is

$$L^2 = 2 \sum_{abc} n_{abc} \log \left(\frac{n_{abc}}{\hat{m}_{abc}} \right). \quad (16)$$

The number of degrees of freedom for a particular model is

df = number of cells

– number of independent u parameters.

Both chi-square statistics have asymptotic, or large sample, chi-square distributions when the postulated model is true. In the case of small sample sizes and sparse tables, the chi-square approximation will generally be poor. Koehler [18] showed that X^2 is valid with smaller sample sizes and sparser tables than L^2 and that the distribution of L^2 is usually poor when the sample size divided by the number of cells is less than 5. Therefore, when sparse tables are analyzed, it is best to use both chi-square

statistics together. When X^2 and L^2 have almost the same value, it is more likely that both chi-square approximations are good. Otherwise, at least one of the two approximations is poor.²

The likelihood-ratio chi-square statistic is actually a conditional test for the significance of the difference in the value of the log-likelihood function for two nested models. Two models are nested when the restricted model has to be obtained by only linearly restricting some parameters of the unrestricted model. Thus, the likelihood-ratio statistic can be used to test the significance of the additional free parameters in the unrestricted model, given that the unrestricted model is true in the population. Assuming multinomial sampling, L^2 can be written more generally as

$$\begin{aligned} L^2_{(r|u)} &= (-2 \log \mathcal{L}_{(r)}) - (-2 \log \mathcal{L}_{(u)}) \\ &= 2 n_{abc} \log \hat{\pi}_{abc(u)} - 2 n_{abc} \log \hat{\pi}_{abc(r)} \\ &= 2 n_{abc} \log \left(\frac{\hat{m}_{abc(u)}}{\hat{m}_{abc(r)}} \right), \end{aligned}$$

where the subscript (u) refers to the unrestricted model and the subscript (r) to the restricted model. Note that in (16), a particular model is tested against the completely unrestricted model, the saturated model. Therefore, in (16), the estimated expected frequency in the numerator is the observed frequency n_{abc} . The $L^2_{(r|u)}$ statistic has a large sample chi-square distribution if the restricted model is approximately true. The approximation of the chi-square distribution may be good for conditional L^2 tests between non-saturated models even if the test against the saturated model is problematic, as in sparse tables. The number of degrees of freedom in conditional tests equals the number of parameters which are fixed in the restricted model compared to the unrestricted model. The $L^2_{(r|u)}$ statistic can also be computed from the unconditional L^2 values of two nested models,

$$L^2_{(r|u)} = L^2_{(r)} - L^2_{(u)},$$

with

$$\text{df}_{(r|u)} = \text{df}_{(r)} - \text{df}_{(u)}.$$

Another approach to model selection is based on information theory. The aim is not to detect the true model but the model that provides the most information about the real world. The best known information criteria are the **Akaike information**

criterion (AIC) [3] and the Schwarz [26] or bayesian information criterion (BIC). These two measures, which can be used to compare both nested and nonnested models, are usually defined as

$$AIC = L^2 - 2 \text{df}. \quad (17)$$

$$BIC = L^2 - \log N \text{df}. \quad (18)$$

Software

Software for log-linear analysis is readily available. Major statistical packages such as SAS and SPSS have modules for log-linear analysis that can be used for estimating hierarchical and general log-linear models, log-rate models, and logit models. Special software is required for estimating log-multiplicative models, log-linear path models, and latent class models. The command language based ℓ_{EM} program developed by Vermunt [27], [28] can deal with any of the models discussed in this article, as well as combinations of these. Vermunt and Magidson's [29] Windows based Latent GOLD can deal with certain types of log-linear models, logit models, and latent class models, as well as combinations of logit and latent class models.

An Application

Consider the four-way cross-tabulation presented in Table 1 containing data taken from four annual waves (1977–1980) of the National Youth Survey [9]. The table reports information on marijuana use of 237 respondents who were age 14 in 1977. The variable of interest is an ordinal variable measuring marijuana use in the past year. It has the three levels ‘never’ (1), ‘no more than once a month’ (2), and ‘more than once a month’ (3). We will denote these four time-specific measures by A , B , C , and D , respectively.

Several types of log-linear models are of interest for this data set. First, we might wish to investigate the overall dependence structure of these repeated responses, for example, whether it is possible to describe the data by a hierarchical log-linear model containing only the two-way associations between consecutive time points; that is, by a first-order Markov structure. Second, we might want to investigate whether it is possible to simplify the model by making use of the ordinal nature of the variables using uniform or RC association structures.

Table 1 Data on marijuana use in the past year taken from four yearly waves of the National Youth Survey (1977–1980)

1977 (A)	1978 (B)	1979 (C)								
		1			2			3		
		1980 (D)			1980 (D)			1980 (D)		
		1	2	3	1	2	3	1	2	3
1	1	115	18	7	6	6	1	2	1	5
1	2	2	2	1	5	10	2	0	0	6
1	3	0	1	0	0	1	0	0	0	4
2	1	1	3	0	1	0	0	0	0	0
2	2	2	1	1	2	1	0	0	0	3
2	3	0	1	0	0	1	1	0	2	7
3	1	0	0	0	0	0	0	0	0	1
3	2	1	0	0	0	1	0	0	0	1
3	3	0	0	0	0	2	1	1	1	6

Table 2 Goodness-of-fit statistics for the estimated models for the data in Table 1

Model	L^2	df	$\hat{p}$	AIC
1. Independence	403.3	72	.00	259.3
2. All two-variable terms	36.9	48	.12	−59.1
3. First-order Markov	58.7	60	.05	−61.3
4. Model 3 + $\lambda_{j\ell}^{BD}$	41.6	56	.30	−70.4
5. Model 4 with uniform associations	83.6	68	.00	−52.4
6. Two-class latent class model	126.6	63	.00	−51.0
7. Three-class latent class model	57.0	54	.12	−56.2

Third, latent class analysis could be used to determine whether it is possible to explain the associations by assuming that there is a small number of groups of children with similar developments in marijuana use.

Table 2 reports the L^2 values for the estimated models. Because the asymptotic P values are unreliable when analyzing sparse frequency tables such as the one we have here, we estimated the P values by means of 1000 parametric bootstrapping replications. The analysis was performed with the Latent GOLD program.

The high L^2 value obtained with Model 1 – the independence model $\{A, B, C, D\}$ – indicates that there is a strong dependence between the 4 time-specific measures. Model 2 is the model with all two-variable associations: $\{AB, AC, AD, BC, BD, CD\}$. As can be seen from its P value, it fits very well, which indicates that higher-order interactions are not needed. The Markov model containing only

associations between adjacent time points – Model 3: $\{AB, BC, CD\}$ – seems to be too restrictive for this data set. It turns out that we need to include one additional term; that is, the association between the second and fourth time point, yielding $\{AB, BC, BD, CD\}$ (Model 4).

Model 5 has the same structure as Model 4, with the only difference that the two-variable terms are assumed to be uniform associations (see 5). This means that each two-way association contains only one instead of four independent parameters. These ‘ordinal’ constraints seems to be too restrictive for this data set.

Models 6 and 7 are latent class models or, equivalently, log-linear models of the form $\{XA, XB, XC, XD\}$, where X is a latent variable with either two or three categories. The fit measures indicate that the associations between the time points can be explained by the existence of three types of trajectories of marijuana use.

Based on the comparison of the goodness of fit measures for the various models, as well as their AIC values that also take into account parsimony, one can conclude that Model 4 is the preferred one. The three-class, however, yields a somewhat simpler explanation for the associations between the time-specific responses.

Notes

1. Log-linear models with symmetric interaction terms may be used for various purposes. In longitudinal research, they may be applied to test the assumption of marginal homogeneity (see [1] and [16]). Other applications of log-linear models with symmetric association parameters are Rasch models for dichotomous (see [24] and [17]) and polytomous items (see [2]).
2. An alternative approach is based on estimating the sampling distributions of the statistics concerned rather than using their asymptotic distributions. This can be done by bootstrap methods [20]. These computationally intensive methods are becoming more and more applicable as computers become faster.

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*, 2nd Edition, 2002, Wiley, New York.
- [2] Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters, *Scandinavian Journal of Statistics* **20**, 63–71.
- [3] Akaike, H. (1987). Factor analysis and AIC, *Psychometrika* **52**, 317–332.
- [4] Bishop, R.J., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.
- [5] Clogg, C.C. (1982). Some models for the analysis of association in multiway cross-classifications having ordered categories, *Journal of the American Statistical Association* **77**, 803–815.
- [6] Clogg, C.C. & Eliason, S.R. (1987). Some common problems in log-linear analysis, *Sociological Methods and Research* **16**, 8–14.
- [7] Clogg, C.C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Data*, Sage Publications, Thousand Oaks.
- [8] Darroch, J.N. & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models, *The Annals of Mathematical Statistics* **43**, 1470–1480.
- [9] Elliot, D.S., Huizinga, D. & Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*, Springer-Verlag, New York.
- [10] Evers, M. & Namboodiri, N.K. (1978). On the design matrix strategy in the analysis of categorical data, in *Sociological Methodology 1979*, K.F. Schuessler, ed., Jossey Bass, San Francisco, pp. 86–111.
- [11] Goodman, L.A. (1972). A modified multiple regression approach for the analysis of dichotomous variables, *American Sociological Review* **37**, 28–46.
- [12] Goodman, L.A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach, *Biometrika* **60**, 179–192.
- [13] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.
- [14] Haberman, S.J. (1978). *Analysis of Qualitative Data*, Vol. 1, *Introduction Topics*, Academic Press, New York, San Francisco, London.
- [15] Haberman, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2, *New Developments*, Academic Press, New York.
- [16] Hagenaars, J.A. (1990). *Categorical Longitudinal Data – Loglinear Analysis of Panel, Trend and Cohort Data*, Sage Publications, Newbury Park.
- [17] Kelderman, H. (1984). Log-linear Rasch model tests, *Psychometrika* **49**, 223–245.
- [18] Koehler, K.J. (1986). Goodness-of-fit statistics for log-linear models in sparse contingency tables, *Journal of the American Statistical Association* **81**, 483–493.
- [19] Laird, N. & Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.
- [20] Langeheine, R., Pannekoek, J. & Van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis, *Sociological Methods and Research* **24**, 492–516.
- [21] Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis, S.A. Stouffer, ed. *Measurement and Prediction*, Princeton University Press, Princeton, 362–412.
- [22] Luijckx, R. (1994). *Comparative Loglinear Analyses of Social Mobility and Heterogamy*, Tilburg University Press, Tilburg.
- [23] McCullagh, P., & Nelder, J.A. (1983). *Generalized Linear Models*, 2nd Edition, 1989, Chapman & Hall, London.
- [24] Mellenbergh, G.J. & Vijn, P. (1981). The Rasch model as a log-linear, *Applied Psychological Measurement* **5**, 369–376.
- [25] Rindskopf, D. (1990). Nonstandard loglinear models, *Psychological Bulletin* **108**, 150–162.
- [26] Schwarz, G. (1978). Estimating the dimensions of a model, *Annals of Statistics* **6**, 461–464.
- [27] Vermunt, J.K. (1997). Log-linear models for event history histories, *Advanced Quantitative Techniques in the Social Sciences Series*, Vol. 8, Sage Publications, Thousand Oaks.
- [28] Vermunt, J.K. (1997). *LEM: A General Program for the Analysis of Categorical Data. User's Manual*, Tilburg University.

12 Log-linear Models

- [29] Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD 2.0 User's Guide*, Statistical Innovations Inc, Belmont. (See also **Measures of Association**)
- [30] Wermuth, N. & Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables, *Biometrika* **70**, 537–552. JEROEN K. VERMUNT
- [31] Xie, Yu. (1992). The log-multiplicative layer effects model for comparing mobility tables, *American Sociological Review* **57**, 380–395.

Log-linear Rasch Models for Stability and Change

THORSTEN MEISER

Volume 2, pp. 1093–1097

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Log-linear Rasch Models for Stability and Change

The latent trait model proposed by Georg Rasch [18, 19] specifies the probability that a person gives the correct response to a test item as a function of the person's ability and the item's difficulty:

$$p(X_{vi} = 1 | \theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \quad (1)$$

In the equation, $X_{vi} = 1$ means that item i is successfully solved by person v , θ_v denotes the ability parameter of person v , and β_i denotes the difficulty parameter of item i . The logistic function in (1) leads to the s-shaped item characteristic curve displayed in Figure 1. The item characteristic curve shows the probability to solve a given item i with its difficulty parameter β_i for varying ability parameters θ . As illustrated in Figure 1, the probability is 0.5 if the ability parameter θ equals the item's difficulty parameter β_i . For smaller values of θ , the probability to solve the item steadily descends to its lower asymptote of 0. At the other extreme, for larger values of θ , the probability ascends to its upper asymptote of 1.

Rasch Models for Longitudinal Data

The Rasch model can be used to measure latent traits like abilities or attitudes at a given point in time. Aside from the measurement of latent constructs at one occasion (see **Latent Class Analysis**), Rasch

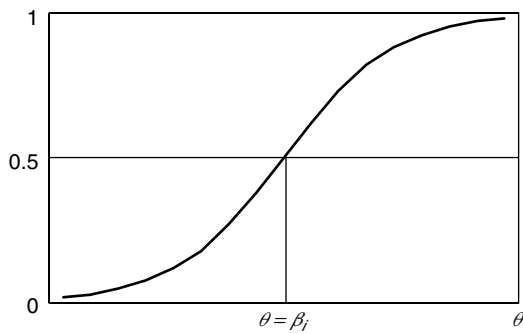


Figure 1 Item characteristic curve of the Rasch model

models also allow one to analyze change in longitudinal research designs (see **Longitudinal Data Analysis**), in which the same items are repeatedly administered. A simple extension of the Rasch model to longitudinal data follows from the assumption of *global change* [7, 22]:

$$p(X_{vit} = 1 | \theta_v, \beta_i, \lambda_t) = \frac{\exp(\theta_v + \lambda_t - \beta_i)}{1 + \exp(\theta_v + \lambda_t - \beta_i)} \quad (2)$$

In (2), X_{vit} denotes the response of person v to item i at occasion t , and the new parameter λ_t reflects the change on the latent dimension that has occurred until occasion t , with $\lambda_1 = 0$. Because the change parameter λ_t is independent of person v and item i , the model specifies a global change process that is homogeneous across persons and items. Homogeneity of change across persons means that all persons move on the latent continuum in the same direction and by the same amount. As a consequence, interindividual differences in the latent trait θ are preserved over time. Figure 2 illustrates homogeneity of change and the resulting stability of interindividual differences. The figure depicts the probability to solve an item with given difficulty parameter for two persons with initial values θ_1 and θ_2 on the latent trait continuum. As indicated in the figure, the difference between the two persons remains constant for any given change parameter λ_t .

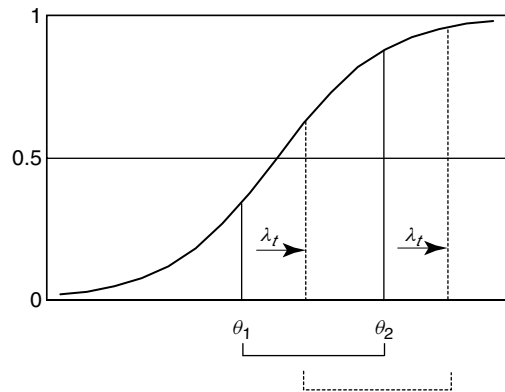


Figure 2 Global change on the latent trait continuum with preservation of interindividual differences. Solid lines symbolize the first measurement occasion and dashed lines symbolize the second measurement occasion

2 Log-linear Rasch Models

The rather restrictive assumption of homogeneous change across persons can be dropped by specifying different person parameters for different occasions [1, 22]:

$$p(X_{vit} = 1 | \theta_{vt}, \beta_i) = \frac{\exp(\theta_{vt} - \beta_i)}{1 + \exp(\theta_{vt} - \beta_i)} \quad (3)$$

Here, the latent trait parameter θ_{vt} reflects an interaction of person v and occasion t , so that the model permits *person-specific change*. Formally, (3) presents a multidimensional latent trait model, because it contains one latent trait dimension for each of t occasions. As a consequence, the model allows for differences in the latent trait between persons at each occasion, and these interindividual differences may vary over time.

Log-linear Representation of Rasch Models for the Analysis of Change

The expected probabilities of response vectors under the Rasch model can be represented by **log-linear models** [5, 11]. The log-linear model formulation facilitates flexible hypothesis tests, which makes it particularly suitable for the analysis of latent change [14, 17].

Imagine that a set of I items is applied at two measurement occasions and that $X_{i1} = 1$ denotes a correct response to item i at the first occasion, whereas $X_{i1} = 0$ denotes an incorrect response. Likewise, let $X_{i2} = 1$ and $X_{i2} = 0$ mean success and failure, respectively, with respect to item i at the second occasion. Then $\mathbf{X} = (X_{11}, \dots, X_{I1}, X_{12}, \dots, X_{I2})$ is the response vector of zeros and ones referring to the I items at the two occasions. Under the Rasch model of global change defined in (2), the logarithm of the probability of a given response vector can be written as [14, 16]

$$\begin{aligned} \ln p(\mathbf{X} = (x_{11}, \dots, x_{I1}, x_{12}, \dots, x_{I2})) \\ = u - \sum_{j=1}^2 \sum_{i=1}^I x_{ij} \beta_i + \sum_{i=1}^I x_{i2} \lambda_2 + u_s. \end{aligned} \quad (4)$$

The parameter u denotes the overall constant, and u_s is an additional constant for all response vectors with total score s , that is, for all $\mathbf{x}$ with $\sum_j \sum_i x_{ij} = s$. The total score parameter s is the sufficient statistic for person parameter θ and carries all the

information about θ from the response vector [2, 6]. Therefore, (4) specifies the logarithm of the expected probability of the response vector across two occasions in terms of a linear combination of the two constants, the set of item parameters, and the change parameter.

Analogously, the probabilities of response vectors under the Rasch model of person-specific change defined in (3) can be represented by a log-linear model of the form

$$\begin{aligned} \ln p(\mathbf{X} = (x_{11}, \dots, x_{I1}, x_{12}, \dots, x_{I2})) \\ = u - \sum_{j=1}^2 \sum_{i=1}^I x_{ij} \beta_i + u_{(s_1, s_2)} \end{aligned} \quad (5)$$

[14, 16]. The term $u_{(s_1, s_2)}$ specifies a constant for all response vectors with total score s_1 at the first occasion and total score s_2 at the second occasion, that is, for all $\mathbf{x}$ with $\sum_i x_{i1} = s_1$ and $\sum_i x_{i2} = s_2$. The combination of total scores s_1 and s_2 is the sufficient statistic for the combination of latent traits θ_1 and θ_2 concerning the first and second occasion. Therefore, the logarithm of the expected probability of a given response vector can be written as a linear combination of the constant terms and the item parameters.

The log-linear representation of Rasch models facilitates the specification and test of hypotheses about change. For instance, the occurrence of global change can be tested by comparing the full log-linear model in (4) with a restricted version that contains the parameter fixation $\lambda_2 = 0$. As the fixation implies that there is no change in the latent trait from the first to the second occasion, a significant difference between the full model and the restricted model indicates that global change takes place. In a similar vein, a statistical comparison between the log-linear model of global change in (4) and the log-linear model of person-specific change in (5) may be used to test for homogeneity of change across persons. The model of person-specific change includes the model of global change as a special case, because setting $u_{(s_1, s_2)} = u_{(s_1 + s_2)} + s_2 \cdot \lambda$ in (5) yields (4) [14, 16]. Hence, a significant difference between the two models indicates that the homogeneity assumption is violated and that change is person specific. Such testing options can be used to investigate the course of development in longitudinal research as well as to analyze the effectiveness of intervention programs in evaluation studies.

Identifiability, Parameter Estimation, and Model Testing

Equations (4) and (5) specify nonstandard log-linear models, which can be presented in terms of their design matrices and parameter vectors (see [4, 20]). Let μ be the vector of expected probabilities and $\mathbf{B}$ the vector of model parameters. Then the design matrix $\mathbf{M}$ defines the linear combination of model parameters for each expected probability, such that $\mu = \mathbf{MB}$. The model parameters are identifiable if $\mathbf{M}$ is of full rank. To achieve identifiability, restrictions have to be imposed on the log-linear parameters in (4) and (5). Appropriate restrictions follow from the common constraints that the item parameters and the total score parameters sum to zero, that is, $\sum_i \beta_i = 0$, $\sum_s u_s = 0$ and $\sum_{s_1} \sum_{s_2} u_{(s_1, s_2)} = 0$.

Nonstandard log-linear models like the Rasch models in (4) and (5) can be implemented in statistical programs which allow user-defined design matrices, such as SAS or LEM [23] (see **Structural Equation Modeling: Software**). These programs provide **maximum likelihood estimates** of the parameters and **goodness-of-fit** indices for the model. The goodness of fit of a log-linear model can be tested by means of the likelihood ratio statistic G^2 , which asymptotically follows a χ^2 distribution [2, 3]. The number of degrees of freedom equals the number of response vectors minus the number of model parameters. If two nonstandard log-linear models are hierarchically nested, then the two models can be compared using the conditional likelihood ratio statistic ΔG^2 . Two models are said to be hierarchically nested if the model with fewer parameters can be derived from the other model by a set of parameter restrictions. If such a hierarchical relation exists, ΔG^2 can be computed as the difference of the likelihood ratios of the two models. The number of degrees of freedom for the model comparison equals the difference in the number of parameters between the two models.

An Empirical Example

The use of log-linear Rasch models for the analysis of change can be illustrated by a recent reanalysis of data from the longitudinal SCHOLASTIK study [16]. The SCHOLASTIK study assessed children at several elementary schools in Germany with various

performance measures from first through fourth grade [25]. The reanalysis with log-linear Rasch models focused on the responses of $N = 1030$ children to specific kinds of arithmetic word problems applied in second and third grade. Figure 3 shows the relative frequencies of correct solutions to three arithmetic items. The log-linear Rasch model of homogeneous change (see (4)) yielded a significant parameter of global learning from Grade 2 to Grade 3 for these items, $\hat{\lambda} = 0.669$. Moreover, differences in item difficulty were indicated by the estimated item parameters. Mirroring the relative frequencies of correct solutions in Figure 3, Item 1 was the least difficult item, $\hat{\beta}_1 = -0.244$, Item 2 was the most difficult item, $\hat{\beta}_2 = 0.310$, and Item 3 fell in between the other two, $\hat{\beta}_3 = -0.066$.

However, the model of global change provided only a poor goodness of fit to the empirical data, $G^2(54) = 72.53$, $p = 0.047$, suggesting that the homogeneity assumption may be violated. The less restrictive model of person-specific change (see (5)) fitted the data well, $G^2(46) = 54.84$, $p = 0.175$. In fact, a statistical comparison of the two models indicated that the model of person-specific change shows a significantly better fit than the model of homogeneous change, $\Delta G^2(8) = 17.69$, $p = 0.024$. We may thus conclude that there are not only interindividual differences in the ability to solve arithmetic problems, but that the children also differ in their speed of development from Grade 2 to Grade 3.

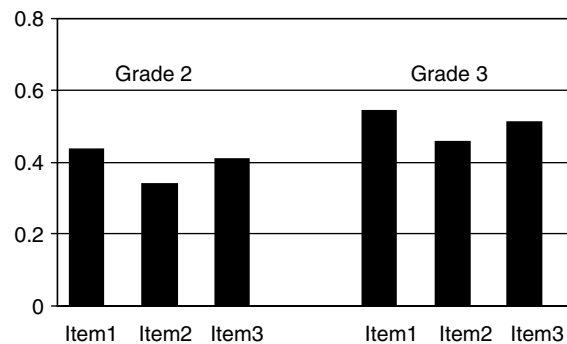


Figure 3 Relative frequencies of correct responses to three arithmetic word problems at two measurement occasions

Model Extensions

The models described in this entry can easily be extended to more than two measurement occasions and in various other ways. First, if effects of different treatments are to be compared in an experimental or quasi-experimental design, a grouping variable can be introduced that allows one to test for differences in the change parameter λ as a function of treatment group [7]. Second, generalizations of the Rasch model to polytomous items with more than two response categories have been adopted for the measurement of change [8–10]. Third, the log-linear representation of Rasch models was extended to models for the simultaneous measurement of more than one latent trait [12]. Accordingly, log-linear Rasch models of global and person-specific change can be formulated that include several ordered response categories per item and multidimensional latent constructs at each measurement occasion [14, 16].

Furthermore, model extensions in terms of mixture distribution models [13] appear especially useful for the analysis of change. Mixture distribution Rasch models extend the Rasch model by specifying several latent subpopulations that may differ with respect to the model parameters [21, 24]. While parameter homogeneity is maintained within each latent subpopulation, heterogeneity is admitted between subpopulations. Applied to longitudinal designs, mixture distribution Rasch models can be used to uncover different developmental trajectories in a population by means of subpopulation-specific change parameters [15]. In addition, *a priori* assumptions about differences in change can be specified by parameter restrictions for the latent subpopulations.

To illustrate the use of mixture distribution Rasch models in the analysis of change, we continued the above empirical example of arithmetic problems in elementary school. Because the assumption of homogeneity of change did not hold, one may surmise that there are two subpopulations of children: one subpopulation of children who profit from the training at school and improve their performance from Grade 2 to Grade 3, and a second subpopulation of children whose performance remains unchanged from Grade 2 to Grade 3. We tested this conjecture by means of a mixture distribution model with two subpopulations [16]. For each subpopulation, a Rasch model of homogeneous change (see (2) and (4)) was specified, and in one of the subpopulations the change

parameter was fixed at zero. The resulting ‘mover-stayer Rasch model’ fitted the empirical data well and provided a parsimonious account of the observed heterogeneity of change.

References

- [1] Andersen, E.B. (1985). Estimating latent correlations between repeated testings, *Psychometrika* **50**, 3–16.
- [2] Andersen, E.B. (1990). *The Statistical Analysis of Categorical Data*, Springer, Berlin.
- [3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.
- [4] Christensen, R. (1990). *Log-linear Models*, Springer, New York.
- [5] Cressie, N. & Holland, P.W. (1983). Characterizing the manifest probabilities of latent trait models, *Psychometrika* **48**, 129–141.
- [6] Fischer, G.H. (1995a). Derivations of the Rasch model, in *Rasch Models. Foundations, Recent Developments and Applications*, G.H. Fischer & I.W. Molenaar, eds, Springer, New York, pp. 15–38.
- [7] Fischer, G.H. (1995b). Linear logistic models for change, in *Rasch Models. Foundations, Recent Developments and Applications*, G.H. Fischer & I.W. Molenaar, eds, Springer, New York, pp. 157–180.
- [8] Fischer, G.H. & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change, *Psychometrika* **56**, 637–651.
- [9] Fischer, G.H. & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change, *Psychometrika* **59**, 177–192.
- [10] Fischer, G.H. & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change, in *Rasch Models. Foundations, Recent Developments and Applications*, G.H. Fischer & I.W. Molenaar, eds, Springer, New York, pp. 353–370.
- [11] Kelderman, H. (1984). Log-Linear Rasch model tests, *Psychometrika* **49**, 223–245.
- [12] Kelderman, H. & Rijkens, C.P.M. (1994). Log-Linear multidimensional IRT models for polytomously scored items, *Psychometrika* **59**, 149–176.
- [13] McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.
- [14] Meiser, T. (1996). Log-Linear Rasch models for the analysis of stability and change, *Psychometrika* **61**, 629–645.
- [15] Meiser, T., Hein-Eggers, M., Rompe, P. & Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models: a comparative and integrative approach, *Applied Psychological Measurement* **19**, 377–391.
- [16] Meiser, T., Stern, E. & Langeheine, R. (1998). Latent change in discrete data: unidimensional, multidimensional, and mixture distribution Rasch models for the

- analysis of repeated observations, *Methods of Psychological Research* **3**, 75–93.
- [17] Meiser, T., von Eye, A. & Spiel, C. (1997). Log-Linear symmetry and quasi-symmetry models for the analysis of change, *Biometrical Journal* **3**, 351–368.
- [18] Rasch, G. (1968). An individualistic approach to item analysis, in *Readings in Mathematical Social Science*, P.F. Lazarsfeld & N.W. Henry, eds, MIT Press, Cambridge.
- [19] Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, The University of Chicago Press, Chicago. (Original published 1960, The Danish Institute of Educational Research, Copenhagen).
- [20] Rindskopf, D. (1990). Nonstandard log-linear models, *Psychological Bulletin* **108**, 150–162.
- [21] Rost, J. (1990). Rasch models in latent classes: an integration of two approaches in item analysis, *Applied Psychological Measurement* **14**, 271–282.
- [22] Spada, H. & McGaw, B. (1985). The assessment of learning effects with linear logistic test models, in *Test Design. Developments in Psychology and Psychometrics*, S. Embretson, ed., Academic Press, Orlando, pp. 169–194.
- [23] Vermunt, J.K. (1997). *LEM 1.0: A General Program for the Analysis of Categorical Data*, Tilburg University, Tilburg.
- [24] von Davier, M. & Rost, J. (1995). Polytomous mixed Rasch models, in *Rasch Models. Foundations, Recent Developments and Applications*, G.H. Fischer & I.W. Molenaar, eds, Springer, New York, pp. 371–379.
- [25] Weinert, F.E. & Helmke, A. (1997). *Entwicklung im Grundschulalter [Development in Elementary School]*, Psychologie Verlags Union, Weinheim.

(See also **Log-linear Models; Rasch Modeling; Rasch Models for Ordered Response Categories**)

THORSTEN MEISER

Longitudinal Data Analysis

BRIAN S. EVERITT

Volume 2, pp. 1098–1101

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Longitudinal Data Analysis

The distinguishing feature of a longitudinal study is that the response variable of interest and a set of explanatory variables (covariates) are measured repeatedly over time. The main objective in such a study is to characterize change in the response variable over time and to determine the covariates most associated with any change. Because observations of the response variable are made on the same individual at different times, it is likely that the repeated responses for the same person will be correlated with each other (*see* **Clustered Data**). This correlation must be taken into account to draw valid and efficient inferences about parameters of scientific interest. It is the likely presence of such correlation that makes modeling longitudinal data more complex than dealing with a single response for each individual. (Time to event data is also thus a form of longitudinal data – *see* **Event History Analysis**; **Survival Analysis**.)

An Example of a Longitudinal Study

Despite the possible complications of analysis, longitudinal studies are popular in psychology and other disciplines since they allow for understanding the development and persistence of behavior or disease and for identifying factors that can alter the course of either one (*see* **Clinical Trials and Intervention Studies**). For example, readers of this article are unlikely to need to be reminded that depression is a major public health problem across the world. Antidepressants are the frontline treatment, but many patients either do not respond to them or do not like taking them. The main alternative is psychotherapy, and the modern ‘talking treatments’ such as cognitive behavioral therapy (CBT) have been shown to be as effective as drugs, and probably more so when it comes to relapse [5]. But there is a problem, namely, availability – there are not enough skilled therapists to meet the demand, and little prospect at all of this situation changing. A number of alternative modes of delivery of CBT have been explored, including interactive systems making use of the new computer technologies. The principles of CBT lend themselves

reasonably well to computerization, and, perhaps surprisingly, patients adapt well to this procedure, and do not seem to miss the physical presence of the therapist as much as one might expect. One such attempt at computerization, known as ‘Beating the Blues (BtB)’, is described in detail in [4]. But, in essence, BtB is an interactive program using multimedia techniques, in particular, video vignettes. The computer-based intervention consists of nine sessions, followed by eight therapy sessions, each lasting about 50 minutes. Nurses are used to explain how the program works, but are instructed to spend no more than 5 minutes with each patient at the start of each session, and are there simply to assist with the technology. In a randomized controlled trial of the program, patients with depression recruited in primary care were randomized to either the BtB program, or to ‘Treatment as Usual (TAU)’. Patients randomized to BtB also received pharmacology and/or general GP support and practical/social help, offered as part of treatment as usual, with the exception of any face-to-face counseling or psychological intervention. Patients allocated to TAU received whatever treatment their GP prescribed. The latter included, besides any medication, discussion of problems with GP, provision of practical/social help, referral to a counselor, referral to a practice nurse, referral to mental health professionals (psychologist, psychiatrist, community psychiatric nurse, counselor), or further physical examination.

A number of outcome measures were used in the trial, with the primary response being the *Beck Depression Inventory II* – [1]. Measurements on this variable were made on the following five occasions:

- prior to treatment;
- two months after treatment began;
- at one, three, and six months follow-up, that is, at three, five, and eight months after treatment.

The data collected in this study have the following features that are fairly typical of the data collected in many clinical trials and intervention studies in psychiatry and psychology:

- there are a considerable number of **missing values** caused by patients dropping out of the study (*see* **Dropouts in Longitudinal Data**; **Dropouts in Longitudinal Studies: Methods of Analysis**);
- there are repeated measurements of the outcome taken on each patient posttreatment, along with a baseline pretreatment measurement;

2 Longitudinal Data Analysis

- the data is multicenter in that they have been collected from a number of different GP surgeries.

Here, the investigator is primarily interested in assessing the effect of treatment on the repeated measurements of the depression score, although a number of other covariates were also available, for example, the length of the current episode of depression and whether or not the patient was taking antidepressants. How should such data be analyzed?

Graphical Methods

Graphical displays of data in general, and longitudinal data in particular, are often useful for exposing patterns in the data and this might be of great help in suggesting which more formal methods of analysis might be applicable to the data (*see Graphical Presentation of Longitudinal Data*). According to [3], there is no single prescription for making effective displays of longitudinal data, although they do offer the following simple guidelines:

- they show as much of the relevant raw data as possible rather than only data summaries;
- they highlight aggregate patterns of potential scientific interest;
- they identify both cross-sectional and longitudinal patterns;
- they make easy the identification of unusual individuals or unusual observations.

Two examples of useful graphs for the data from the BtB study are shown in Figures 1 and 2. In the first example, the profiles of individual subjects are plotted, and in the second one, the mean profiles of each treatment group. The diagrams show a decline in depression scores over time, with the BtB group having generally lower scores.

The Analysis of Longitudinal Data

As in other areas, it cannot be overemphasized that statistical analyses of longitudinal data should be no more complex than necessary. So it will not always be essential to use one of the modeling approaches mentioned later since a simpler procedure may often not only be statistically and scientifically adequate but can also be more persuasive and easier to communicate than more ambitious analysis. One simple

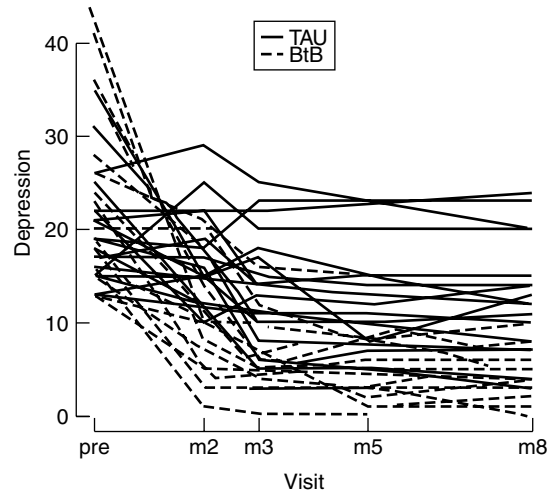


Figure 1 Plot of individual patient profiles for data from the BtB study

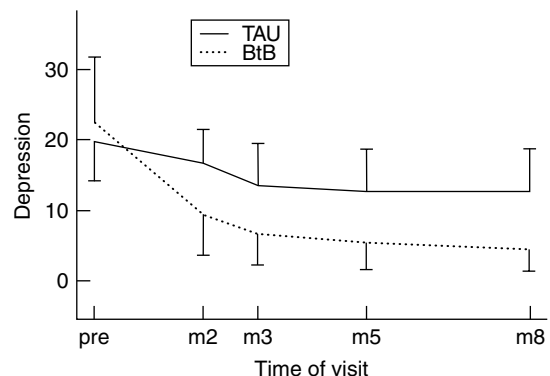


Figure 2 Plot of treatment profiles for both BtB and TAU groups

(but not simplistic) method for the analysis of longitudinal data is the so-called summary measure analysis process (*see Summary Measure Analysis of Longitudinal Data*). Here, the essential multivariate nature of the longitudinal data is transformed to univariate by the calculation of an appropriate summary of the response profile of each individual. For the BtB data, for example, we might use the mean of the available depression scores for each patient (*see Summary Measure Analysis of Longitudinal Data*).

There may, however, be occasions when there are more complex questions to be asked about

how response measures evolve over time and more problematic data collected than can be adequately handled by the summary measure technique. In such cases, there are a rich variety of regression type models now available that can be usefully employed, the majority of which are implemented in most of the major statistical software packages.

There are a number of desirable general features that methods used to analyze data from studies in which the outcome variable is measured at several time points should aim for – these include the following:

- the specification of the mean response profile needs to be sufficiently flexible to reflect both time trends within each treatment group and any differences in these time trends between treatments;
- repeated measurements of the chosen outcome are likely to be correlated rather than independent. The specification of the pattern of correlations or covariances of the repeated measurements by the chosen model needs to be flexible, but economical;
- the method of analysis should accommodate virtually arbitrary patterns of irregularly spaced time sequences within individuals.

Two very commonly used methods of dealing with longitudinal data, one based on **analysis of variance** (ANOVA), and one on **multivariate analysis of variance** (MANOVA) – (see **Repeated Measures Analysis of Variance** for details) – fail one or other of the criteria above. For an ANOVA repeated measures analysis to be valid, it is necessary both for the variances across time points to be the same, and the covariances between the measurements at each pair of time points to be equal to each other, the so-called compound symmetry requirement (see **Sphericity Test**). Unfortunately, it is unlikely that this compound symmetry structure will apply in general to longitudinal data, since observations from closely aligned time points are likely to be more highly correlated than observation made at time points further apart in time. Also variances often increase over time. The ANOVA approach to the analysis of longitudinal data fails the flexibility requirement of the second bulleted point above.

The MANOVA approach allows for *any* pattern of covariances between the repeated measurements, but

only by including too many parameters; MANOVA falls down on economy.

Fortunately, there are readily available alternatives to ANOVA and MANOVA for the analysis of longitudinal data that are both more satisfactory from a statistical viewpoint and far more flexible. These alternatives are essentially regression models that have two components, the first of which models the average response over time and the effect of covariates such as treatment group on this average response, and the second of which models the pattern of covariances between the repeated measures. Each component involves a set of parameters that have to be estimated from the data. In most applications, it is the parameters reflecting the effects of covariates on the average response that will be of most interest. The parameters defining the covariance structure of the observations will not, in general, be of direct concern (they are often regarded as so-called nuisance parameters). In [2], Diggle suggests that overparameterization (i.e., too many parameters, a problem that affects the MANOVA approach – see above) will lead to inefficient estimation and potentially poor assessment of standard errors for estimates of the mean response profiles; too restrictive a specification (too few parameters to do justice to the actual covariance structure in the data, a problem of the ANOVA method for repeated measures) may also invalidate inferences about the mean response profiles when the assumed covariance structure does not hold.

With a single value of the response variable available for each subject, modeling is restricted to the expected or average value of the response among persons with the same values of the covariates. But with repeated observations on each individual, there are other possibilities. In marginal models, (see **Marginal Models for Clustered Data**), interest focuses on the regression parameters of each response separately (see **Generalized Estimating Equations (GEE)**). But in longitudinal studies, it may also be of interest to consider the *conditional expectation* of each response, given either the values of previous responses or a set of random effects that reflect natural heterogeneity amongst individuals owing to unmeasured factors. The first possibility leads to what are called *transition models* – see [3], and the second to **linear multilevel models**, **nonlinear mixed models**, and **generalized linear mixed models**.

The distinction between marginal models and conditional models for longitudinal data is only

of academic importance for normally distributed responses, but of real importance for repeated nonnormal responses, for example, binary responses, where different assumptions about the source of the correlation can lead to regression coefficients with distinct interpretations (*see* **Generalized Linear Mixed Models**).

Dropouts

The design of most longitudinal studies specifies that all patients are to have the measurement of the outcome variable(s) made at a common set of time points leading to what might be termed a *balanced data set*. However, although balanced longitudinal data is generally the aim, it is rarely achieved in most studies carried out in psychiatry and psychology because of the occurrence of missing values, in the sense that intended measurements on some patients are not taken, are lost, or are otherwise unavailable. In longitudinal studies, missing values may occur intermittently, or they may arise because a participant ‘drops out’ of the study before the end as specified in the study plan. Dropping out of a study implies that once an observation at a particular time point is missing, so are all the remaining planned observations. Many studies will contain missing values of both types, although in practice it is missing values that result from participants dropping out that cause most problems when it comes to analyzing the resulting data set. Missing observations are a nuisance and the very best way to avoid problems with missing values is not to have any! If only a few values are missing, it is unlikely that these will cause any major difficulties for analysis. But in psychiatric studies, the researcher is often faced with the problem of how to

deal with a data set in which a substantial proportion of the intended observations are missing. In such cases, more thought needs to be given to how to best deal with the data. There are three main possibilities:

- discard incomplete cases and analyze the remainder – *complete case analysis*;
- *impute* or fill in the missing values and then analyze the filled-in data;
- analyze the incomplete data by a method that does not require a complete (rectangular) data set.

These possibilities are discussed in **missing data and dropouts in longitudinal studies: methods of analysis**.

References

- [1] Beck, A., Steer, R. & Brown, G. (1996). *BDI-II Manual*, 2nd Edition, The Psychological Corporation, San Antonio.
- [2] Diggle, P.J. (1988). An approach to the analysis of repeated measures, *Biometrics* **44**, 959–971.
- [3] Diggle, P.J., Heagerty, P.J., Liang, K.Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford.
- [4] Proudfoot, J., Goldberg, D., Mann, A., et al. (2003). Computerised, interactive, multimedia CBT reduced anxiety and depression in general practice: a RCT, *Psychological Medicine* **33**, 217–227.
- [5] Watkins, E. & Williams, R. (1998). The efficacy of cognitive behavioural therapy, *Journal of Counseling and Clinical Psychology* **27**, 31–39.

(*See also* **Age–Period–Cohort Analysis**; **Latent Transition Models**; **State Dependence**)

BRIAN S. EVERITT

Longitudinal Designs in Genetic Research

THOMAS S. PRICE

Volume 2, pp. 1102–1104

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Longitudinal Designs in Genetic Research

Longitudinal Experimental Designs

Experiments with a longitudinal design are often called **cohort** or *follow-up* studies. (see **Longitudinal Data Analysis**) They consist of repeated measurements of the same group of experimental subjects (see **Repeated Measures Analysis of Variance**). Longitudinal studies can be contrasted with *cross-sectional* or **case-control** designs, in which multiple measurements are made, but of different groups of subjects. Only longitudinal studies can test changes over time – or changes in response to some treatment or other experimentally controlled variable – in a particular population.

Longitudinal studies come with certain disadvantages. One problem is the age of the subject is confounded with the date of measurement, so that the real effects of age in the measurement sample are confounded with the changes in the whole population over time – so called *secular trends* or *cohort effects*. A second problem is that subjects may drop out of the study before all the measurements have been administered, giving rise to problems of **attrition** (see **Missing Data**).

Longitudinal Correlations and Causality

The correlation between two measurements in a longitudinal sample may be termed a *longitudinal correlation*, and is a measure of how well their relationship conforms to a linear dependency. A correlation between two variables is equivalent to their *covariance* between them, standardized by their *variances*. This encyclopedia entry only considers the kind of change that can be measured by correlations between variables. For information about longitudinal models that consider changes in mean levels as well as correlations between different measurements, see **growth curve models**.

The existence of a longitudinal correlation – or any other correlation – does not imply any particular mode of causality. Correlations can arise by one variable acting causally on the other (Figure 1); or by mutual causality (Figure 2);



Figure 1

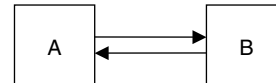


Figure 2

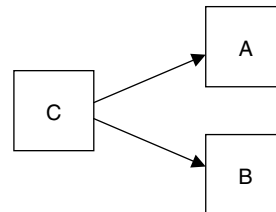


Figure 3

or through a shared dependency on a third, causally prior, variable (Figure 3).

When analyzing longitudinal datasets, it is often assumed *a priori* that temporally prior measurements also have causal priority, which ignores the possibility of an underlying shared dependency on a third, unmeasured variable. Whether there are adequate grounds for making this assumption is not a question that can be resolved by the data themselves.

Longitudinal Models as Multivariate Models

Behavioral genetic models typically assume that the variance of an observed variable can be divided up into genetic and environmental *components of variance*. In the case of multivariate behavioral genetic models (see **Multivariate Genetic Analysis**), the relationships among multiple variables are divided up into genetic and environmental pathways that together are responsible for the covariances among the variables. Longitudinal behavioral genetic models are a special case of these multivariate models: the covariances among repeated measurements in a longitudinal design are decomposed in the same

2 Longitudinal Designs in Genetic Research

way. So for example, just as in a standard multivariate behavioral genetic model, the degree of overlap between the genetic influences on two longitudinal measurements can be summarized by the genetic correlation (*see Correlation Issues in Genetics Research*). Likewise, the proportion of the longitudinal correlation between them that is attributable to this shared **heritability** is given by the bivariate heritability (*see Multivariate Genetic Analysis*).

Examples of Longitudinal Behavioral Genetic Models

The Cholesky decomposition. The **Cholesky** (or ‘triangular’) **decomposition** is a *saturated* model, meaning that it has the maximum possible number of parameters that can be estimated from the data. The model assumes that the factors influencing on the phenotype on any given measurement occasion continue to influence the phenotype on all subsequent occasions, and are causally prior to any new influences.

In terms of matrix algebra, if the observed covariance matrix is given by Σ , then the decomposition consists in finding the upper triangular matrix Λ such that

$$\Sigma = \Lambda \Lambda^T \quad (1)$$

In a behavioral genetic model, we can, for example, substitute for Λ the upper triangular matrices of genetic and environmental pathways, A and E , such that

$$\Sigma = AA^T + EE^T \quad (2)$$

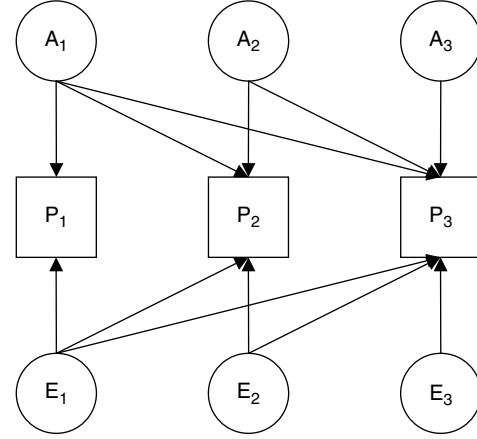


Figure 4

This model is represented in Figure 4 for phenotypes measured on three occasions. The uncorrelated latent genetic and environmental variables (A_i and E_j , respectively) act on the observed variables P_t wherever $i, j \leq t$.

The simplex model. The simplex model can be estimated whenever there are three or more measurement occasions. When there are four or more measurement occasions, or when further simplifying assumptions are made, it is an *unsaturated* model – that is, fewer parameters are estimated from the data than is theoretically possible – and therefore imposes a certain structure on the data. Its theoretical starting point is the assumption that the relationships among

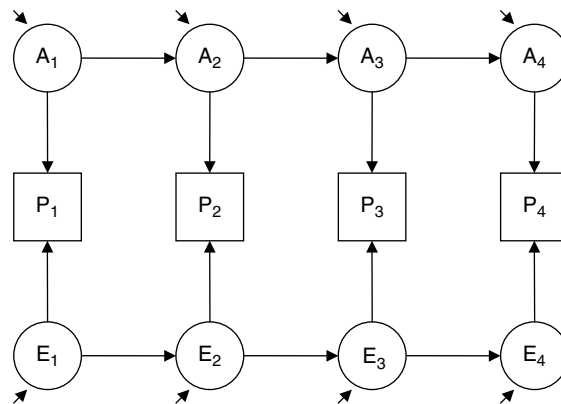


Figure 5

phenotypes measured longitudinally can be represented by chains of causal paths, known as *simplex coefficients*, linking the phenotypes on successive measurement occasions. The simplex model requires that the correlations among successive measurements be greater than those between widely spaced measurements. Thus, the correlation matrix between phenotypes conforming to a simplex structure contains coefficients that are greatest between successive measurements (those adjacent to the leading diagonal) and fall off as the measurements become spaced further apart (i.e., as the distance to the leading diagonal increases).

Figure 5 represents a behavioral genetic example for a phenotype P measured on four occasions. The simplex coefficients are represented by horizontal paths, modeling the continuity between successive

latent genetic (A) and environmental (E) variables. The slanted paths on the diagram represent new genetic and environmental variance on the phenotype, uncorrelated with previous measurements – the *innovations*. Note that the continuity between widely spaced latent variables (A₁ and A₄, say) is mediated by the latent variables on intermediate measurement occasions (in this case, A₂ and A₃).

In terms of matrix algebra, the covariance matrix Σ is a function of diagonal matrices of genetic and environmental innovations (I_A and I_E , respectively) and sub lower matrices of simplex coefficients (S_A and S_E):

$$\Sigma = I_A S_A I_A^T + I_E S_E I_E^T \quad (3)$$

THOMAS S. PRICE

Lord, Frederic Mather

RONALD K. HAMBLETON

Volume 2, pp. 1104–1106

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Lord, Frederic Mather

Born: November 12, 1912, in Hanover, USA.

Died: February 5, 2000, in Florida, USA.

Dr. Frederic Lord was described as the ‘Father of Modern Testing’ by the National Council on Measurement in Education in the United States when he received their career achievement award in 1991. The Educational Testing Service in the United States bestowed upon Dr. Lord their highest honor for ‘... the tremendous achievements and contributions, not just to the Educational Testing Service (ETS) but nationally and internationally, as the principal developer of the statistical underpinnings supporting modern mental testing, providing the theoretical structure for **item response theory** (IRT), which allowed for effective test equating and made possible modern computer-adaptive tests...’ The American Psychological Association in 1997 selected Dr. Lord to receive one of its achievement awards for a lifetime of seminal and important contributions to the field of test theory. And these are just three of many of the awards Dr. Lord receiving during his life. It is clear that Dr. Lord’s contributions to psychometric theory and practices place him among the most influential and productive psychometricians of the twentieth century.

Dr. Lord was born in Hanover, New Hampshire, in 1912, earning a BA degree in Sociology from Dartmouth College in 1936; a masters degree in Educational Psychology from the University of Minnesota in 1943; and finally achieving a PhD from Princeton in 1951 under the direction of Harold Gulliksen, another of the outstanding contributors to psychometric theory and practices in the twentieth century. Beginning in 1944, Dr. Lord gained some early experience in psychometric methods by working with the predecessor of the Educational Testing Service (ETS), the Graduate Record Office of the Carnegie Foundation, and he joined ETS when it was officially founded in 1949. Dr. Lord spent his whole career at ETS, except for sabbaticals that took him to several universities in the United States. He retired from ETS in 1982.

Dr. Lord will be forever influential in the field of psychometric methods because of three seminal publications and over 100 scholarly journal articles and

chapters. In 1952, his dissertation research introducing modern test theory, today called ‘Item Response Theory’, was published as *Psychometric Monograph No. 7* [2]. This publication was the first formal one that provided mathematical models for linking candidate responses on achievement test items to underlying constructs or abilities that were measured by a test. He followed up his monograph with papers in the 1953 editions of *Psychometrika* and *Educational and Psychological Measurement* that showed how IRT was linked to true score theory, and how **maximum likelihood** ability estimates could be obtained and confidence bands (*see Confidence Intervals*) about these ability estimates determined. But then Lord stopped this line of research for about 15 years, presumably because the computer power was not available to solve the complicated and numerous nonlinear equations needed for IRT model parameter estimation. Clearly, the research published by Lord in 1952 and 1953 did not satisfy him. He was forced to assume that ability scores were normally distributed, and that, for all practical purposes, guessing on multiple-choice achievement test items did not occur. These unrealistic assumptions allowed him to simplify the maximum likelihood equations and to solve them in 1952, but he was not happy with the assumptions he needed to make, and he was reluctant to move forward in his research with these flawed assumptions. He would need to wait for more computer power before moving forward with his IRT program of research.

For the next 15 years or so, Dr. Lord studied many psychometric problems such as those associated with equating, norming, measurement of change, formula scoring, and most importantly, true score theory. It was in this era that he built his reputation for scholarship and creative problem solving. His impact on ETS was huge, with many of their standard operating procedures, such as test score equating (*see Classical Test Score Equating*), being based on his research and suggestions. ETS then, and today, remains the mecca for testing theory and practices, and so his work at the time had an immense impact on the ways testing was carried out around the world.

In 1968, Dr. Lord, with the great assistance of Melvin Novick and Alan Birnbaum, produced his second seminal publication *Statistical theories of mental test scores*, which has had an immense impact on the testing field for over 30 years [4]. This book extended the work of Gulliksen’s *Theory of Mental*

Tests [1] by providing a statistical framework for developing tests and analyzing test scores. It was the culmination for Dr. Lord's efforts between 1952 and 1967 to extend the models and results of **classical test theory**. This book became required reading for a generation of psychometricians. I carried my copy of 'Lord and Novick', as we called it, just about everywhere I went for the first 10 years of my career!

The Lord–Novick text remains the standard reference for persons interested in learning about classical test theory. But the book will be long remembered for something else too – it was the first text that provided extended coverage of IRT: five chapters in total. With huge improvements in the speed with which computers could solve large numbers of nonlinear equations and with his keen interest in computer-adaptive testing, a topic that required IRT as a measurement model, Dr. Lord became committed again to advancing IRT. In the late 1960s, model parameter estimation became possible without a highly restrictive normal distribution assumption of ability. Also, Lord's colleague Allan Birnbaum in the late 1950s wrote a series of reports in which he substituted the logistic model for the normal ogive model as the basis for modeling data and introduced an additional parameter into his logistic models to handle the guessing problem that had worried Lord in the early 1950s. Lord contributed Chapter 16 to *Statistical Theories of Mental Test Scores*, and Birnbaum provided Chapters 17 to 20 [4]. Because of the influence of Lord and Novick's book along with the work of Georg Rasch, item response theory took off as an area for research and development.

Through the 1970s, Lord made breakthrough contributions to IRT topics such as equating, estimation of ability, the detection of potentially biased items, test development, and his favorite IRT application, **computer-adaptive testing**. All of this and more was included in his third seminal publication, *Applications of Item Response Theory to Practical Testing Problems* [3]. No book has been more important than this one in advancing the IRT field. Lord retired in 1982 but continued his research at ETS until his career was cut short by a serious car accident in South Africa in 1985. Dr. Lord spent the remaining years of his life in retirement with his wife, Muriel Bessemer, in Florida.

I will conclude with one brief personal story. In 1977, I invited Dr. Lord to speak at the annual meeting of the American Educational Research Association as a discussant of a monograph prepared by

a colleague and myself and three graduate students from the University of Massachusetts. I sent him the 130-page monograph, and soon after I received it back with editorial comments. This was unexpected. First, Dr. Lord felt we had been far too generous in our recognition of his contributions to the field of IRT and he wanted us to revise the monograph to more accurately reflect the work of his colleagues. We disagreed but revised the text to reflect Lord's wishes. Second, he felt that we had made some minor errors, and he preferred that our monograph be revised before any distribution. I think his comment was that 'the field would be less confused if there were fewer misunderstandings and so it would be in the best interests of everyone if the monograph were revised'. He wanted no recognition for his input to our work, he only wanted to be sure that the monograph be correct in every respect when it was read by researchers and students. When he delivered his remarks at the conference, he read them from a carefully prepared speech, not the usual kind of presentation given by a discussant working from rough notes, some jotted down during the presentations. All of the speakers and the conference attendees, and the conference room itself jammed with people wanting to hear Dr. Lord, hung on every word he had to say. He prepared his text in advance because he wanted to be sure that what he said was exactly what he wanted to say, no more, no less, and he wanted to deliver his message with great care and technical precision. What came through that day for me was that Dr. Lord was a man who cared about getting things right, exactly right, and he spoke clearly, thoughtfully, and with great precision about how to move the testing field forward. He was a superb example for all of us working in the assessment field.

References

- [1] Gulliksen, H.O. (1950). *Theory of Mental Tests*, Wiley, New York.
- [2] Lord, F.M. (1952). *A Theory of Test Scores*, Psychometric Monograph No. 7, Psychometric Society.
- [3] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.
- [4] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

RONALD K. HAMBLETON

Lord's Paradox

PAUL W. HOLLAND

Volume 2, pp. 1106–1108

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Lord's Paradox

Following Lord in [3], suppose a researcher wishes to compare two groups on some criterion variable, Y , and would like to make allowance for the fact that the groups being compared differ on some important independent variable X . The situation is such that the observed differences in Y might be caused by the differences in X , and the researcher wishes to rule out this possibility.

With measurements on Y and X made for each person in each of two groups, the thought of using the **analysis of covariance** is irresistible to some. Lord's intended purpose was to warn them that the interpretation of the results of such analyses could be problematic.

He illustrated his points with a simple example. Taking small liberties with Lord's language, suppose a university wants to investigate the effects of the diet provided in the dining halls and any differences in these effects between men and women. To this end, they collect the weight of each student at registration in the fall and again later at the end of the spring term. Two of the university's distinguished statisticians, Henry and Howard, offer to analyze the results, and they use different methods.

Henry studies the distribution of weight for each gender group and finds that while men on average weigh more than women, the average weight of each group is unchanged from fall and spring. The correlation between the fall and spring weights is both positive and substantial in each gender group. But, while the weights of individual students often change between the two time points, some go up and some go down. A sort of dynamic equilibrium appears to be maintained over the year within each gender group. Henry concludes that with respect to the distribution of weight nothing systematic has occurred, either for men or for women.

Howard sees this as a natural place to apply the analysis of covariance and uses **multiple regression** to do the calculations. There are two groups, men and women, and two measurements for each, with Y being the spring weight and X being the fall weight. He estimates a regression function that predicts Y from both X and an indicator variable M , that is 1 for males and 0 for females. (He also checks that if he adds the interaction or product term, $X \times M$,

as a predictor it has no extra contribution so that the regressions of Y on X are parallel for the two groups.) There is a significant and substantial positive coefficient on M . This coefficient is the regression-adjusted (for X) difference in spring weights between men and women. Howard correctly interprets this as the average amount more that a male will weigh in the spring, compared to a female whose fall weight is the same. Whatever the diet is doing to their weight, it is adding more weight to men than it does to equally weighing women! In [3], Lord gives a diagram showing that the results found by Henry and Howard are compatible.

The paradox is that while Henry seems reasonable in his conclusion that the diet did not have any systematic effect between the fall and the spring, either for men or women, Howard observes that indeed men will gain more on the diet than will equally weighing women.

Lord's explanation of the paradox is that 'with the data usually available in such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups'. In program evaluation, it is common to use nonrandomized observational studies in which treatment and control groups do exhibit 'uncontrolled preexisting differences', and therefore Lord's paradox quickly became part of the lore surrounding program evaluation, [4].

As it stands, there are several points of ambiguity in the language that is used to describe Lord's paradox. What is meant by 'making allowance for...'? What meaning should we attach to the idea that differences in X could cause differences in Y ? What exactly is the researcher trying to rule out? What are 'pre-existing group differences'? Will any treatment group do? Holland and Rubin give an extended explanation of Lord's paradox in [2], and my treatment is based on theirs but I will focus only on the example described above.

Asking about the effect of a treatment, in this case, the effect of the diet, is a causal question, and the causal effect of a treatment is always defined relative to another treatment, [1]. But, this example has only one treatment condition, that is, the diet that is used in the dining halls. There is no control diet here, yet if we would like to talk about the *effect* of the diet it must be relative to some other, control, diet. In the notation of the Neyman-Rubin model

[1], in order to make causal sense of the situation, there need to be two potential spring weights, Y_t and Y_c for each student. Y_t is the student's spring weight that results from the dining hall diet (t) and Y_c is the student's spring weight that would have resulted from the control diet (c), whatever that might be. In this example, only Y_t is observed. For each student, Y_c is **counterfactual reasoning** and we are left to speculate as to what Y_c is. It turns out that if they make causal inferences about the effect of the diet, Henry and Howard must have different Y_c 's in mind.

The causal effect of t relative to c on the spring weight of a student is the difference, $Y_t - Y_c$. The *average causal effect* (ACE) is the average or expectation of these individual causal effects over the population of students: $ACE = E(Y_t - Y_c) = E(Y_t) - E(Y_c)$. In this example, there is also an interest in differences in effects between men and women. Using the indicator variable, M , defined above, this means that not only are we interested in the ACE, we also want the values of the 'conditional' ACEs of the form: $ACE(\text{men}) = E(Y_t - Y_c | M = 1)$, and $ACE(\text{women}) = E(Y_t - Y_c | M = 0)$. If the effect of the diet is different for men and women, this will be reflected by differences between the average causal effects for men and for women.

With no clear control diet in hand, it is a daunting task to obtain values for either the ACE or for $ACE(\text{men})$ and $ACE(\text{women})$. However, most examples of causal inference must make some assumptions that are often not directly testable, and we may interpret causal interpretations of Henry's and Howard's analyses as based on different assumptions.

Henry assumes that $Y_c = X$, that is, the response to the control diet is the same as the student's fall weight. For Henry, the causal effect for any student is the gain score, $Y_t - X$. This is an entirely untestable assumption, but look where it gets Henry!

$$ACE_{\text{Henry}} = E(Y_t - Y_c) = E(Y_t) - E(X). \quad (1)$$

and

$$\begin{aligned} ACE_{\text{Henry}}(\text{men}) &= E(Y_t | M = 1) \\ &\quad - E(X | M = 1) = 0, \end{aligned} \quad (2)$$

$$\begin{aligned} ACE_{\text{Henry}}(\text{women}) &= E(Y_t | M = 0) \\ &\quad - E(X | M = 0) = 0. \end{aligned} \quad (3)$$

From his assumption, the effect of the diet is nil, both for men and for women.

Howard, on the other hand, assumes that $Y_c = a + bX$, where b is the slope on X from his estimated regression function. Again, this is an entirely untestable assumption, but look where it gets Howard!

$$\begin{aligned} ACE_{\text{Howard}} &= E(Y_t) - E(a + bX) \\ &= E(Y_t) - a - bE(X). \end{aligned} \quad (4)$$

and

$$\begin{aligned} ACE_{\text{Howard}}(\text{men}) &= E(Y_t | M = 1) \\ &\quad - a - bE(X | M = 1) \end{aligned} \quad (5)$$

$$\begin{aligned} ACE_{\text{Howard}}(\text{women}) &= E(Y_t | M = 0) \\ &\quad - a - bE(X | M = 0) \end{aligned} \quad (6)$$

so that the difference between the gender specific ACEs is

$$\begin{aligned} ACE_{\text{Howard}}(\text{men}) - ACE_{\text{Howard}}(\text{women}) &= E(Y_t | M = 1) - E(Y_t | M = 0) \\ &\quad - b[E(X | M = 1) - E(X | M = 0)]. \end{aligned} \quad (7)$$

The last expression is the mean difference between male and female spring weights that is regression-adjusted by their fall weight. This is the parameter estimated by the analysis of covariance.

Who is right? In an example like this, where the control condition is fictitious, it is hard to say. Perhaps some will favor Henry's assumption while others may prefer Howard's or some other assumption. It is certain that there is nothing in the data at hand to contradict either or any assumption about Y_c .

Another approach to resolving the paradox is simply to avoid making the causal inferences. This lowers our sights to descriptions of the data rather than attempting to draw causal inferences from them. From this point of view, there is just Y and X , no t or c , and both Henry's description of an apparent 'dynamic equilibrium' and Howard's description of spring weights for men and women conditional on fall weight are both equally valid. Both statisticians are 'right' but they are talking about very different things, and neither has any relevance to the *effects* of the dining hall diet.

References

- [1] Holland, P.W. (1986). Statistics and causal inference (with discussion), *Journal of the American Statistical Association* **81**, 945–970.
- [2] Holland, P.W. & Rubin, D.B. (1983). On Lord's paradox, in *Principals of Modern Psychological Measurement*, H. Wainer & S.J. Messick, eds, Erlbaum, Hillsdale, pp. 3–25.
- [3] Lord, F.M. (1967). A paradox in the interpretation of group comparisons, *Psychological Bulletin* **68**, 304–305.
- [4] Lord, F.M. (1973). Lord's paradox, in *Encyclopedia of Educational Evaluation*, S.B. Anderson, S. Ball, R. Murphy & associates, eds, Jossey-Bass, San Francisco.

Further Reading

- Lord, F.M. (1969). Statistical adjustments when comparing preexisting groups, *Psychological Bulletin* **72**, 336–337.

PAUL W. HOLLAND

Encyclopedia of Statistics in

BEHAVIORAL SCIENCE

Editors in Chief **Brian Everitt** **David Howell**

VOLUME

3

m-qu

VOLUME 3

M Estimators of Location	1109-1110	Median Test.	1193-1194
Mahalanobis Distance.	1110-1111	Mediation.	1194-1198
Mahalanobis, Prasanta Chandra.	1112-1114	Mendelian Genetics Rediscovered.	1198-1204
Mail Surveys.	1114-1119	Mendelian Inheritance and Segregation Analysis.	1205-1206
Malloves Cp Statistic.	1119-1120	Meta-Analysis.	1206-1217
Mantel-Haenszel Methods.	1120-1126	Microarrays.	1217-1221
Marginal Independence.	1126-1128	Mid-p Values.	1221-1223
Marginal Models for Clustered Data.	1128-1133	Minimum Spanning Tree.	1223-1229
Markov Chain Monte Carlo and Bayesian Statistics.	1134-1143	Misclassification Rates.	1229-1234
Markov Chain Monte-Carlo Item Response Theory Estimation.	1143-1148	Missing Data.	1234-1238
Markov Chains.	1149-1152	Model Based Cluster Analysis.	1238
Markov, Andrei Andreevich.	1148-1149	Model Evaluation.	1239-1242
Martingales.	1152-1154	Model Fit: Assessment of.	1243-1249
Matching.	1154-1158	Model Identifiability.	1249-1251
Mathematical Psychology.	1158-1164	Model Selection.	1251-1253
Maximum Likelihood Estimation.	1164-1170	Models for Matched Pairs.	1253-1256
Maximum Likelihood Item Response Theory Estimation.	1170-1175	Moderation.	1256-1258
Maxwell, Albert Ernest.	1175-1176	Moments.	1258-1260
Measurement: Overview.	1176-1183	Monotonic Regression.	1260-1261
Measures of Association.	1183-1192	Monte Carlo Goodness of Fit Tests.	1261-1264
Median.	1192-1193	Monte Carlo Simulation.	1264-1271
Median Absolute Deviation.	1193	Multidimensional Item Response Theory Models.	1272-1280
		Multidimensional Scaling.	1280-1289

Multidimensional Unfolding.	1289-1294	Neyman, Jerzy.	1401-1402
Multigraph Modeling.	1294-1296	Neyman-Pearson Inference.	1402-1408
Multilevel and SEM Approaches to Growth Curve Modeling.	1296-1305	Nightingale, Florence.	1408-1409
Multiple Baseline Designs.	1306-1309	Nonequivalent Control Group Design.	1410-1411
Multiple Comparison Procedures.	1309-1325	Nonlinear Mixed Effects Models.	1411-1415
Multiple Comparison Tests: Nonparametric and Resampling Approaches.	1325-1331	Nonlinear Models.	1416-1419
Multiple Imputation.	1331-1332	Nonparametric Correlation (r_s).	1419-1420
Multiple Informants.	1332-1333	Nonparametric Correlation (tau).	1420-1421
Multiple Linear Regression.	1333-1338	Nonparametric Item Response Theory Models.	1421-1426
Multiple Testing.	1338-1343	Nonparametric Regression.	1426-1430
Multi-trait Multi-method Analyses.	1343-1347	Nonrandom Samples.	1430-1433
Multivariate Analysis of Variance.	1359-1363	Nonresponse in Sample Surveys	1433-1436
Multivariate Analysis: Bayesian.	1348-1352	Nonshared Environment.	1436-1439
Multivariate Analysis: Overview.	1352-1359	Normal Scores & Expected Order Statistics.	1439-1441
Multivariate Genetic Analysis.	1363-1370	Nuisance Variables.	1441-1442
Multivariate Multiple Regression.	1370-1373	Number Needed to Treat.	1448-1450
Multivariate Normality Tests.	1373-1379	Number of Clusters.	1442-1446
Multivariate Outliers.	1379-1384	Number of Matches and Magnitude of Correlation.	1446-1448
Neural Networks.	1387-1393	Observational Study.	1451-1462
Neuropsychology.	1393-1398	Odds and Odds Ratios.	1462-1467
New Item Types and Scoring.	1398-1401	One Way Designs: Nonparametric and Resampling Approaches.	1468-1474

Optimal Design for Categorical Variables. 1474-1479	Pie Chart. 1547-1548
Optimal Scaling. 1479-1482	Pitman Test. 1548-1550
Optimization Methods. 1482-1491	Placebo Effect. 1550-1552
Ordinal Regression Models. 1491-1494	Point Biserial Correlation. 1552-1553
Outlier Detection. 1494-1497	Polychoric Correlation. 1554-1555
Outliers. 1497-1498	Polynomial Model. 1555-1557
Overlapping Clusters. 1498-1500	Population Stratification. 1557
P Values. 1501-1503	Power. 1558-1564
Page's Ordered Alternatives Test. 1503-1504	Power Analysis for Categorical Methods. 1565-1570
Paired Observations, Distribution Free Methods. 1505-1509	Power and Sample Size in Multilevel Linear Models. 1570-1573
Panel Study. 1510-1511	Prediction Analysis of Cross- Classifications. 1573-1579
Paradoxes. 1511-1517	Prevalence. 1579-1580
Parsimony/Occham's Razor. 1517-1518	Principal Component Analysis. 1580-1584
Partial Correlation Coefficients. 1518-1523	Principal Components and Extensions. 1584-1594
Partial Least Squares. 1523-1529	Probability Plots. 1605-1608
Path Analysis and Path Diagrams. 1529-1531	Probability: An Introduction. 1600-1605
Pattern Recognition. 1532-1535	Probits. 1608-1610
Pearson Product Moment Correlation. 1537-1539	Procrustes Analysis. 1610-1614
Pearson, Egon Sharpe. 1536	Projection Pursuit. 1614-1617
Pearson, Karl. 1536-1537	Propensity Score. 1617-1619
Percentiles. 1539-1540	Proscriptive and Retrospective Studies. 1619-1621
Permutation Based Inference. 1540-1541	Proximity Measures. 1621-1628
Person Misfit. 1541-1547	Psychophysical Scaling. 1628-1632

Qualitative Research.	1633-1636
Quantiles.	1636-1637
Quantitative Methods in Personality Research.	1637-1641
Quartiles.	1641
Quasi-Experimental Designs.	1641-1644
Quasi-Independence.	1644-1647
Quasi-Symmetry in Contingency Tables.	1647-1650
Quetelet, Adolphe.	1650-1651

M Estimators of Location

RAND R. WILCOX

Volume 3, pp. 1109–1110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

M Estimators of Location

Under normality, the sample mean has a lower standard error than the mean and median. A consequence is that hypothesis testing methods based on the mean have more **power**; the probability of rejecting the null hypothesis is higher versus using the median. But under nonnormality, there are general conditions under which this is no longer true, a result first derived by **LaPlace** over two centuries ago. In fact, any method based on means can have poor power.

This raises the issue of whether an alternative to the mean and median can be found that maintains relatively high power under normality but continues to have high power in situations in which the mean performs poorly. Three types of estimators aimed at achieving this goal have received considerable attention: M-estimators, L-estimators, and R-estimators. L-estimators contain trimmed and Winsorized means, and the median, as special cases (*see Trimmed Means; Winsorized Robust Measures*).

To describe M-estimators, first consider the **least squares** approach to estimation. Given some data, how might we choose a value, say c , that is typical of what we observe? The least squares approach is to choose c so as to minimize the sum of the squared differences between the observations and c . In symbols, if we observe $X_1, \dots, X_n$, the goal is to choose c so as to minimize

$$\sum (X_i - c)^2.$$

This is accomplished by setting $c = \bar{X}$, the sample mean. If we replace squared differences with absolute values, we get the median instead. That is, if the goal is to choose c so as to minimize

$$\sum |X_i - c|,$$

the answer is the usual sample median. But the sample mean can have a relatively large standard error under small departures from normality [1, 3–6], and the median performs rather poorly if sampling is indeed from a normal distribution. So an issue of some practical importance is whether some measure of the difference between c and the observations can be found that not only performs relatively well under normality but also continues to perform well in situations where the mean is unsatisfactory.

Several possibilities have been proposed; see, for example, [2–4]. One that seems to be particularly useful was derived by Huber [2]. There is no explicit equation for computing his estimator, but there are two practical ways of dealing with this problem. The first is to use an iterative estimation method. Software for implementing the method is easily written and is available, for example, in [6]. The second is to use an approximation of this estimator that inherits its positive features and is easier to compute. This approximation is called a one-step M-estimator.

In essence, a one-step M-estimator searches for any **outliers**, which are values that are unusually large or small. This is done using a method based on the median of the data plus a measure of dispersion called the **Median Absolute Deviation** (MAD). (Outlier detection methods based on the mean and usual standard deviation are known to be unsatisfactory; see [5] and [6].) Then, any outliers are discarded and the remaining values are averaged. If no outliers are found, the one-step M-estimator becomes the mean. However, if the number of outliers having an unusually large value differs from the number of outliers having an unusually small value, an additional adjustment is made in order to achieve an appropriate approximation of the M-estimator of location. There are some advantages in ignoring this additional adjustment, but there are negative consequences as well [6].

To describe the computational details, consider any n observations, say $X_1, \dots, X_n$. Let M be the median and compute

$$|X_1 - M|, \dots, |X_n - M|.$$

The median of these n differences is MAD, the **Median Absolute Deviation**. Let $\text{MADN} = \text{MAD}/0.6745$, let i_1 be the number of observations X_i for which $(X_i - M)/\text{MADN} < -K$, and let i_2 be the number of observations such that $(X_i - M)/\text{MADN} > K$, where typically $K = 1.28$ is used to get a relatively small standard error under normality. The one-step M-estimator of location (based on Huber's Ψ) is

$$\hat{\mu}_{\text{os}} = \frac{K(\text{MADN})(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2}. \quad (1)$$

2 M Estimators of Location

Computing a one-step M-estimator (with $K = 1.28$) is illustrated with the following ($n = 19$) observations:

77 87 88 114 151 210 219 246 253 262 296 299
306 376 428 515 666 1310 2611.

It can be seen that $M = 262$ and that $MADN = MAD/0.6745 = 114/0.6745 = 169$. If for each observed value we subtract the median and divide by MADN, we get

-1.09 -1.04 -1.035 -0.88 -0.66 -0.31
-0.25 -0.095 -0.05 0.00 0.20 0.22 0.26
0.67 0.98 1.50 2.39 6.2 13.90

So there are four values larger than the median that are declared outliers: 515, 666, 1310, 2611. That is, $i_2 = 4$. No values less than the median are declared outliers, so $i_1 = 0$. The sum of the values not declared outliers is

$$77 + 87 + \cdots + 428 = 3411.$$

So the value of the one-step M-estimator is

$$\frac{1.28(169)(4 - 0) + 3411}{19 - 0 - 4} = 285.$$

References

- [1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.
- [2] Huber, P.J. (1964). Robust estimation of location parameters, *Annals of Mathematical Statistics* **35**, 73–101.
- [3] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [4] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [5] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.
- [6] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

RAND R. WILCOX

Mahalanobis Distance

CARL J. HUBERTY

Volume 3, pp. 1110–1111

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mahalanobis Distance

It may be recalled from studying the Pythagorean theorem in a geometry course that the (Euclidean) distance between two points, (x_1, y_1) and (x_2, y_2) (in a two-dimensional space), is given by

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

From a statistical viewpoint, the two variables involved are uncorrelated; also, each variable has a standard deviation of 1.0. It was in 1936 that a statistician from India, **Prasanta C. Mahalanobis**, (1893–1972) introduced a generalization of the distance concept while investigating anthropometric problems [1]. It is a generalization in the sense of dealing with more than two variables whose inter-correlations may range from -1.0 to 1.0 , and whose standard deviations may vary.

Let there be a set of p variables. For a single analysis unit (or, subject), let $\mathbf{x}_i$ denote a vector of p variable scores for unit i . Also, let $\mathbf{S}$ denote the $p \times p$ **covariance matrix** that reflects the p variable interrelationships. Then, the Mahalanobis distance between unit i and unit j is given by

$$D_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \quad (2)$$

where the prime denotes a matrix transpose and $\mathbf{S}^{-1}$ denotes the inverse of $\mathbf{S}$ – the inverse standardizes the distance. (The radicand is a $(1 \times p)(p \times p)(p \times 1)$ triple product that results in a scalar.) Geometrically, $\mathbf{x}_i$ and $\mathbf{x}_j$ represent points for the two units in a p -dimensional space and D_{ij} represents the distance between the two points (in the p -dimensional space).

Now, suppose there are two groups of units. Let $\bar{\mathbf{x}}_1$ represent the vector of p means in group 1; similarly for group 2, $\bar{\mathbf{x}}_2$. Then, the distance between the two

mean vectors (called *centroids*) is analogously given by the scalar

$$D_{12} = \sqrt{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}. \quad (3)$$

This D_{12} represents the distance between the two group centroids. The $\mathbf{S}$ matrix used is generally the two-group error (or, pooled) covariance matrix.

A third type of Mahalanobis distance is that between a point representing an individual unit and a point representing a group centroid. The distance between unit i and the group 1 centroid is given by

$$D_{i1} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_1)' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_1)}. \quad (4)$$

The $\mathbf{S}$ matrix here may be the covariance matrix for group 1 or, in the case of multiple groups, the error covariance matrix based on all of the groups.

In sum, there are three types of Mahalanobis distance indices:

- D_{ij} – unit-to-unit,
- D_{12} – group-to-group, and
- D_{i1} – unit-to-group.

These distance indices are utilized in a little variety of multivariate analyses. A summary of some uses is provided in Table 1.

As alluded to earlier, a Mahalanobis D value may be viewed as a standardized distance. Jacob Cohen (1923–1998) in 1969 [2]. The *Cohen d*, is applied in a two-group, one outcome-variable context. Let x denote the outcome variable, s denote the standard deviation of one group of x scores (or the error standard deviation for the two groups), and $\bar{x}_1$ denote the mean of the outcome variable scores for group 1. The Cohen index, then, is

$$d_{12} = \frac{\bar{x}_1 - \bar{x}_2}{s} = (\bar{x}_1 - \bar{x}_2)s^{-1} \quad (5)$$

Table 1 The use of distance indices in some multivariate analyses

	Unit-to-unit	Group-to-group	Unit-to-group
Hotellings T^2 ; multivariate analysis of variance (MANOVA) Contrasts		X	
Predictive discriminant analysis		X	X
Cluster analysis	X		X
Pattern recognition			X
Multivariate outlier detection			X

2 Mahalanobis Distance

which is a special case of D_{12} . This d_{12} is sometimes considered as an *effect size* index (see **Effect Size Measures**).

References

- [1] Mahalanobis, P.C. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Science, Calcutta* **2**, 49–55.

- [2] Cohen, J. (1969). *Statistical Power for the Behavioral Sciences*, Academic Press, New York.

(See also **Hierarchical Clustering**; **k-means Analysis**; **Multidimensional Scaling**)

CARL J. HUBERTY

Mahalanobis, Prasanta Chandra

GARETH HAGGER-JOHNSON

Volume 3, pp. 1112–1114

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mahalanobis, Prasanta Chandra

Born: June 29, 1893, in Calcutta, [1].

Died: June 28, 1972, in Calcutta, [1].

Mahalanobis once invoked a mental construct of ‘five concentric circles’ to characterize the domains of science and statistics [3]. At the center is physics and at the outer layers are survey methods or areas where the variables are mostly unidentifiable, uncontrollable and ‘free’. As a student of physics and mathematics at Cambridge [4], he began his own career at the central geometric point of this construct. By the end of his life, he had made theoretical and applied contributions to every sphere. In 1922, he became professor of Physics at the Presidency College in Calcutta [7]. As ‘a physicist by training, statistician by instinct and a planner by conviction’ [9], his interests led him to found the Indian Statistical Institute (ISI) in 1931 [10] to promote interdisciplinary research. In 1933, he launched the internationally renowned journal *Sankyā*, serving as editor for forty years. Among Mahalanobis’ various achievements at the ISI were the establishment of undergraduate, postgraduate and Ph.D. level courses in statistics, and the reform in 1944 to make the ISI fully coeducational [5, 10]. His work made Delhi the ‘Mecca’ of statisticians, economists and planners world-wide [9]. At the national level, the Professor was chair of the National Income Committee set up by the Government of India in 1949 [1]. Mahalanobis was involved in the establishment of the Central Statistical Organization, the Perspective Planning Division in the Planning Commission and the National Sample Survey (perhaps his biggest achievement) [2]. In 1954, Prime Minister Nehru enlisted Mahalanobis to plan studies at the ISI to inform India’s second and third Five Year Plans [10] and he became honorary Statistical Adviser to the Government [4]. To economists, this was an important contribution [3, 7] and it is noteworthy that his lack of training in economics did not prevent him from undertaking this major responsibility [10]. He disliked accepted forms of knowledge and enjoyed problems that offered sufficient subject challenge, regardless of the topic [7]. When dealing with economic problems, he was used to saying that

he was glad he was not exposed to formal education in economics [8]! He developed the *Mahalanobis Model*, which was initially rejected by the planning commission. However, Nehru approved the model and their combined efforts have been described as ‘the golden age of planning’ [3]. The ISI was declared an Institute of national importance by the Parliament in 1959 with guaranteed funding. It had previously survived on project and *ad hoc* grants. The Indian Statistical Service (ISS) was established in 1961 [10] and the Professor was elected Fellow of the Royal Society of Great Britain in 1945 for his outstanding contributions to theoretical and applied statistics [9]. Other international awards included Honorary Member of the Academy of Sciences in the USSR and Honorary Member of the American Academy of Sciences. He was frequently called upon to collaborate with scientific research and foreign scientists [2, 4]. The ISI itself worked with scientists from the USSR, UK, USA and Australia.

The Professor’s contributions to the theory and practice of large-scale surveys have been the most celebrated [7]. His orientation in physics served as a starting point [8]. The three main contributions were [7]: (a) to provide means of obtaining an optimal sample design, which would either minimize the sampling variance of the estimate for a given cost, or minimize the cost for a given standard error of the estimate; (b) to show how one or more pilot surveys could be utilized to estimate the parameters of the variance and cost functions; (c) to suggest and use various techniques for measurement and control of sampling and nonsampling errors. He developed ‘interpenetrating network of samples in sample surveys’, which can help control for observational errors and judge the validity of survey results. The technique can also be used to measure variation across investigators, between different methods of data collection and input, and variation across seasons [7]. Total variation is split into three components: sampling error, ascertainment error, and tabulation error. In modern terminology, Mahalanobis’ four types of sampling (unitary unrestricted, unitary configurational, zonal unrestricted, and zonal configuration) correspond to unrestricted simple random sampling, unrestricted cluster sampling, stratified simple random sampling, and stratified cluster sampling [8] (*see Survey Sampling Procedures*). His watchwords were randomization, statistical control, and cost [8]. However, he also believed that samples should be independently

investigated and analyzed, or at the very least be split into two subsamples for analysis, despite the increase in cost.

The statistic **Mahalanobis distance** is perhaps the most widely known aspect of Mahalanobis' work today. This statistic is used in problems of taxonomical classification [10], in cluster analysis (*see Cluster Analysis: Overview; Hierarchical Clustering*) and for detecting multivariate **outliers** in datasets. It is helpful when drawing inferences on interrelationships between populations and speculating on their origin, or for measuring divergence between two or more groups. Alternative measures of quantitative characteristics such as social class can be compared using it [5]. **Pearson** rejected a paper on this area of work for the journal *Biometrika*, but **Fisher** soon recognized the importance of the work and provided the term *Mahalanobis D^2* . In later life, Mahalanobis developed *Fractile Graphical Analysis* (FGA), a nonparametric method for the comparison of two samples. For example, it can be used to compare the characteristics of a group at different time-points, or two groups at different places. It can also be used to test the normality or log normality or a frequency distribution [6]. The Professor also developed educational tests, studies of the correlations between intelligence or aptitude tests and success in school leaving certificates and other examinations. He made important contributions to factor analysis, which appear in early volumes of *Sankya*. His work on soil heterogeneity led him to meet Fisher and they became friends. They shared views on the foundations and the methodological aspects of statistics, and also on the role of statistics as a new technology.

Perhaps unusually, Mahalanobis had numerous interests outside his scientific pursuits. He was interested in art, poetry (particularly of Rabindranath Tagore) and literature [7], anthropology [11], and architecture. The various social, cultural, and intellectual movements in Bengal were also sources of interest. He was not 'an ivory-tower scholar or arm-chair intellectual', and 'hardly any issue of the time failed to make him take a stand one way or the other' [7]. Economic development was one of many broader objectives he believed in: including social change, modernization, national security and international peace [11]. Because he was on such good terms with Indian and world politicians [10], it has been said that Mahalanobis was the only scientist in the world who, when the Cold War was at its height,

was received with as much warmth in Washington and London as in Moscow and Beijing.

A genius for locating talent, Mahalanobis' approach to recruitment prevented a 'brain drain' from the ISI. He employed a strategy called 'brain irrigation', paying low salaries for posts earmarked for individual people [11]. When they left, the post disappeared. He felt that job security bred inefficiency and used this technique as a screening, quality control mechanism. His personality was unmistakably conscientious and driven, but a flair for argument and an impatience for bureaucracy 'made the Professor a fighter all his life' [3]. He believed in the public sector but not the typical civil servant [11] who he saw as inefficient, with little idea of the function and use of science. He struggled with governmental bureaucrats continuously in order to retain the autonomy of the ISI. He could talk with great effectiveness in small groups, using his histrionic talents to command attention, but was less effective in larger gatherings [6]. His character has been summarized as tough, courageous, tenacious, bold [6], intellectual, dynamic, devoted, loving, proud [3], odd (particularly in his attitude to money), but periodically despondent and depressed [11]. Four principal qualities outlined by Chatterjee [3] were: (a) practical mindedness: a preference for things tangible rather than abstract; (b) breadth of vision and farsightedness, a knack for looking beyond problems to envisage the long-term implications of possible solutions; (c) extraordinary organizing ability, although his wife Rani helped alleviate occasional absent mindedness [11]; (d) an innate sense of humanism and nationalism: strengthened in later life by his contacts with Nehru and Tagore [7]. The resourcefulness of his personality and the variety of his works set Mahalanobis apart from other scientists of his time [5].

Many types of research were welcomed at the ISI, but Rudra [11] recalled the Professor saying, 'There is one kind of research that I shall not allow to be carried out at the Institute. I will not support anybody working on problems of aeronavigation in viscous fluid'. He therefore did not approve of research that neither had any contribution to pure theory nor to solving practical problems. Knowledge was for socially useful purposes, and this conviction found expression in the later phases of his life [7]. Mahalanobis paid almost equal attention to both theoretical and applied statistical research.

According to him, statistics was an applied science: its justification centered on the help it can give in solving a problem [7].

References

- [1] Bhattacharyya, N. (1996a). Professor Mahalanobis and large scale sample surveys, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 34–56.
- [2] Bhattacharyya, D. (1996b). Introduction, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. ix–xv.
- [3] Chatterjee, S.K. (1996). P.C. Mahalanobis's journey through statistics to economic planning, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 99–109.
- [4] Deshmukh, C.D. (1965). Foreword, in *Contributions to Statistics. Presented to Professor P. C. Mahalanobis on the Occasion of his 70th Birthday*, C.R. Rao & D.B. Lahiri, eds, Statistical Publishing Society/Pergamon Press, Calcutta, Oxford.
- [5] Goon, A.M. (1996). P.C. Mahalanobis: scientist, activist and man of letters, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 89–98.
- [6] Iyengar, N.S. (1996). A tribute: Professor P.C. Mahalanobis and fractile graphical analysis, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 57–62.
- [7] Mukherjee, M. (1996). The Professor's demise: our reactions, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 6–14.
- [8] Murthy, M.N. (1965). On Mahalanobis' contributions to the development of sample survey theory and methods, in *Contributions to Statistics. Presented to Professor P. C. Mahalanobis on the Occasion of his 70th Birthday*, C.R. Rao & D.B. Lahiri, eds, Statistical Publishing Society/Pergamon Press, Calcutta, Oxford, pp. 283–315.
- [9] Rao, C.R. (1973). Prasanta Chandra Mahalanobis, *Biographical Memoirs of Fellows of the Royal Society* **19**, 455–492.
- [10] Rao, C.R. (1996). Mahalanobis era in statistics, in *Science, Society and Planning. A Centenary Tribute to Professor P.C. Mahalanobis*, D. Bhattacharyya, ed., Progressive Publishers, Calcutta, pp. 15–33.
- [11] Rudra, A. (1996). *Prasanta Chandra Mahalanobis. A Biography*, Oxford University Press, Delhi, Oxford.

Further Reading

- Mahalanobis, P.C. (1944). On large scale sample surveys, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **231**, 329–451.
- Mahalanobis, P.C. (1946a). Recent experiments in statistical sampling in the Indian Statistical Institute, *Journal of the Royal Statistical Society. Series A* **109**, 325–378.
- Mahalanobis, P.C. (1946b). Use of small-size plots in sample surveys for crop yields, *Nature* **158**, 798–799.
- Mahalanobis, P.C. (1953). On some aspects of the Indian National Sample Survey, *Bulletin of the International Statistical Institute* **34**, 5–14.
- Mahalanobis, P.C. & Lahiri, D.B. (1961). Analysis of errors in censuses and surveys with special emphasis on the experience in India, *Bulletin of the International Statistical Institute* **38**, 401–433.

GARETH HAGGER-JOHNSON

Mail Surveys

THOMAS W. MANGIONE

Volume 3, pp. 1114–1119

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mail Surveys

A mailed survey is one of several methods of collecting data that researchers can use to answer questions from a sample of the population. Mailed surveys usually involve the research team mailing a questionnaire to a potential respondent who then fills it out and returns the survey by mail. The major advantage of the mailed survey approach is its relatively low cost for data collection compared to telephone surveys or in-person interview surveys. A disadvantage of mailed surveys is that they often achieve a much lower response rate – percent of persons returning the survey from all of those asked to fill out the survey – than other data collection methods. Research studies conducted over the past few decades, however, have found ways to improve response rates to mailed surveys in many situations [5, 17–19, 44].

One general way to improve response rates to mailed surveys is to use this methodology only when it is appropriate. Mailed surveys are a good data collection choice when:

1. the budget for the study is relatively modest;
2. the sample of respondents are widely distributed geographically;
3. the data collection for a large sample needs to be completed in a relatively short time frame;
4. the validity of the answers to the questions would be improved if respondents could answer questions at their own pace;
5. the extra privacy of not having to give the answers to an interviewer would improve the veracity of the answers;
6. the study has a modest number of questions; and
7. the research sample has a moderate to high interest in the survey topic.

All mailed survey studies should incorporate three basic elements. The study mailing should include a well-crafted respondent letter, a preaddressed and postage-paid return envelope, and include a promise of confidentiality of answers or, preferably, anonymity of answers.

How the respondent letter is written is important because it is usually the sole mechanism for describing the study's purpose, explaining the procedures to be followed, and motivating the respondent to participate [2, 12, 43, 46, 74]. The following features

contribute to a well-crafted letter: it is not too long (limit to one page if possible); it begins with an engaging sentence; it clearly tells the respondent why the study is important; it explains who the sample is and how people were selected; it explains how confidentiality will be maintained; it indicates that participation is voluntary but emphasizes the importance of participation; it is printed on letterhead that clearly identifies the research institution; it tells the respondent how to return the survey; and it is easy to read in terms of type size, layout, and language level.

Early studies of mail surveys showed that including a preaddressed and postage-paid envelope is critically important to the success of a mail survey [3, 6, 38, 61, 67, 81]. Interestingly, research also has shown that using pretty commemorative stamps on both the initial delivered package and the return envelope improve response rates slightly [41, 50, 59].

Confidentiality is provided by not putting names on the surveys but instead using an ID number. Furthermore, confidentiality is maintained by keeping returned surveys under lock and key, keeping the list which links ID numbers to names in a separate locked place or password protected file, and presenting the data in reports in such a way that individuals are not identified. Anonymity can be achieved by not putting an ID number or other identifier on the surveys so that when they are returned the researcher does not know who returned it. Using this procedure, however, makes it difficult to send reminders to those who did not return their survey [7, 8, 14, 15, 25, 31, 34, 35, 54, 62, 66, 69, 78].

Beyond these three basic elements, there are two major strategies for improving response rates – sending reminders to those who have not responded and providing incentives to return the survey. The goal of a reminder is to increase motivation to respond. Reminders are best sent just to those who have not returned the survey so that the language of the reminder letter can be focused on those who have not returned their survey. Reminders should be sent out approximately 10 to 14 days after the previous mailing. This interval is not too short and hence will not waste a reminder on someone who intends to return the survey. Also, the interval is not too long, so that nonparticipating respondents will still remember what this is about. The first reminder should just be a postcard or letter and encourage the respondent to complete their survey. The second reminder should include answers to probable concerns the respondents

2 Mail Surveys

might have in not returning the survey as well as a replacement copy of the survey itself. The third reminder should again be a letter or postcard and the content should focus on the message that this is the last chance to participate. Some studies alter the last reminder by using a telephone reminder or delivering the letter with some type of premium postage like special delivery or overnight mail [16, 20, 23, 24, 26, 28, 33, 46, 49, 51, 53, 56, 81].

There is one important variation on these methods if you want to both provide anonymity to respondents and be able to send reminders to those who have not responded. The surveys are sent without any sort of ID number on them. However, in addition, the initial mailed packet includes a postcard. The person's name or ID number are printed on the postcard. On the back of the postcard is a message from the respondent to the study director. It says, 'I am returning my survey, so I do not need any more reminders'. The respondents are instructed in the respondent letter and maybe also at the end of the survey, to return the survey *and* the postcard, *but* to return them separately in order to maintain anonymity. By using this dual return mechanism, study directors can send reminders only to those who have not returned their surveys while at the same time maintaining anonymity of the returned surveys. The major concern in using this procedure is that many people will return the postcard but *not* the surveys. This turns out not to be the case. In general, about 90% of those returning their surveys also return a postcard. Almost certainly there are some people among the 90% who do not return their surveys, but the proportion is relatively small.

Incentives are the other strategy to use to improve response rates. Lots of different incentive materials have been used such as movie tickets, ball-point pens, coffee mugs, and cash or checks. The advantage of money or checks is that the value of the incentive is clear. On the other hand, other types of incentives may be perceived of having value greater than what was actually spent [22, 29, 32, 40, 42, 51, 56, 73, 81, 82].

There are three types of incentives based on who gets them and when. The three types are 'promised-rewards', 'lottery awards', and 'up-front rewards'. Promised rewards set up a 'contract' with the potential respondent such as: 'each person who sends back the completed survey will be paid \$5'. Although improvements in response rates compared to no incentive studies have been observed for

this approach, the improvements are generally modest [82].

Lottery methods are a form of promised-reward, but with a contingency – only one, or a few, of the participating respondents will be randomly selected to receive a reward. This method also differs from the promised rewards method in that the respondents who are selected receive a relatively large reward – \$100 to \$1000 or sometimes more. Generally speaking, the effectiveness of the lottery method over the basic promised-reward strategy depends on the respondents' perceptions of the chances of winning and the attractiveness of the 'prize' [36, 42, 58].

Up-front incentives work the best. For this method, everyone who is asked to participate in a study is given a 'small' incentive that is enclosed with the initial mailing. The incentive is usually described as a 'token of our appreciation for your participation'. Everyone can keep the incentive whether or not they respond. This unconditional reward seems to establish a sense of trust and admiration for the institution carrying out the study and thereby motivates respondents to 'help out these good people' by returning their survey. The size of the up-front incentive does not have to be that large to produce a notable increase in response rates. It is common to see incentives in the \$5–\$10 range [6, 37, 72, 79].

To get the best response rates, it is recommended that researchers use both reminders and incentives in their study designs [47]. When applying both procedures, it is not unusual to achieve response rates in the 60 to 80% range. By comparison, studies that only send out a single mailing with no incentives achieve response rates of 30% or less.

In addition to these two major procedures to improve response rates, there are a variety of other things that can be done by the researcher to achieve a small improvement in response rates. For example, somewhat better response rates can be achieved by producing a survey that looks professional and is laid out in a pleasing manner, has a reading level that is not too difficult, and includes instructions that are easy to follow.

Past research has shown that response rates can be increased by:

1. Using personalization in correspondence such as inserting the person's name in the salutation or by having the researcher actually sign the letter in ink [2, 11, 21, 30, 45, 52, 54, 55, 70, 74.]

2. Sending a prenotification letter a week or two before the actual survey is sent out to 'warn' the respondent of their selection as a participant and to be on the look-out for the survey [1, 9, 27, 33, 39, 48, 54, 63, 65, 71, 75, 77, 80, 81].
3. Using deadlines in the respondent letters to give respondents a heightened sense of priority to respond. Soft deadlines such as 'please respond within the next two weeks so we don't have to send you a reminder' provides a push without preventing the researcher from sending a further reminder [34, 41, 51, 56, 64, 68, 76].
4. Using a questionnaire of modest length, say 10 to 20 pages, rather than an overly long surveys will improve response rates (see **Survey Questionnaire Design**). Research has produced mixed results concerning the length of a survey. Obviously shorter surveys are less burdensome; but longer surveys may communicate a greater sense of the importance of the research issue. In general, the recommendation is to include no more material than is necessary to address the central hypotheses of the research [4, 10, 12, 13, 57, 60, 73].

Mail surveys when used appropriately and conducted utilizing procedures that have been shown to improve response rates, offer an attractive alternative to more expensive telephone or in-person interviews.

References

- [1] Allen, C.T., Schewe, C.D. & Wijk, G. (1980). More on self-perception theory's foot technique in the pre-call/mail survey setting, *Journal of Marketing Research* **17**, 498–502.
- [2] Andreasen, A.R. (1970). Personalizing mail questionnaire correspondence, *Public Opinion Quarterly* **34**, 273–277.
- [3] Armstrong, J.S. & Lusk, E.J. (1987). Return postage in mail surveys, *Public Opinion Quarterly* **51**, 233–248.
- [4] Berdie, D.R. (1973). Questionnaire length and response rate, *Journal of Applied Psychology* **58**, 278–280.
- [5] Biemer, P.R., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. & Sudman, S. (1991). *Measurement Errors in Surveys*, John Wiley & Sons, New York.
- [6] Blumberg, H.H., Fuller, C. & Hare, A.P. (1974). Response rates in postal surveys, *Public Opinion Quarterly* **38**, 113–123.
- [7] Boek, W.E. & Lade, J.H. (1963). Test of the usefulness of the postcard technique in a mail questionnaire study, *Public Opinion Quarterly* **27**, 303–306.
- [8] Bradt, K. (1955). Usefulness of a postcard technique in a mail questionnaire study, *Public Opinion Quarterly* **19**, 218–222.
- [9] Brunner, A.G. & Carroll, Jr, S.J. (1969). Effect of prior notification on the refusal rate in fixed address surveys, *Journal of Advertising Research* **9**, 42–44.
- [10] Burchell, B. & Marsh, C. (1992). Effect of questionnaire length on survey response, *Quality and Quantity* **26**, 233–244.
- [11] Carpenter, E.H. (1975). Personalizing mail surveys: a replication and reassessment, *Public Opinion Quarterly* **38**, 614–620.
- [12] Champion, D.J. & Sear, A.M. (1969). Questionnaire response rates: a methodological analysis, *Social Forces* **47**, 335–339.
- [13] Childers, T.L. & Ferrell, O.C. (1979). Response rates and perceived questionnaire length in mail surveys, *Journal of Marketing Research* **16**, 429–431.
- [14] Childers, T.L. & Skinner, S.J. (1985). Theoretical and empirical issues in the identification of survey respondents, *Journal of the Market Research Society* **27**, 39–53.
- [15] Cox III, E.P., Anderson Jr, W.T. & Fulcher, D.G. (1974). Reappraising mail survey response rates, *Journal of Marketing Research* **11**, 413–417.
- [16] Denton, J., Tsai, C. & Chevrete, P. (1988). Effects on survey responses of subject, incentives, and multiple mailings, *Journal of Experimental Education* **56**, 77–82.
- [17] Dillman, D.A. (1972). Increasing mail questionnaire response in large samples of the general public, *Public Opinion Quarterly* **36**, 254–257.
- [18] Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*, John Wiley & Sons, New York.
- [19] Dillman, D.A. (1999). *Mail and Internet Surveys: The Tailored Design Method*, John Wiley & Sons, New York.
- [20] Dillman, D., Carpenter, E., Christenson, J. & Brooks, R. (1974). Increasing mail questionnaire response: a four state comparison, *American Sociological Review* **39**, 744–756.
- [21] Dillman, D.A. & Frey, J.H. (1974). Contribution of personalization to mail questionnaire response as an element of a previously tested method, *Journal of Applied Psychology* **59**, 297–301.
- [22] Duncan, W.J. (1979). Mail questionnaires in survey research: a review of response inducement techniques, *Journal of Management* **5**, 39–55.
- [23] Eckland, B. (1965). Effects of prodding to increase mail back returns, *Journal of Applied Psychology* **49**, 165–169.
- [24] Etzel, M.J. & Walker, B.J. (1974). Effects of alternative follow-up procedures on mail survey response rates, *Journal of Applied Psychology* **59**, 219–221.
- [25] Fillion, F.L. (1975). Estimating bias due to nonresponse in mail surveys, *Public Opinion Quarterly* **39**, 482–492.
- [26] Fillion, F.L. (1976). Exploring and correcting for nonresponse bias using follow-ups on nonrespondents, *Pacific Sociological Review* **19**, 401–408.
- [27] Ford, N.M. (1967). The advance letter in mail surveys, *Journal of Marketing Research* **4**, 202–204.

- [28] Ford, R.N. & Zeisel, H. (1949). Bias in mail surveys cannot be controlled by one mailing, *Public Opinion Quarterly* **13**, 495–501.
- [29] Fox, R.J., Crask, M.R. & Kim, J. (1988). Mail survey response rate: a meta-analysis of selected techniques for inducing response, *Public Opinion Quarterly* **52**, 467–491.
- [30] Frazier, G. & Bird, K. (1958). Increasing the response of a mail questionnaire, *Journal of Marketing* **22**, 186–187.
- [31] Fuller, C. (1974). Effect of anonymity on return rate and response bias in a mail survey, *Journal of Applied Psychology* **59**, 292–296.
- [32] Furse, D.H. & Stewart, D.W. (1982). Monetary incentives versus promised contribution to charity: new evidence on mail survey response, *Journal of Marketing Research* **19**, 375–380.
- [33] Furse, D.H., Stewart, D.W. & Rados, D.L. (1981). Effects of foot-in-the-door, cash incentives, and follow-ups on survey response, *Journal of Marketing Research* **18**, 473–478.
- [34] Futrell, C. & Swan, J.E. (1977). Anonymity and response by salespeople to a mail questionnaire, *Journal of Marketing Research* **14**, 611–616.
- [35] Futrell, C. & Hise, R.T. (1982). The effects on anonymity and a same-day deadline on the response rate to mail surveys, *European Research* **10**, 171–175.
- [36] Gajraj, A.M., Faria, A.J. & Dickinson, J.R. (1990). Comparison of the effect of promised and provided lotteries, monetary and gift incentives on mail survey response rate, speed and cost, *Journal of the Market Research Society* **32**, 141–162.
- [37] Hancock, J.W. (1940). An experimental study of four methods of measuring unit costs of obtaining attitude toward the retail store, *Journal of Applied Psychology* **24**, 213–230.
- [38] Harris, J.R. & Guffey Jr, H.J. (1978). Questionnaire returns: stamps versus business reply envelopes revisited, *Journal of Marketing Research* **15**, 290–293.
- [39] Heaton Jr, E.E. (1965). Increasing mail questionnaire returns with a preliminary letter, *Journal of Advertising Research* **5**, 36–39.
- [40] Heberlein, T.A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature, *American Sociological Review* **43**, 447–462.
- [41] Henley Jr, J.R. (1976). Response rate to mail questionnaires with a return deadline, *Public Opinion Quarterly* **40**, 374–375.
- [42] Hopkins, K.D. & Gullickson, A.R. (1992). Response rates in survey research: a meta-analysis of the effects of monetary gratuities, *Journal of Experimental Education* **61**, 52–62.
- [43] Hornik, J. (1981). Time cue and time perception effect on response to mail surveys, *Journal of Marketing Research* **18**, 243–248.
- [44] House, J.S., Gerber, W. & McMichael, A.J. (1977). Increasing mail questionnaire response: a controlled replication and extension, *Public Opinion Quarterly* **41**, 95–99.
- [45] Houston, M.J. & Jefferson, R.W. (1975). The negative effects of personalization on response patterns in mail surveys, *Journal of Marketing Research* **12**, 114–117.
- [46] Houston, M.J. & Nevin, J.R. (1977). The effects of source and appeal on mail survey response patterns, *Journal of Marketing Research* **14**, 374–377.
- [47] James, J.M. & Bolstein, R. (1990). Effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys, *Public Opinion Quarterly* **54**, 346–361.
- [48] Jolson, M.A. (1977). How to double or triple mail response rates, *Journal of Marketing* **41**, 78–81.
- [49] Jones, W.H. & Lang, J.R. (1980). Sample composition bias and response bias in a mail survey: a comparison of inducement methods, *Journal of Marketing Research* **17**, 69–76.
- [50] Jones, W.H. & Linda, G. (1978). Multiple criteria effects in a mail survey experiment, *Journal of Marketing Research* **15**, 280–284.
- [51] Kanuk, L. & Berenson, C. (1975). Mail surveys and response rates: a literature review, *Journal of Marketing Research* **12**, 440–453.
- [52] Kawash, M.B. & Aleamoni, L.M. (1971). Effect of personal signature on the initial rate of return of a mailed questionnaire, *Journal of Applied Psychology* **55**, 589–592.
- [53] Kephart, W.M. & Bressler, M. (1958). Increasing the responses to mail questionnaires, *Public Opinion Quarterly* **22**, 123–132.
- [54] Kerin, R.A. & Peterson, R.A. (1977). Personalization, respondent anonymity, and response distortion in mail surveys, *Journal of Applied Psychology* **62**, 86–89.
- [55] Kimball, A.E. (1961). Increasing the rate of return in mail surveys, *Journal of Marketing* **25**, 63–65.
- [56] Linsky, A.S. (1975). Stimulating responses to mailed questionnaires: a review, *Public Opinion Quarterly* **39**, 82–101.
- [57] Lockhart, D.C. (1991). Mailed surveys to physicians: the effect of incentives and length on the return rate, *Journal of Pharmaceutical Marketing & Management* **6**, 107–121.
- [58] Lorenzi, P., Friedmann, R. & Paolillo, J. (1988). Consumer mail survey responses: more (unbiased) bang for the buck, *Journal of Consumer Marketing* **5**, 31–40.
- [59] Martin, J.D. & McConnell, J.P. (1970). Mail questionnaire response induction: the effect of four variables on the response of a random sample to a difficult questionnaire, *Social Science Quarterly* **51**, 409–414.
- [60] Mason, W.S., Dressel, R.J. & Bain, R.K. (1961). An experimental study of factors affecting response to a mail survey of beginning teachers, *Public Opinion Quarterly* **25**, 296–299.
- [61] McCrohan, K.F. & Lowe, L.S. (1981). A cost/benefit approach to postage used on mail questionnaires, *Journal of Marketing* **45**, 130–133.

-
- [62] McDaniel, S.W. & Jackson, R.W. (1981). An investigation of respondent anonymity's effect on mailed questionnaire response rate and quality, *Journal of Market Research Society* **23**, 150–160.
- [63] Myers, J.H. & Haug, A.F. (1969). How a preliminary letter affects mail survey return and costs, *Journal of Advertising Research* **9**, 37–39.
- [64] Nevin, J.R. & Ford, N.M. (1976). Effects of a deadline and a veiled threat on mail survey responses, *Journal of Applied Psychology* **61**, 116–118.
- [65] Parsons, R.J. & Medford, T.S. (1972). The effect of advanced notice in mail surveys of homogeneous groups, *Public Opinion Quarterly* **36**, 258–259.
- [66] Pearlin, L.I. (1961). The appeals of anonymity in questionnaire response, *Public Opinion Quarterly* **25**, 640–647.
- [67] Price, D.O. (1950). On the use of stamped return envelopes with mail questionnaires, *American Sociological Review* **15**, 672–673.
- [68] Roberts, R.E., McCrory, O.F. & Forthofer, R.N. (1978). Further evidence on using a deadline to stimulate responses to a mail survey, *Public Opinion Quarterly* **42**, 407–410.
- [69] Rosen, N. (1960). Anonymity and attitude measurement, *Public Opinion Quarterly* **24**, 675–680.
- [70] Rucker, M., Hughes, R., Thompson, R., Harrison, A. & Vanderlip, N. (1984). Personalization of mail surveys: too much of a good thing? *Educational and Psychological Measurement* **44**, 893–905.
- [71] Schegelmilch, B.B. & Diamantopoulos, S. (1991). Prenotification and mail survey response rates: a quantitative integration of the literature, *Journal of the Market Research Society* **33**, 243–255.
- [72] Schewe, C.D. & Cournoyer, N.D. (1976). Prepaid vs. promised incentives to questionnaire response: further evidence, *Public Opinion Quarterly* **40**, 105–107.
- [73] Scott, C. (1961). Research on mail surveys, *Journal of the Royal Statistical Society, Series A, Part 2* **124**, 143–205.
- [74] Simon, R. (1967). Responses to personal and form letters in mail surveys, *Journal of Advertising Research* **7**, 28–30.
- [75] Stafford, J.E. (1966). Influence of preliminary contact on mail returns, *Journal of Marketing Research* **3**, 410–411.
- [76] Vocino, T. (1977). Three variables in stimulating responses to mailed questionnaires, *Journal of Marketing* **41**, 76–77.
- [77] Walker, B.J. & Burdick, R.K. (1977). Advance correspondence and error in mail surveys, *Journal of Marketing Research* **14**, 379–382.
- [78] Wildman, R.C. (1977). Effects of anonymity and social settings on survey responses, *Public Opinion Quarterly* **41**, 74–79.
- [79] Wotruba, T.R. (1966). Monetary inducements and mail questionnaire response, *Journal of Marketing Research* **3**, 398–400.
- [80] Wynn, G.W. & McDaniel, S.W. (1985). The effect of alternative foot-in-the-door manipulations on mailed questionnaire response rate and quality, *Journal of the Market Research Society* **27**, 15–26.
- [81] Yammarino, F.J., Skinner, S.J. & Childers, T.L. (1991). Understanding mail survey response behavior, *Public Opinion Quarterly* **55**, 613–639.
- [82] Yu, J. & Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires, *Journal of Marketing Research* **20**, 36–44.

(See also **Survey Sampling Procedures**)

THOMAS W. MANGIONE

Mallows' C_p Statistic

CHARLES E. LANCE

Volume 3, pp. 1119–1120

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mallows' C_p Statistic

A common problem in applications of **multiple linear regression analysis** in the behavioral sciences is that of predictor subset selection [10]. The goal of subset selection, also referred to as **variable selection** (e.g., [4]) or **model selection** (e.g., [13]) is to choose a smaller subset of predictors from a relatively larger number that is available so that the resulting regression model is parsimonious, yet has good predictive ability [3, 4, 12]. The problem of subset selection arises, for example, when a researcher seeks a predictive model that cross-validates well (see **Cross-validation** and [1, 11]), or when there is redundancy amongst the predictors leading to multicollinearity [3].

There are several approaches to predictor subset selection, including forward selection, backward elimination, and stepwise regression [2, 10]. A fourth, 'all possible subsets' procedure fits all $2^k - 1$ distinct models to determine a best fitting model (BFM) on the basis of some statistical criterion. A number of such criteria for choosing a BFM can be considered [5], including R^2 , one of several forms of R^2 adjusted for the number of predictors [11], the mean squared error, or one of a number of criteria based on information theory (e.g., **Akaike's** criterion, [6]). Mallows' [8, 9] C_p statistic is one such criterion that is related to this latter class of indices.

For any model containing a subset of p predictors from the total number of k predictors, Mallows' C_p can be written as:

$$C_p = \frac{RSS_p}{MSE_k} + 2p - n \quad (1)$$

where RSS_p is the residual sum of squares for the p -variable model, MSE_k is the mean squared error for the full (k -variable) model, and n is sample size. As such, C_p indexes the mean squared error of prediction for 'subset' models relative to the full model with a penalty for inclusion of unimportant predictors [7]. Because MSE_k is usually estimated as $RSS_k/(n - k)$, C_p can also be written as:

$$C_p = (n - k) \frac{RSS_p}{RSS_k} + 2p - n \quad (2)$$

Note that if the R^2 for a p -variable model is substantially less than the R^2 for the full k -variable model (i.e., 'important' variables are excluded from the subset), RSS_p/RSS_k will be large compared to the situation in which the p -variable subset model includes all or most of the 'important' predictors. In this case, the RSS_p/RSS_k ratio approaches 1.00, and if the model is relatively parsimonious, a large 'penalty' (in the form of $2p$) is not invoked and C_p is small. Note, however, that for the full model, $p = k$ so that $RSS_p/RSS_k = 1$ and C_p necessarily $= p$. In practice, for models with $p < k$ predictors, variable subsets with lower C_p values (around p or less) indicate preferred subset models. It is commonly recommended to plot models' C_p as a function of p and choose the predictor subset with minimum C_p as the preferred subset model (see [3], [8], and [9] for examples).

References

- [1] Camstra, A. & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: an overview, *Sociological Methods & Research* **21**, 89–115.
- [2] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.
- [3] Fox, J. (1991). *Regression Diagnostics*, Sage Publications, Newbury Park.
- [4] George, E.I. (2000). The variable selection problem, *Journal of the American Statistical Association* **95**, 1304–1308.
- [5] Hocking, R.R. (1972). Criteria for selection of a subset regression: which one should be used? *Technometrics* **14**, 967–970.
- [6] Honda, Y. (1994). Information criteria in identifying regression models, *Communications in Statistics—Theory and Methods* **23**, 2815–2829.
- [7] Kobayashi, M. & Sakata, S. (1990). Mallows' C_p criterion and unbiasedness of model selection, *Journal of Econometrics* **45**, 385–395.
- [8] Mallows, C.L. (1973). Some comments on C_p , *Technometrics* **15**, 661–675.
- [9] Mallows, C.L. (1995). More comments on C_p , *Technometrics* **37**, 362–372.
- [10] Miller, A.J. (1990). *Subset Selection in Regression*, Chapman and Hall, London.
- [11] Raju, N.S., Bilgic, R., Edwards, J.E. & Fleer, P.F. (1997). Methodology review: estimation of population validity and cross-validity, and the use of equal weights

2 Mallows' C_p Statistic

- in prediction, *Applied Psychological Measurement* **21**, 291–305.
- [12] Thompson, M.L. (1978). Selection of variables in multiple regression: part I. A review and evaluation, *International Statistical Review* **46**, 1–19.
- [13] Zuccaro, C. (1992). Mallows' C_p statistic and model selection in multiple linear regression, *Journal of the Market Research Society* **34**, 163–172.

CHARLES E. LANCE

Mantel–Haenszel Methods

ÁNGEL M. FIDALGO

Volume 3, pp. 1120–1126

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mantel–Haenszel Methods

Numerous studies within the behavioral sciences have attempted to determine the degree of association between an explanatory variable, called *factor* (e.g., treatments, potentially harmful exposure), and another variable that is assumed to be determined by or dependent upon the factor variable, called *response variable* (e.g., degree of improvement, state of health). Moreover, quite frequently, researchers wish to control the modulating effect of other variables, known as *stratification variables* or *covariables* (e.g., age, level of illness, gender), on the relationship between factor and response (see **Analysis of Covariance**). **Stratification** variables may be the result of the research design, as in a multicenter clinical trial, in which the strata correspond to the different hospitals where the treatments have been applied or of *a posteriori* considerations made after the study data have been obtained. In any case, when factor and response variables are reported on categorical measurement scales (see **Categorizing Data**), either nominal or ordinal, the resulting data can be summarized in **contingency tables**, and the methods based on the work of **Cochran** [4] and Mantel and Haenszel [16] are commonly used for their analysis. In a general way, it can be said that these methods provide measures and significance tests of two-way association that control for the effects of covariables. The null hypothesis (H_0) they test is that of ‘partial no-association’, which establishes that in each one of the strata of the covariable, the response variable is distributed randomly with respect to the levels of the factor. In the simplest case, both factor and response are dichotomous variables, and the data can be summarized in a set of Q contingency tables 2×2 (denoted by $Q : 2 \times 2$), where each table corresponds to a stratum or level of the covariable, or to each combination of levels in the case of there being several covariables (see **Two by Two Contingency Tables**). To establish notation, let n_{hij} denote the frequency count of observations occurring at the i th factor level (row), ($i = 1, \dots, R = 2$), the j th level of the response variable (column), ($j = 1, \dots, C = 2$), and the h th level of the covariable or stratum, ($h = 1, \dots, Q$). In Table 1, this notation is applied to a typical study on risk factors and state of health.

Table 1 The 2×2 contingency table for the h th stratum

Exposure status	Health status		Total
	Condition present	Condition absent	
Not exposed	n_{h11}	n_{h12}	$N_{h1\cdot}$
Exposed	n_{h21}	n_{h22}	$N_{h2\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$	$N_{h\cdot\cdot}$

In sets of tables 2×2 , Mantel–Haenszel (MH) statistics are asymptotic tests that follow a chi-squared distribution (see **Catalogue of Probability Density Functions**) with one degree of freedom (df) and that, in general, can be expressed thus:

$$\chi_{MH}^2 = \frac{\left(\sum_{h=1}^Q a - \sum_{h=1}^Q E(A) \right)^2}{\sum_{h=1}^Q \text{var}(A)}. \quad (1)$$

Here, a is the observed frequencies in one of the cells of the table, $E(A)$ is the frequencies expected under H_0 , and $\text{var}(A)$ is the variance of the frequencies expected under H_0 . All of this summed across the strata. The strategy used for testing H_0 is as simple as comparing the frequencies observed in the contingency table (n_{hij}) with the frequencies that could be expected in the case of there being no relationship between factor and response. Naturally, in order to carry out the test, we need to establish a probabilistic model that determines the probability of obtaining a under H_0 . According to the assumed probability model, $E(A)$ and $\text{var}(A)$ will adopt different forms. Mantel and Haenszel [16] conditioning on the row and column totals ($N_{h1\cdot}$, $N_{h2\cdot}$, $N_{h\cdot 1}$, and $N_{h\cdot 2}$), that is, taking those values as fixed, establish that, under H_0 , the observed frequencies in each table (n_{hij}) follow the multiple hypergeometric probability model (see **Catalogue of Probability Density Functions**):

$$\Pr(n_{hij} | H_0) = \frac{\prod_{i=1}^R N_{hi\cdot}! \prod_{j=1}^C N_{h\cdot j}!}{N_{h\cdot\cdot}! \prod_{i=1}^R \prod_{j=1}^C n_{hij}!}. \quad (2)$$

2 Mantel–Haenszel Methods

Furthermore, on assuming that the marginal totals $(N_{hi\cdot})$ and $(N_{h\cdot j})$ are fixed, this distribution can be expressed in terms of count n_{h11} alone, since it determines the rest of the cell counts (n_{h12} , n_{h21} , and n_{h22}). Under H_0 , the hypergeometric mean and variance of n_{h11} are $E(n_{h11}) = N_{h1\cdot}N_{h\cdot 1}/N_{h\cdot\cdot}$ and $\text{var}(n_{h11}) = N_{h1\cdot}N_{h2\cdot}N_{h\cdot 1}N_{h\cdot 2}/N_{h\cdot\cdot}^2(N_{h\cdot\cdot} - 1)$ respectively. We thus obtain the MH chi-squared test:

$$\chi_{\text{MH}}^2 = \frac{\left(\left| \sum_{h=1}^Q n_{h11} - \sum_{h=1}^Q E(n_{h11}) \right| - 0.5 \right)^2}{\sum_{h=1}^Q \text{var}(n_{h11})}. \quad (3)$$

Cochran [4], taking only the total sample in each stratum $(N_{h\cdot\cdot})$ as fixed and assuming a binomial distribution, proposed a statistic that was quite similar to χ_{MH}^2 . The Cochran statistic can be rewritten in terms of (3), eliminating from it the continuity correction (0.5) and substituting $\text{var}(n_{h11})$ by $\text{var}^*(n_{h11}) = N_{h1\cdot}N_{h2\cdot}N_{h\cdot 1}N_{h\cdot 2}/N_{h\cdot\cdot}^3$. The essential equivalence between the two statistics when we have moderate sample sizes in each stratum means that the techniques expressed here are frequently referred to as the *Cochran–Mantel–Haenszel methods*. Nevertheless, statistic (3) is preferable, since it only demands that the combined row sample sizes $(N_{i\cdot} = \sum_{h=1}^Q \sum_{j=1}^C n_{hij})$ be large (e.g., $N_{i\cdot} > 30$), while that of Cochran demands moderate to large sample sizes (e.g., $N_{h\cdot\cdot} > 20$) in each table. A more precise criterion in terms of sample demands [15] recommends applying χ_{MH}^2 only if the result of the expressions

$$\sum_{h=1}^Q E(n_{h11}) - \left[\sum_{h=1}^Q \max(0, N_{h1\cdot} - N_{h\cdot 2}) \right] \text{ and } \left[\sum_{h=1}^Q \min(N_{h\cdot 1}, N_{h1\cdot}) \right] - \sum_{h=1}^Q E(n_{h11}) \quad (4)$$

is over five. When the sample requirements are not fulfilled, either because of the small sample size or because of the highly skewed observed table margins, we will have to use exact tests, such as those of Fisher or Birch [1].

In the case of rejecting the H_0 at the α significance level (if $\chi_{\text{MH}}^2 \geq_{\text{df}(1)} \chi_{\alpha}^2$), the following step

is to determine the degree of association between factor and response. Among the numerous measures of association available for contingency tables (rate ratio, **relative risk**, **prevalence ratio**, etc.), Mantel and Haenszel [16] employed **odds ratios**. The characteristics of this measure of association is examined as follows. It should be noted that if the variables were independent, the *odds* of the probability (π) of responding in column 1 instead of column 2 (π_{h11}/π_{h12}) would be equal at all the levels of the factor. Therefore, the ratio of the odds, referred to as the *odds ratio* $\{\alpha = (\pi_{h11}/\pi_{h12})/(\pi_{h21}/\pi_{h22})\}$ or *cross-product ratio* ($\alpha = \pi_{h11}\pi_{h22}/\pi_{h12}\pi_{h21}$), will be 1. Assuming homogeneity of the odds ratios of each stratum ($\alpha_1 = \alpha_2 = \dots = \alpha_Q$), the MH measure of association calculated across all 2×2 contingency tables is the *common odds ratio estimator* ($\hat{\alpha}_{\text{MH}}$), given by

$$\hat{\alpha}_{\text{MH}} = \frac{\sum_{h=1}^Q R_h}{\sum_{h=1}^Q S_h} = \frac{\sum_{h=1}^Q n_{h11}n_{h22}/N_{h\cdot\cdot}}{\sum_{h=1}^Q n_{h12}n_{h21}/N_{h\cdot\cdot}}. \quad (5)$$

The range of $\hat{\alpha}_{\text{MH}}$ varies between 0 and ∞ . A value of 1 represents the hypothesis of no-association. If $1 < \hat{\alpha}_{\text{MH}} < \infty$, the first response is more likely in row 1 than in row 2; if $0 < \hat{\alpha}_{\text{MH}} < 1$, the first response is less likely in row 1 than in row 2. Because of the skewness of the distribution of $\hat{\alpha}_{\text{MH}}$, it is more convenient to use $\ln(\hat{\alpha}_{\text{MH}})$, the natural logarithm of $\hat{\alpha}_{\text{MH}}$. In this case, the independence corresponds to $\ln(\hat{\alpha}_{\text{MH}}) = 0$, the $\ln$ of the common odds ratio being symmetrical about this value. This means that, for $\ln(\hat{\alpha}_{\text{MH}})$, two values that are equal except for their sign, such as $\ln(2) = 0.69$ and $\ln(0.5) = -0.69$, represent the same degree of association. For constructing **confidence intervals** around $\hat{\alpha}_{\text{MH}}$, we need an estimator of its variance, and that which presents the best properties [12] is that of Robins, Breslow, and Greenland [18]:

$$\text{var}(\hat{\alpha}_{\text{MH}}) = \frac{(\hat{\alpha}_{\text{MH}})^2}{2} \frac{\sum_{h=1}^Q (P_h R_h)}{\left(\sum_{h=1}^Q R_h \right)^2}$$

$$+ \frac{\sum_{h=1}^Q (P_h S_h + Q_h R_h)}{\sum_{h=1}^Q R_h \sum_{h=1}^Q S_h} + \frac{\sum_{h=1}^Q Q_h S_h}{\left(\sum_{h=1}^Q S_h\right)^2}, \quad (6)$$

where $P_h = (n_{h11} + n_{h22})/N_{h..}$ and $Q_h = (n_{h12} + n_{h21})/N_{h..}$.

If we construct the intervals on the basis of $\ln(\hat{\alpha}_{MH})$, we must make the following adjustment: $\text{var}[\ln(\hat{\alpha}_{MH})] = \text{var}(\hat{\alpha}_{MH})/(\hat{\alpha}_{MH})^2$. Thus, the $100(1 - \alpha)\%$ confidence interval for $\ln(\hat{\alpha}_{MH})$ will be equal to $\ln(\hat{\alpha}_{MH}) \pm z_{\alpha/2} \sqrt{\text{var}[\ln(\hat{\alpha}_{MH})]}$. For $\hat{\alpha}_{MH}$, the $100(1 - \alpha)\%$ confidence limits will be equal to $\exp(\ln(\hat{\alpha}_{MH}) \pm z_{\alpha/2} \sqrt{\text{var}[\ln(\hat{\alpha}_{MH})]})$.

In relation to these aspects, it should be pointed out that nonfulfillment of the assumption of homogeneity of the odds ratios ($\alpha_1 = \alpha_2 = \dots = \alpha_Q$) does not invalidate $\hat{\alpha}_{MH}$ as a measure of association; even so, given that it is a weighted average of the stratum-specific odds ratios, it makes its interpretation difficult. In the case that the individual odds ratios differ substantially in direction, it is preferable to use the information provided by these odds ratios than to use $\hat{\alpha}_{MH}$. Likewise, nonfulfillment of the assumption of homogeneity does not invalidate χ^2_{MH} as a test of association, though it does reduce its statistical power [2]. Breslow and Day [3, 20] provide a test for checking the mentioned assumption (see **Breslow–Day Statistic**).

The statistics shown above are applied to the particular case of sets of contingency tables 2×2 . Fortunately, from the outset, various extensions have been proposed for these statistics [14, 16], all of them being particular cases of the analysis of sets of contingency tables with dimensions $Q : R \times C$. The

data structure for this general contingency table is shown in Table 2.

In the general case, the H_0 of no-association will be tested against different alternative hypotheses (H_1) that will be a function of the scale on which factor and response are measured. Thus, we shall have a variety of statistics that will serve for detecting the general association (both variables are nominal), mean score differences (factor is nominal and response ordinal), and linear correlation (both variables are ordinal).

The standard generalized Mantel–Haenszel test is defined, in terms of matrices, by Landis, Heyman, and Koch [13], as

$$Q_{GMH} = \left\{ \sum_{h=1}^Q (\mathbf{n}_h - \mathbf{m}_h)' \mathbf{A}_h' \right\} \left\{ \sum_{h=1}^Q \mathbf{A}_h \mathbf{V}_h \mathbf{A}_h' \right\}^{-1} \times \left\{ \sum_{h=1}^Q \mathbf{A}_h (\mathbf{n}_h - \mathbf{m}_h) \right\}, \quad (7)$$

where $\mathbf{n}_h$, $\mathbf{m}_h$, $\mathbf{V}_h$, and $\mathbf{A}_h$ are, respectively, the vector of observed frequencies, the vector of expected frequencies, the covariances matrix, and a matrix of linear functions defined in accordance with the H_1 of interest. From Table 2, these vectors are defined as $\mathbf{n}_h = (n_{h11}, n_{h21}, \dots, n_{hRC})'$; $\mathbf{m}_h = N_{h..}(\mathbf{p}_{h.*} \otimes \mathbf{p}_{h*})$, $\mathbf{p}_{h.*}$ and $\mathbf{p}_{h*}$ being vectors with the marginal row proportions ($p_{hi.} = N_{hi.}/N_{h..}$) and the marginal column proportions ($p_{h.j} = N_{h.j}/N_{h..}$), and $\otimes$ denoting the Kronecker product multiplication; $\mathbf{V}_h = N_{h..}^2/(N_{h..} - 1) \{(\mathbf{D}_{p_{h.*}} - \mathbf{p}_{h.*} \mathbf{p}_{h.*}') \otimes (\mathbf{D}_{p_{h*}} - \mathbf{p}_{h*} \mathbf{p}_{h*}')\}$, where $\mathbf{D}_{p_h}$ is a diagonal matrix with elements of the vector $\mathbf{p}_h$ on its main diagonal.

As it has been pointed out, depending on the measurement scale of factor and response, (7) will be resolved, via definition of the matrix $\mathbf{A}_h$ ($\mathbf{A}_h =$

Table 2 Data structure in the h th stratum

Factor levels	Response variable categories						Total
	1	2	·	j	·	C	
1	n_{h11}	n_{h12}	·	n_{h1j}	·	n_{h1C}	$N_{h1.}$
2	n_{h21}	n_{h22}	·	n_{h2j}	·	n_{h2C}	$N_{h2.}$
·	·	·	·	·	·	·	·
i	n_{hi1}	n_{hi2}	·	n_{hij}	·	n_{hiC}	$N_{hi.}$
·	·	·	·	·	·	·	·
R	n_{hR1}	n_{hR2}	·	n_{hRj}	·	n_{hRC}	$N_{hR.}$
Total	$N_{h.1}$	$N_{h.2}$	·	$N_{h.j}$	·	$N_{h.C}$	$N_{h..}$

$C_h \otimes R_h$), in a different statistic for detecting each H_1 . Briefly, these are as follows:

$Q_{GMH(1)}$. When the variable row and the variable column are nominal, the H_1 specifies that the distribution of the response variable differs in nonspecific patterns across levels of the row factor. Here, $R_h = [I_{R-1}, -J_{R-1}]$ and $C_h = [I_{C-1}, -J_{C-1}]$, where I_{R-1} is an identity matrix, and J_{R-1} is an $(R-1) \times 1$ vector of ones. Under H_0 , $Q_{GMH(1)}$ follows approximately a chi-squared distribution with $df = (R-1)(C-1)$.

$Q_{GMH(2)}$. When only the variable column is ordinal, the H_1 establishes that the mean responses differ across the factor levels, R_h being the same as that used in the previous case and $C_h = (c_{h1}, \dots, c_{hC})$, where c_{hj} is an appropriate score reflecting the ordinal nature of the j th category of response for the h th stratum. Selection of the values of C_h admits different possibilities that are well described in [13]. Under H_0 , $Q_{GMH(2)}$ has approximately a chi-squared distribution with $df = (R-1)$.

$Q_{GMH(3)}$. If both variables are ordinal, the H_1 establishes the existence of a linear progression (linear trend) in the mean responses across the levels of the factor (*see Trend Tests for Counts and Proportions*). In this case, C_h can be defined as the same as that for the mean responses difference and $R_h = (r_{h1}, \dots, r_{hR})$, where r_{hi} is an appropriate score reflecting the ordinal nature of the i th factor level for the h th stratum. Under H_0 , $Q_{GMH(3)}$ has approximately a chi-squared distribution with $df = 1$.

It should be noted how the successive H_1 s specify more and more restrictive patterns of association, so that each statistic increases the statistical power with respect to the previous ones for detecting its particular pattern of association. For example, $Q_{GMH(1)}$ can detect linear patterns of association, but it will do so with less power than $Q_{GMH(3)}$. Furthermore, the increase in power of $Q_{GMH(3)}$ compared to $Q_{GMH(1)}$ is achieved at the cost of an inability to detect more complex patterns of association. Obviously, when $C = R = 2$, $Q_{GMH(1)} = Q_{GMH(2)} = Q_{GMH(3)} = \chi_{MH}^2$, except for the lack of the continuity correction.

While MH methods have satisfactorily resolved the testing of the H_0 in the general case, to date

there is no estimator of the degree of association that is completely generalizable for $Q : R \times C$ tables. The interested reader can find generalizations, always complex, of $\hat{\alpha}_{MH}$ to $Q : 2 \times C$ tables ($C > 2$) in [10], [17], and [21].

The range of application of the methodology presented in this article is enormous, and it is widely used in epidemiology, **meta-analysis**, analysis of survival data (*see Survival Analysis*) (where it is known by the name of *logrank-test*), and psychometric research on differential item functioning [5–7, 9]. This is undoubtedly due to its simplicity and flexibility and due to its minimal demands for guaranteeing the validity of its results: on the one hand, it requires only that the sample size summed across the strata be sufficiently large for asymptotic results to hold (such that the MH test statistic can perfectly well be applied for matched **case-control studies** with only two subjects in each of the Q tables, as long as the number of tables is large); on the other hand, it permits the use of samples of convenience on not assuming a known sampling link to a larger reference population. This is possible, thanks to the fact that the H_0 of interest – that the distribution of the responses is random with respect to the levels of the factor – induces a probabilistic structure (the multiple hypergeometric distribution) that allows for judgment of its compatibility with the observed data without the need for external assumptions. Thus, it can be applied to **experimental designs**, group designs, and repeated-measures designs [12, 19, 22, 23], (*see Repeated Measures Analysis of Variance*) and also to designs based on observation or of a historical nature, such as **retrospective studies**, nonrandomized studies, or case–control studies [11], regardless of how the sampling was carried out. This undoubted advantage is offset by a disadvantage: given that the probability distributions employed are determined by the observed data (see (1)), the conclusions obtained will apply only to the sample under study. Consequently, generalization of the results to the target population should be based on arguments about the representativeness of the sample [11].

Example

We shall illustrate the use of the Mantel–Haenszel methods on the basis of the results of a survey on bullying and harassment at work (mobbing) carried out in Spain [8]. Table 3 shows the number of people

Table 3 Characteristics of mobbing victims

Gender	Job category	No bullying	bullied	Total	% bullied
Men	Manual	148	28	176	15.9
	Clerical	65	13	78	16.7
	Specialized technician	121	18	139	12.9
	Middle manager	95	7	102	6.9
	Manager/executive	29	2	31	6.5
Women	Manual	98	22	120	18.3
	Clerical	144	32	176	18.2
	Specialized technician	43	10	53	18.9
	Middle manager	38	7	45	15.6
	Manager/executive	8	1	9	11.1

in the survey who over a period of 6 months or more have been bullied at least once a week (bullied), and the number of those not fulfilling this criterion (no bullying).

In order to employ all the statistics considered, we shall analyze the data in different ways. Let us suppose that the main objective of the research is to determine whether there is a relationship between job category and mobbing. Moreover, as it is suspected that gender may be related to mobbing, we decided to control the effect of this variable. The result of applying the statistic χ^2_{MH} , without the continuity correction, to the following contingency table

	No bullying	Bullied	Total
Men	458	68	526
Women	331	72	403
Total	789	140	929

indicates the pertinence of this adjustment. Indeed, we find that χ^2_{MH} takes the value 4.34, which has

a p value of 0.037 with 1 df . Thus, assuming a significance level of .05, we reject the H_0 of ‘no partial association’ in favor of the alternative that suggests that bullying behavior is not distributed homogeneously between men and women. A value of 1.47 in $\hat{\alpha}_{MH}$ indicates that women have a higher probability of suffering mobbing than men do. The 95% confidence interval for the common odds ratio estimator, using the variance in (6) with a result of 0.07, is (1.02, 2.10). In the case of transforming $\hat{\alpha}_{MH}$ to the logarithmic scale [$\ln(\hat{\alpha}_{MH}) = 0.38$], the variance would equal 0.03, and the confidence interval, (0.022, 0.74). Note how $1.02 = \exp(0.02)$ and $2.10 = \exp(0.74)$. It should also be noted how our data satisfy the sample demands of the Mantel–Fleiss criterion (4): $\sum_{h=1}^1 E(n_{h11}) = 446.73$, $\sum_{h=1}^1 \max(0, 526 - 140) = 386$ and $\sum_{h=1}^1 \min(789, 526) = 526$, so that $(446.73 - 386) > 5$ and $(526 - 446.73) > 5$.

In line with the research objective, Table 4 shows the results of the Mantel–Haenszel statistics for association between job category and mobbing

Table 4 Results of the Mantel–Haenszel statistics with the respective linear operator matrices

Statistic	Value	df	p	$\mathbf{A}_h = \mathbf{C}_h \otimes \mathbf{R}_h$
$Q_{GMH(1)}$	5.74	4	.22	$[1 \ -1] \otimes \begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$
$Q_{GMH(2)}$	5.74	4	.22	$[1 \ 2] \otimes \begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$
$Q_{GMH(3)}$	4.70	1	.03	$[1 \ 2] \otimes [1 \ 2 \ 3 \ 4 \ 5]$

adjusted for gender; also shown is the linear operator matrix defined in accordance with the H_1 s of general association ($Q_{GMH(1)}$), mean row scores difference ($Q_{GMH(2)}$), and trend in mean scores ($Q_{GMH(3)}$).

Considering the data in Table 4, a series of points can be made. First point is that when the response variable is dichotomous, as is the case here, $Q_{GMH(1)}$ and $Q_{GMH(2)}$ offer the same results. Second is that if we suppose that both the response variable and the factor are ordinal variables, then there exists a statistically significant linear relationship ($p < .05$) between job category and level of harassment or bullying, once the effect of gender is controlled; more specifically, that the higher the level of occupation, the less bullying there is. Third point is that, as we can see, $Q_{GMH(3)}$ is the most powerful statistic, via reduction of df, for detecting the linear relationship between factor and response variable. Finally, we can conclude that almost 82% (4.70/5.74) of the nonspecific difference in mean scores can be explained by the linear tendency.

Note: The software for calculating the generalized Mantel-Haenszel statistics is available upon request from the author.

References

- [1] Agresti, A. (1992). A survey of exact inference for contingency tables, *Statistical Science* **7**, 131–177.
- [2] Birch, M.W. (1964). The detection of partial association, I: the 2×2 case, *Journal of the Royal Statistical Society. Series B* **26**, 313–324.
- [3] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research I: The analysis of Case-Control Studies*, International Agency for research on Cancer, Lyon.
- [4] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 test, *Biometrics* **10**, 417–451.
- [5] Fidalgo, A.M. (1994). MHDIF: a computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure, *Applied Psychological Measurement* **18**, 300.
- [6] Fidalgo, A.M., Ferreres, D. & Muñiz, J. (2004). Utility of the Mantel-Haenszel procedure for detecting differential item functioning with small samples, *Educational and Psychological Measurement* **64**(6), 925–936.
- [7] Fidalgo, A.M., Mellenbergh, G.J. & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures, *Methods of Psychological Research Online* **5**(3), 43–53.
- [8] Fidalgo, A.M. & Piñuel, I. (2004). La escala Cisneros como herramienta de valoración del mobbing [Cisneros scale to assess psychological harassment or mobbing at work], *Psicothema* **16**, 615–624.
- [9] Holland, W.P. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure, in *Test Validity*, H. Wainer & H.I. Braun, eds, LEA, Hillsdale, pp. 129–145.
- [10] Greenland, S. (1989). Generalized Mantel-Haenszel estimators for $K \times J$ tables, *Biometrics* **45**, 183–191.
- [11] Koch, G.G., Gillings, D.B. & Stokes, M.E. (1980). Biostatistical implications of design, sampling and measurement to health science data analysis, *Annual Review of Public Health* **1**, 163–225.
- [12] Kuritz, S.J., Landis, J.R. & Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: applications and recent developments, *Annual Review of Public Health* **9**, 123–160.
- [13] Landis, J.R., Heyman, E.R. & Koch, G.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.
- [14] Mantel, N. (1963). Chi-square tests with one degree of freedom; extension of the Mantel-Haenszel procedure, *Journal of the American Statistical Association* **58**, 690–700.
- [15] Mantel, N. & Fleiss, J.L. (1980). Minimum expected cell size requirements for the Mantel-Haenszel one-degree of freedom chi-square test and a related rapid procedure, *American Journal of Epidemiology* **112**, 129–143.
- [16] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [17] Mickey, R.M. & Elashoff, R.M. (1985). A generalization of the Mantel-Haenszel estimator of partial association for $2 \times J \times K$ tables, *Biometrics* **41**, 623–635.
- [18] Robins, J., Breslow, N. & Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics* **42**, 311–324.
- [19] Stokes, M.E., Davis, C.S. & Koch, G. (2000). *Categorical Data Analysis Using the SAS® System*, 2nd Edition, SAS Institute Inc, Cary.
- [20] Tarone, R.E., Gart, J.J. & Hauck, W.W. (1983). On the asymptotic inefficiency of certain noniterative estimators of a common relative risk or odds ratio, *Biometrika* **70**, 519–522.
- [21] Yanagawa, T. & Fujii, Y. (1995). Projection-method Mantel-Haenszel estimator for $K \times J$ tables, *Journal of the American Statistical Association* **90**, 649–656.
- [22] Zhang, J. & Boos, D.D. (1996). Mantel-Haenszel test statistics for correlated binary data, *Biometrics* **53**, 1185–1198.
- [23] Zhang, J. & Boos, D.D. (1997). Generalized Cochran-Mantel-Haenszel test statistics for correlated categorical data, *Communications in Statistics: Theory and Methods* **26**(8), 1813–1837.

ÁNGEL M. FIDALGO

Marginal Independence

VANCE W. BERGER AND JIALU ZHANG

Volume 3, pp. 1126–1128

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Marginal Independence

The definition of statistical independence is rather straightforward. Specifically, if $f(x,y|\theta)$ denotes the joint distribution of random variables X and Y , then the marginal distributions of X and Y can be written as $f(x|\theta) = \int f(x,y|\theta) dy$ and $f(y|\theta) = \int f(x,y|\theta) dx$, respectively, where θ is some parameter in the distribution functions, and in the discrete case one would use summation instead of integration (*see Contingency Tables*). Then the variables X and Y are independent if their joint distribution is the product of the two marginal distributions, $f(x,y|\theta) = f(x|\theta)f(y|\theta)$. The intuition here is that knowledge of one of the variables offers no knowledge of the other. Yet the simplicity of this definition belies the complexity that may arise when it is applied in practice.

For example, it may seem intuitively obvious that shoe size and intelligence should not be correlated or associated. That is, they should be independent. This may be true when considering a given cohort defined by age group, but when the age is allowed to vary, it is not hard to see that older children would be expected to have both larger feet and more education than younger children. This confounding variable, age, could then lead to the finding that children with larger shoe sizes appear to be more intelligent than children with smaller shoe sizes.

Conversely, one would expect height and weight to be positively associated, and not independent, because all things considered, taller individuals tend to also be heavier. However, this assumes that we are talking about the relation between height and weight among a set of distinct individuals. It is also possible to consider the relation between height and weight for a given individual over time. While many individuals will gain weight as they age and get taller over a period of years, one may also consider the relation between height and weight for a given individual over the period of one day. That is, the height and weight of a given individual can be measured every hour during a given day. Among these measurements, height and weight may be independent.

Finally, consider income and age. For any given individual who does not spend much time out of work, income would tend to rise over time, as the age also increases. As such, these two variables should have a positive association over time for a given

individual. Yet, to the extent that educational opportunities improve over time, a younger generation may have access to better paying positions than their older counterparts. This trend could compensate for, or even reverse, the association within individuals.

We see, then, that the association between two variables depends very much on the context, and so the context must always be borne in mind when discussing association, causality, or independence. In fact, it is possible to create the appearance of a positive relation, either unwittingly or otherwise, when in fact no association exists. Berkson [4] demonstrated that when two events are independent in a population at large, they become negatively associated in a subpopulation characterized by the requirement of one of these events. That is, imagine a trauma center that handles only victims of gun-shot wounds and victims of car accidents, and suppose that in the population at large the two are independent. Then the probability of having been shot would not depend on whether one has been in a car accident. However, if one wanted to over-sample the cases, and study the relationship between these two events in our hypothetical trauma center, then one might find a negative association due to the missing cell corresponding to neither event. See Table 1.

In the trauma center, there would be an empty upper-left cell, corresponding to neither the X -event nor the Y -event. That is, n_{11} would be replaced with zero, and this would create the appearance of a negative association when in fact none exists. It is also possible to intentionally create the appearance of such a positive association between treatment and outcome with strategic patient selection [2]. Unfortunately, there is no universal context, and marginal independence without consideration of covariates does not tell us much about association within levels defined by the covariates. Suppose, for example, that Table 1 turned out as shown in Table 2.

In Table 2 X and Y are marginally independent because each cell entry (50) can be expressed

Table 1 Creating dependence between variables X and Y

$x \backslash y$	0	1	Sum
0	$n_{11} = 0$	n_{12}	$n_{1+} = n_{12}$
1	n_{21}	n_{22}	n_{2+}
Sum	$n_{+1} = n_{21}$	n_{+2}	n_{++}

2 Marginal Independence

Table 2 Independence between variables X and Y

$y \backslash x$	0	1	Sum
0	50	50	100
1	50	50	100
Sum	100	100	200

Table 3 Independence between variables X and Y by gender

Males

$y \backslash x$	0	1	Sum
0	25	25	50
1	25	25	50
Sum	50	50	100

Females

$y \backslash x$	0	1	sum
0	25	25	50
1	25	25	50
Sum	50	50	100

as the product of its row and column proportions, and the grand total, 200. That is, $50 = (100/200)(100/200)(200)$. In general, for discrete variables, X and Y are marginally independent if $P(x, y) = P(x)P(y)$. But now consider a covariate. To illustrate, suppose that X is the treatment received, Y is the outcome, and Z is gender. From Table 2 we see that treatment and outcome are independent, but what does this really tell us? When gender is considered, the gender-specific tables may turn out to confirm this independence, as in Table 3.

In Table 3, each treatment is equally effective for each gender, with a 50% response rate. It is also possible that Table 2 conceals a true treatment*gender interaction, so that each treatment is better for one gender, and overall they compensate so as to balance out. Consider, for example, Table 4.

In Table 4, Treatment 0 is better for males, with a 60% response rate (compared to 40% for Treatment 1), but this is reversed for females. What may be more surprising is that, even with marginal independence, either treatment may still be better for all levels of the covariate. Consider, for example, Table 5.

In Table 5 we see that more males take Treatment 0 (62%), more females take Treatment 1 (62%), and

Table 4 Compensating dependence between variables X and Y by gender

Males

$y \backslash x$	0	1	Sum
0	20	30	50
1	30	20	50
Sum	50	50	100

Females

$y \backslash x$	0	1	Sum
0	30	20	50
1	20	30	50
Sum	50	50	100

Table 5 Separation of variables X and Y by gender

Males

$y \backslash x$	0	1	Sum
0	18	2	20
1	44	36	80
Sum	62	38	100

Females

$y \backslash x$	0	1	Sum
0	32	48	80
1	6	14	20
Sum	38	62	100

males have a better response rate than females for either treatment (44/62 vs. 6/38 for Treatment 0, and 36/38 vs. 14/62 for Treatment 1). However, it is also apparent that Treatment 1 is more effective than Treatment 0, both for males (36/38 vs. 44/62) and females (14/62 vs. 6/38). Such reinforcing effects can be masked when the two tables are combined to form Table 2; this is possible precisely because more of the better responders (males) took the worse treatment (Treatment 0), thereby making it look better than it should relative to the better treatment (Treatment 1). See [1], **Paradoxes** and [3] for more information regarding Simpson's paradox (*see Two by Two Contingency Tables*). Clearly, different conclusions might be reached from the same data, depending on whether the marginal table or the partial tables are considered.

References

- [1] Baker, S.G. & Kramer, B.S. (2001). Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies, *Journal of Women's Health and Gender-Based Medicine* **10**, 867–872.
- [2] Berger, V.W., Rezvani, A. & Makarewicz, V.A. (2003). Direct effect on validity of response run-in selection in clinical trials, *Controlled Clinical Trials* **24**, 156–166.
- [3] Berger, V.W. (2004). Valid adjustment of randomized comparisons for binary covariates, *Biometrical Journal* **46**(5), 589–594.
- [4] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data, *Biometric Bulletin* **2**, 47–53.

VANCE W. BERGER AND JIALU ZHANG

Marginal Models for Clustered Data

GARRETT M. FITZMAURICE

Volume 3, pp. 1128–1133

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Marginal Models for Clustered Data

Introduction

Clustered data commonly arise in studies in the behavioral sciences. For example, clustered data arise when observations are obtained on children nested within schools. Other common examples of naturally occurring clusters in the population are households, clinics, medical practices, and neighborhoods. Longitudinal and repeated measures studies (*see* **Longitudinal Data Analysis** and **Repeated Measures Analysis of Variance**) also give rise to clustered data. The main feature of clustered data that needs to be accounted for in any analysis is the fact that units from the same cluster are likely to respond in a more similar manner than units from different clusters. This **intracluster correlation** invalidates the crucial assumption of independence that is the cornerstone of so many standard statistical techniques. As a result, the straightforward application of standard regression models (e.g., **multiple linear regression** for a continuous response or **logistic regression** for a binary response) to clustered data is no longer valid unless some allowance for the clustering is made.

There are a number of ways to extend regression models to handle clustered data. All of these procedures account for the within-cluster correlation among the responses, though they differ in approach. Moreover, the method of accounting for the within-cluster association has important implications for the interpretation of the regression coefficients in models for discrete response data (e.g., binary data or counts). The need to distinguish models according to the interpretation of their regression coefficients has led to the use of the terms ‘marginal models’ and ‘mixed effects models’; the former are often referred to as ‘population-average models’, the latter as ‘cluster-specific models’ (see [3, 6, 8]). In this article, we focus on marginal models for clustered data; the meaning of the term ‘marginal’, as used in this context, will soon be apparent.

Marginal Models

Marginal models are primarily used to make inferences about population means. A distinctive feature

of marginal models is that they separately model the mean response and the within-cluster association among the responses. The goal is to make inferences about the former, while the latter is regarded as a nuisance characteristic of the data that must be accounted for to make correct inferences about the population mean response. The term *marginal* in this context is used to emphasize that we are modeling each response within a cluster separately and that the model for the mean response depends only on the covariates of interest, and not on any random effects or on any of the other responses within the same cluster. This is in contrast to *mixed effects models* where the mean response depends not only on covariates but also on random effects (*see* **Generalized Linear Mixed Models; Hierarchical Models**).

In our discussion of marginal models, the main focus is on discrete response data, for example, binary responses and counts. However, it is worth stressing that marginal models can be applied equally to continuous response data. Before we describe the main features of marginal models, we need to introduce some notation for clustered data. Let i index units within a cluster and j index clusters. We let Y_{ij} denote the response on the i th unit within the j th cluster; the response can be continuous, binary, or a count. For example, Y_{ij} might be the outcome for the i th individual in the j th clinic or the i th repeated measurement on the j th subject. Associated with each Y_{ij} is a collection of p covariates, $X_{ij1}, \dots, X_{ijp}$. We can group these into a $p \times 1$ vector denoted by $X_{ij} = (X_{ij1}, \dots, X_{ijp})'$. A marginal model has the following three-part specification:

- (1) The conditional mean of each response, denoted by $\mu_{ij} = E(Y_{ij}|X_{ij})$, is assumed to depend on the p covariates through a known ‘link function’,

$$g(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \dots + \beta_p X_{ijp}. \quad (1)$$

The link function applies a known transformation to the mean, $g(\mu_{ij})$, and then links the transformed mean to a linear combination of the covariates (*see* **Generalized Linear Models (GLM)**).

2 Marginal Models for Clustered Data

- (2) Given the covariates, the variance of each Y_{ij} is assumed to depend on the mean according to

$$\text{Var}(Y_{ij}|X_{ij}) = \phi v(\mu_{ij}), \quad (2)$$

where $v(\mu_{ij})$ is a known ‘variance function’ (i.e., a known function of the mean, μ_{ij}) and ϕ is a scale parameter that may be known or may need to be estimated.

- (3) The within-cluster association among the responses, given the covariates, is assumed to be a function of additional parameters, α (and also depends upon the means, μ_{ij}).

In less technical terms, marginal models take a suitable transformation of the mean response (e.g., a logit transformation for binary responses or a log transformation for count data) and relate the transformed mean response to the covariates. Note that the model for the mean response allows each response within a cluster to depend on covariates but not on any random effects or on any of the other responses within the same cluster. The use of nonlinear link functions, for example, $\log(\mu_{ij})$, ensures that the model produces predictions of the mean response that are within the allowable range. For example, when analyzing a binary response, μ_{ij} has interpretation in terms of the probability of ‘success’ (with $0 < \mu_{ij} < 1$). If the mean response, here the probability of success, is related directly to a linear combination of the covariates, the regression model can yield predicted probabilities outside of the range from 0 to 1. The use of certain nonlinear link functions, for example, $\text{logit}(\mu_{ij})$, ensures that this cannot happen.

This three-part specification of a marginal model can be considered a natural extension of **generalized linear models**. Generalized linear models are a collection of regression models for analyzing diverse types of univariate responses (e.g., continuous, binary, counts). They include as special cases the standard linear regression and **analysis of variance** (ANOVA) models for a normally distributed continuous response, logistic regression models for a binary response, and **log-linear models** or Poisson regression models for counts (see **Generalized Linear Models (GLM)**). Although generalized linear models encompass a much broader range of regression models, these three are among the most widely used regression models in applications. Marginal models can be thought of as extensions

of generalized linear models, developed for the analysis of independent observations, to the setting of clustered data. The first two parts correspond to the standard generalized linear model, albeit with no distributional assumptions about the responses. It is the third component, the incorporation of the within-cluster association among the responses, that represents the main extension of generalized linear models. To highlight the main components of marginal models, we consider some examples using the three-part specification given earlier.

Example 1. Marginal model for counts Suppose that Y_{ij} is a count and we wish to relate the mean count (or expected rate) to the covariates. Counts are often modeled as Poisson random variables (see **Catalogue of Probability Density Functions**), using a log link function. This motivates the following illustration of a marginal model for Y_{ij} :

- (1) The mean of Y_{ij} is related to the covariates through a log link function,

$$\log(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}. \quad (3)$$

- (2) The variance of each Y_{ij} , given the effects of the covariates, depends on the mean response,

$$\text{Var}(Y_{ij}|X_{ij}) = \phi \mu_{ij}, \quad (4)$$

where ϕ is a scale parameter that needs to be estimated.

- (3) The within-cluster association among the responses is assumed to have an exchangeable correlation (or equi-correlated) pattern,

$$\text{Corr}(Y_{ij}, Y_{i'j}|X_{ij}, X_{i'j}) = \alpha \text{ (for } i \neq i'), \quad (5)$$

where i and i' index two distinct units within the j th cluster.

The marginal model specified above is a log-linear regression model (see **Log-linear Models**), with an extra-Poisson variance assumption. The within-cluster association is specified in terms of a single correlation parameter, α . In this example, the extra-Poisson variance assumption allows the variance to be inflated (relative to Poisson variability) by a factor ϕ (when $\phi > 1$). In many applications, count data have variability that far exceeds that predicted by

the Poisson distribution; a phenomenon referred to as *overdispersion*.

Example 2. Marginal model for a binary response

Next, suppose that Y_{ij} is a binary response, taking values of 0 (denoting ‘failure’) or 1 (denoting ‘success’), and it is of interest to relate the mean of Y_{ij} , $\mu_{ij} = E(Y_{ij}|X_{ij}) = \Pr(Y_{ij} = 1|X_{ij})$, to the covariates. The distribution of each Y_{ij} is Bernoulli (*see Catalogue of Probability Density Functions*) and the probability of success is often modeled using a logit link function. Also, for a Bernoulli random variable, the variance is a known function of the mean. This motivates the following illustration of a marginal model for Y_{ij} :

- (1) The mean of Y_{ij} , or probability of success, is related to the covariates by a logit link function,

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}. \quad (6)$$

- (2) The variance of each Y_{ij} depends only on the mean response (i.e., $\phi = 1$),

$$\text{Var}(Y_{ij}|X_{ij}) = \mu_{ij}(1 - \mu_{ij}). \quad (7)$$

- (3) The within-subject association among the vector of repeated responses is assumed to have an exchangeable log odds ratio pattern,

$$\log \text{OR}(Y_{ij}, Y_{i'j}|X_{ij}, X_{i'j}) = \alpha, \quad (8)$$

where

$$\begin{aligned} & \text{OR}(Y_{ij}, Y_{i'j}|X_{ij}, X_{i'j}) \\ &= \frac{\Pr(Y_{ij} = 1, Y_{i'j} = 1|X_{ij}, X_{i'j}) \times \Pr(Y_{ij} = 0, Y_{i'j} = 0|X_{ij}, X_{i'j})}{\Pr(Y_{ij} = 1, Y_{i'j} = 0|X_{ij}, X_{i'j}) \times \Pr(Y_{ij} = 0, Y_{i'j} = 1|X_{ij}, X_{i'j})}. \end{aligned} \quad (9)$$

The marginal model specified above is a logistic regression model, with a Bernoulli variance assumption, $\text{Var}(Y_{ij}|X_{ij}) = \mu_{ij}(1 - \mu_{ij})$. An exchangeable within-cluster association is specified in terms of pairwise log odds ratios rather than correlations, a natural metric of association for binary responses (*see Odds and Odds Ratios*).

These two examples are purely illustrative. They demonstrate how the choices of the three components of a marginal model might differ according to the type of response variable. However, in principle, any suitable link function can be chosen and alternative assumptions about the variances and within-cluster associations can be made.

Finally, our description of marginal models does not require distributional assumptions for the observations, only a regression model for the mean response. In principle, this three-part specification of a marginal model can be extended by making full distributional assumptions about the responses within a cluster. For discrete data, the joint distribution requires specification of the mean vector and pairwise (or two-way) associations, as well as the three-, four- and higher-way associations among the responses (see, for example, [1, 2, 4]). Furthermore, as the number of responses within a cluster increases, the number of association parameters proliferates rapidly. In short, with discrete data there is no simple analog of the multivariate normal distribution (*see Catalogue of Probability Density Functions*). As a result, specification of the joint distribution for discrete data is inherently difficult and **maximum likelihood estimation** can be computationally difficult except in very simple cases. Fortunately, assumptions about the joint distribution are not necessary for estimation of the parameters of the marginal model. The avoidance of distributional assumptions leads to a method of estimation known as *generalized estimating equations* (GEE) (*see Generalized Estimating Equations (GEE)*). The GEE approach provides a convenient alternative to maximum likelihood estimation for estimating the parameters of marginal models (see [5, 7]) and has been implemented in many of the commercially available statistical software packages, for example, SAS, Stata, S-Plus, SUDAAN, and GenStat (*see Software for Statistical Analyses*).

Contrasting Marginal and Mixed Effects Models

A crucial aspect of marginal models is that the mean response and within-cluster association are modeled separately. This separation has important implications for interpretation of the regression coefficients. In particular, the regression coefficients in the model for the mean have *population-averaged* interpretations.

4 Marginal Models for Clustered Data

That is, the regression coefficients describe the effects of covariates on the mean response, where the mean response is *averaged* over the clusters that comprise the target population; hence, they are referred to as *population-averaged* effects. For example, the regression coefficients might have interpretation in terms of contrasts of the overall mean responses in certain subpopulations (e.g., different intervention groups).

In contrast, *mixed effects models* account for clustering in the data by assuming there is natural heterogeneity across clusters in a subset of the regression coefficients. Specifically, a subset of the regression coefficients are assumed to vary across clusters according to some underlying distribution; these are referred to as *random effects* (see **Fixed and Random Effects**). The correlation among observations within a cluster arises from their sharing of common random effects. Although the introduction of random effects can simply be thought of as a means of accounting for the within-cluster correlation, it has important implications for the interpretation of the regression coefficients. Unlike marginal models, the regression coefficients in mixed effects models have cluster-specific, rather than population-averaged, interpretations. That is, due to the nonlinear link functions (e.g., logit or log) that are usually adopted for discrete responses, the regression coefficients in mixed effects models describe covariate effects on the mean response for a typical cluster.

The subtle distinctions between these two classes of models for clustered data can be illustrated in the following example based on a pre-post study design with a binary response (e.g., denoting ‘success’ or ‘failure’). Suppose individuals are measured at baseline (pretest) and following an intervention intended to increase the probability of success (posttest). The ‘cluster’ is comprised of the pair of binary responses obtained on the same individual at baseline and post-baseline. These clustered data can be analyzed using a mixed effects logistic regression model,

$$\text{logit}\{E(Y_{ij}|X_{ij}, b_j)\} = \beta_0^* + \beta_1^* X_{ij} + b_j, \quad (10)$$

where b_j is normally distributed, with mean zero and variance, σ_b^2 . The single covariate in this model takes values $X_{ij} = 0$ at baseline and $X_{ij} = 1$ post-baseline (see **Dummy Variables**). In this mixed effects model, β_0^* and β_1^* are the *fixed effects* and b_j is the *random effect*. This model assumes that individuals

(or clusters) differ in terms of their underlying propensity for success; this heterogeneity is expressed in terms of the variability of the random effect, b_j . For a ‘typical’ individual from the population (where a ‘typical’ individual can be thought of as one with unobserved random effect $b_j = 0$, the center of the distribution of b_j), the log odds of success at baseline is β_0^* ; the log odds of success following the intervention is $\beta_0^* + \beta_1^*$.

The log odds of success at baseline and postbaseline are displayed in Figure 1, for the case where $\beta_0^* = -1.75$, $\beta_1^* = 3.0$, and $\sigma_b^2 = 1.5$. At baseline, the log odds has a normal distribution with mean and median of -1.75 (see the unshaded density for the log odds in Figure 1). From Figure 1 it is clear that there is heterogeneity in the odds of success, with approximately 95% of individuals having a baseline log odds of success that varies from -4.15 to 0.65 (or $-1.75 \pm 1.96\sqrt{1.5}$). When the odds of success is translated to the probability scale (see vertical axis of

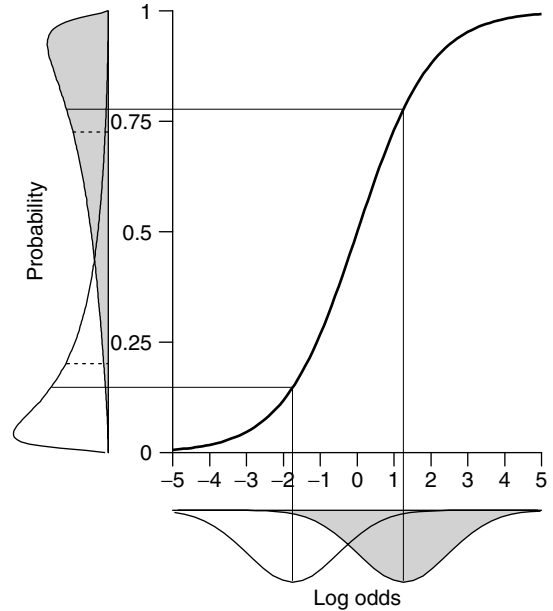


Figure 1 Subject-specific probability of success as a function of subject-specific log odds of success at baseline (unshaded densities) and post-baseline (shaded densities). Solid lines represent medians of the distributions; dashed lines represent means of the distributions.

Figure 1),

$$\begin{aligned} E(Y_{ij}|X_{ij}, b_j) &= \Pr(Y_{ij} = 1|X_{ij}, b_j) \\ &= \frac{e^{\beta_0^* + \beta_1^* X_{ij} + b_j}}{1 + e^{\beta_0^* + \beta_1^* X_{ij} + b_j}}, \end{aligned} \quad (11)$$

the baseline probability of success for a typical individual (i.e., an individual with $b_j = 0$) from the population is 0.148 (or $e^{-1.75}/1 + e^{-1.75}$). Furthermore, approximately 95% of individuals have a baseline probability of success that varies from 0.016 to 0.657.

From Figure 1 it is transparent that the symmetric, normal distribution for the baseline log odds of success does not translate into a corresponding symmetric, normal distribution for the probability of success. Instead, the subject-specific probabilities have a positively skewed distribution with median, but not mean, of 0.148 (see solid line in Figure 1). Because of the skewness, the mean of the subject-specific baseline probabilities is pulled towards the tail and is equal to 0.202 (see dashed line in Figure 1). Thus, the probability of success for a ‘typical’ individual from the population (0.148) is not the same as the prevalence of success in the same population (0.202), due to the nonlinearity of the relationship between subject-specific probabilities and log odds. Similarly, although the log odds of success postbaseline has a normal distribution (see the shaded density for the log odds in Figure 1), the subject-specific post-baseline probabilities have a negatively skewed distribution with median, but not mean, of 0.777 (see solid line in Figure 1). Because of the skewness, the mean is pulled towards the tail and is equal to 0.726 (see dashed line in Figure 1).

Figure 1 highlights how the effect of intervention on the log odds of success for a typical individual (or cluster) from the population, $\beta_1^* = 3.0$, is not the same as the contrast of population log odds. The latter is what is estimated in a marginal model, say

$$\text{logit}\{E(Y_{ij}|X_{ij})\} = \beta_0 + \beta_1 X_{ij}, \quad (12)$$

and can be obtained by comparing or contrasting the log odds of success in the population at baseline, $\log(0.202/0.798) = -1.374$, with the log odds of success in the population postbaseline, $\log(0.726/0.274) = 0.974$. This yields a *population-averaged* measure of effect, $\beta_1 = -2.348$, which is approximately 22% smaller than β_1^* , the *subject-specific* (or *cluster-specific*) effect of intervention.

This simple example highlights how the choice of method for accounting for the within-cluster association has consequences for the interpretation of the regression model parameters.

Summary

Marginal models are widely used for the analysis of clustered data. They are most useful when the focus of inference is on the overall population mean response, averaged over all the clusters that comprise the population. The distinctive feature of marginal models is that they model each response within the cluster separately. They assume dependence of the mean response on covariates but not on any random effects or other responses within the same cluster. This is in contrast to mixed effects models where the mean response depends not only on covariates but on random effects.

There are a number of important distinctions between marginal and mixed effects models that go beyond simple differences in approaches to accounting for the within-cluster association among the responses. In particular, these two broad classes of regression models have somewhat different targets of inference and have regression coefficients with distinct interpretations. In general, the choice of method for analyzing discrete clustered data cannot be made through any automatic procedure. Rather, it must be made on subject-matter grounds. Different models for discrete clustered data have somewhat different targets of inference and thereby address subtly different scientific questions regarding the dependence of the mean response on covariates.

References

- [1] Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items, in *Studies in Item Analysis and Prediction*, H. Solomon, ed., Stanford University Press, Stanford, pp. 158–168.
- [2] George, E.O. & Bowman, D. (1995). A full likelihood procedure for analyzing exchangeable binary data, *Biometrics* **51**, 512–523.
- [3] Graubard, B.I. & Korn, E.L. (1994). Regression analysis with clustered data, *Statistics in Medicine* **13**, 509–522.
- [4] Kupper, L.L. & Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics* **34**, 69–76.
- [5] Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**, 13–22.

6 Marginal Models for Clustered Data

- [6] Neuhaus, J.M., Kalbfleisch, J.D. & Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review* **59**, 25–35.
- [7] Zeger, S.L. & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**, 121–130.
- [8] Zeger, S.L., Liang, K.-Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**, 1049–1060.

GARRETT M. FITZMAURICE

Markov Chain Monte Carlo and Bayesian Statistics

PETER CONGDON

Volume 3, pp. 1134–1143

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Markov Chain Monte Carlo and Bayesian Statistics

Introduction

Bayesian inference (*see* **Bayesian Statistics**) differs from classical inference in treating parameters as random variables, one consequence being probability values on hypotheses, and confidence intervals on parameters (or ‘credible intervals’ for Bayesians), that are concordant with commonsense interpretations [31]. Thus, in the Bayesian paradigm, credible intervals are areas under the probability distributions of the parameters involved. Despite inferential and philosophical advantages of this kind, practical application of Bayesian methods was formerly prohibited by the need for numerical integration, except in straightforward problems with a small number of parameters. This problem has been overcome in the last 15 years by the advent of simulation methods known as Markov Chain Monte Carlo (MCMC) algorithms; *see* [2, 21, 49]. MCMC sample-based estimation methods overcome problems associated with numerical procedures that were in use in the 1980s. They can handle high-dimensional problems and explore the distributions of parameters, regardless of the forms of the distributions of the likelihood and the parameters. Using MCMC methods, one can obtain compute posterior summary statistics (means, variances, 95% intervals) for parameters or other structural quantities. Starting from postulated or ‘prior’ distributions of the parameters, improved or ‘posterior’ estimates of the distributions are obtained by randomly sampling from parameter distributions in turn and updating the parameter values until stable distributions are generated.

The implementation of these methods has resulted in Bayesian methods actually being easier to apply than ‘classical’ methods to some complex problems with large numbers of parameters, for example, those involving multiple random effects (*see* **Generalized Linear Mixed Models**) [6]. Such methods are now available routinely via software such as WINBUGS (*see* the page at <http://www.mrc-bsu.ac.uk> for programs and examples). Bayesian data analysis (especially via modern MCMC methods) has

a range of other advantages such as the ability to combine inferences over different models when no model is preeminent in terms of fit (‘model averaging’) and permitting comparisons of fit between nonnested models. By contrast, classical hypothesis testing is appropriate for testing nested models that differ only with respect to the inclusion/exclusion of particular parameters.

Bayesian inference can be seen as a process of learning about parameters. Thus, Bayesian learning does not undertake statistical analysis in isolation but draws on existing knowledge in prior framing of the model, which can be quite important (and beneficial) in clinical, epidemiological, and health applications [48, Chapter 5]. The estimation process then combines existing evidence with the actual study data at hand. The result can be seen as a form of evidence accumulation. Prior knowledge about parameters and updated knowledge about them are expressed in terms of densities: the data analysis converts existing knowledge expressed in a *prior* density to updated knowledge expressed in a *posterior* density (*see* **Bayesian Statistics**). The Bayesian analysis also produces predictions (e.g., the predicted response for new values of independent variables, as in the case of predicting a relapse risk for a hypothetical patient with an ‘average’ profile). It also provides information about functions of parameters and data (e.g., a ratio of two regression coefficients). One may also assess hypotheses about parameters (e.g., that a regression coefficient is positive) by obtaining a posterior probability relating to the hypothesis being true.

Modern sampling methods also allow for representing densities of parameters that may be far from Normal (e.g., skew or multimodal), whereas **maximum likelihood estimation** and classical tests rely on asymptotic Normality approximations under which a parameter has a symmetric density. In addition to the ease with which exact densities of parameters (i.e., possibly asymmetric) are obtained, other types of inference may be simplified by using the sampling output from an MCMC analysis. For example, a hypothesis that a regression parameter is positive is assessed by the proportion of iterations where the sampled value of the parameter is positive. Bayesian methods are advantageous for random effects models in which ‘pooling strength’ acts to provide more reliable inferences about individual cases. This has relevance in applications such as multilevel

analysis (*see* **Generalized Linear Mixed Models**) and **meta-analysis** in which, for example, institutions or treatments are being compared. A topical UK illustration relates to a Bayesian analysis of child surgery deaths in Bristol, summarized in a Health Department report [16]; *see* [40].

Subsequent sections consider modern Bayesian methods and inference procedures in more detail. We first consider the principles of Bayesian methodology and the general principles of MCMC estimation. Then, sampling algorithms are considered as well as possible ‘problems for the unwary’ in terms of convergence and identifiability raised when models become increasingly complex. We next consider issues of prior specification (a distinct feature of the Bayes approach) and questions of model assessment. A worked example using data on drug regimes for schizophrenic patients involves the question of missing data: this frequently occurs in **panel studies** (in the form of ‘attrition’), and standard approaches such as omitting subjects with missing data lead to biased inference.

Bayesian Updating

In more formal terms, the state of existing knowledge about a parameter, or viewed another way, a statement about the uncertainty about a parameter set $\theta = (\theta_1, \theta_2, \dots, \theta_d)$ is expressed in a prior density $p(\theta)$. The likelihood of a particular set of data y , given θ , is denoted $p(y|\theta)$. For example, the standard constant variance linear regression (*see* **Multiple Linear Regression**) with p predictors (including the intercept) involves a Normal likelihood with $\theta = (\beta, \sigma^2)$, where β is a vector of length p and σ^2 is the error variance. The updated or posterior uncertainty about θ , having taken account of the data, is expressed in the density $p(\theta|y)$. From the usual conditional probability rules (*see* **Probability: An Introduction**), this density can be written

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)}. \quad (1)$$

Applying conditional probability again gives

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \quad (2)$$

The divisor $p(y)$ is a constant and the updating process therefore takes the form

$$p(\theta|y) \propto p(y|\theta)p(\theta). \quad (3)$$

This can be stated as ‘the posterior is proportional to likelihood times prior’ or, in practical terms, updated knowledge about parameters combines existing knowledge with the sample data at hand.

The Place of MCMC Methods

The genesis of estimation via MCMC sampling methods lies in the need to obtain expectations of, or densities for, functions of parameters $g(\theta)$, and of model predictions z , that take account of the information in both the data and the prior. Functions $g(\theta)$ might, for instance, be differences in probabilities of relapse under different treatments, where the probabilities are predicted by a **logistic regression** or the total of sensitivity and specificity, sometimes known as the *Youden Index* [10]. These are standard indicators of the effectiveness of screening instruments, for example [33]; the sensitivity is the proportion of cases correctly identified and the specificity is the probability that the test correctly identifies that a healthy individual is disease free.

The expectation of $g(\theta)$ is obtained by integrating over the posterior density for θ

$$E_{\theta|y}[g(\theta)] = \int g(\theta)p(\theta|y) d\theta, \quad (4)$$

while the density for future observations (‘replicate data’) is

$$p(z|y) = \int p(z|\theta, y)p(\theta|y) d\theta. \quad (5)$$

Often, the major interest is in the marginal densities of the parameters themselves, where the marginal density of the j th parameter θ_j is obtained by integrating out all other parameters

$$p(\theta_j|y) = \int p(\theta|y) d\theta_1 d\theta_2 \dots d\theta_{j-1} d\theta_{j+1} \dots d\theta_d. \quad (6)$$

Such expectations or densities may be obtained analytically for conjugate analyses such as a binomial likelihood where the probability has a beta prior (*see* **Catalogue of Probability Density Functions**). (A conjugate analysis occurs when the prior and posterior densities for θ_j have the same form, e.g., both Normal or both beta.) Results can be obtained by asymptotic approximations [5], or by analytic approximations [34]. Such approximations are not appropriate for posteriors that are non-Normal or

where there is multimodality. An alternative strategy facilitated by contemporary computer technology is to use sampling-based approximations based on the Monte Carlo principle. The idea here is to use repeated sampling to build up a picture of marginal densities such as $p(\theta_j|y)$: modal values of the density will be sampled most often, those in the tails relatively infrequently.

The Monte Carlo method in general applications assumes a sample of independent simulations $u^{(1)}, u^{(2)} \dots u^{(T)}$ from a density $p(u)$, whereby $E[g(u)]$ is estimated as

$$\bar{g}_T = \frac{\sum_{t=1}^T g(u^{(t)})}{T}. \quad (7)$$

However, in MCMC applications, independent sampling from the posterior target density $p(\theta|y)$ is not typically feasible. It is valid, however, to use dependent samples $\theta^{(t)}$, provided the sampling satisfactorily covers the support of $p(\theta|y)$ [28]. MCMC methods generate pseudorandom-dependent draws from probability distributions such as $p(\theta|y)$ via Markov chains. Sampling from the Markov chain converges to a stationary distribution, namely, $\pi(\theta) = p(\theta|y)$, if certain requirements on the chain are satisfied, namely irreducibility, aperiodicity, and positive recurrence¹; see [44]. Under these conditions, the average $\bar{g}_T$ tends to $E_\pi[g(u)]$ with probability 1 as $T \rightarrow \infty$, despite dependent sampling. Remaining practical questions include establishing an MCMC sampling scheme and establishing that convergence to a steady state has been obtained [14].

MCMC Sampling Algorithms

The Metropolis–Hastings (M–H) algorithm [29] is the baseline for MCMC sampling schemes. Let $\theta^{(t)}$ be the current parameter value in an MCMC sampling sequence. The M–H algorithm assesses whether a move to a potential new parameter value θ^* should be made on the basis of the transition probability

$$\alpha(\theta^*|\theta^{(t)}) = \min \left(1, \frac{p(\theta^*|y)f(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|y)f(\theta^*|\theta^{(t)})} \right). \quad (8)$$

The density f is known as a proposal or jumping density. The rate at which a proposal generated by f is accepted (the acceptance rate) depends on how

close θ^* is to $\theta^{(t)}$ and this depends on the variance σ^2 assumed in the proposal density. For a Normal proposal density, a higher acceptance rate follows from reducing σ^2 but with the risk that the posterior density will take longer to explore. If the proposal density is symmetric $f(\theta^{(t)}|\theta^*) = f(\theta^*|\theta^{(t)})$, then the M–H algorithm reduces to an algorithm used by Metropolis et al. [41] for indirect simulation of energy distributions, whereby

$$\alpha(\theta^*|\theta^{(t)}) = \min \left[1, \frac{p(\theta^*|y)}{p(\theta^{(t)}|y)} \right]. \quad (9)$$

While it is possible for the proposal density to relate to the entire parameter set, it is often computationally simpler to divide θ into blocks or components, and use componentwise updating.

Gibbs sampling is a special case of the componentwise M–H algorithm, whereby the proposal density for updating θ_j is the full conditional density $f_j(\theta_j|\theta_{k \neq j})$, so that proposals are accepted with probability 1. This sampler was originally developed by Geman and Geman [24] for Bayesian image reconstruction, with its full potential for simulating marginal distributions by repeated draws recognized by Gelfand and Smith [21]. The Gibbs sampler involves parameter-by-parameter updating, which when completed forms the transition from $\theta^{(t)}$ to $\theta^{(t+1)}$:

1. $\theta_1^{(t+1)} \sim f_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_d^{(t)});$
2. $\theta_2^{(t+1)} \sim f_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_d^{(t)});$
- d. $\theta_d^{(t+1)} \sim f_d(\theta_d|\theta_1^{(t+1)}, \theta_3^{(t+1)}, \dots, \theta_{d-1}^{(t+1)}).$

Repeated sampling from M–H samplers such as the Gibbs sampler generates an autocorrelated sequence of numbers that, subject to regularity conditions (ergodicity, etc.), eventually ‘forgets’ the starting values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})$ used to initialize the chain and converges to a stationary sampling distribution $p(\theta|y)$.

Full conditional densities can be obtained by abstracting out from the full model density (likelihood times prior) those elements including θ_j and treating other components as constants [26]. Consider a conjugate model for Poisson count data y_i with means μ_i that are themselves gamma distributed. This is a model appropriate for overdispersed count data with actual variability $\text{var}(y)$ exceeding that under

the Poisson model. This sort of data often occurs because of variations in frailty, susceptibility, or ability between subjects; a study by Bockenholt et al. provides an illustration involving counts of emotional experiences [7]. Thus, neuroticism is closely linked to proneness to experience unpleasant emotions, while extraversion is linked with sociability, enthusiasm, and pleasure arousal: hence, neuroticism and extraversion are correlated with counts of intense unpleasant and pleasant emotions. For example, $y_i \sim Po(\mu_i t_i)$ might be the numbers of emotions experienced by a particular person in a time interval t_i and μ_i the average emotion count under the model. Suppose variations in proneness follow a gamma density $\mu_i \sim Ga(\alpha, \beta)$, namely,

$$f(\mu_i|\alpha, \beta) = \frac{\mu_i^{\alpha-1} e^{-\beta\mu_i} \beta^\alpha}{\Gamma(\alpha)} \quad (10)$$

and further that $\alpha \sim E(a)$, $\beta \sim Ga(b, c)$, where a , b , and c are preset constants; this prior structure is used by George et al. [25]. The posterior density $p(\theta|y)$, of $\theta = (\mu_1, \dots, \mu_n, \alpha, \beta)$ given y , is then proportional to

$$e^{-a\alpha} \beta^{b-1} e^{-c\beta} \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \right]^n \left[\prod_{i=1}^n e^{-\mu_i t_i} \mu_i^{y_i} \mu_i^{\alpha-1} e^{\beta\mu_i} \right],$$

since all constants (such as the denominator $y_i!$ in the Poisson likelihood) are included in the proportionality constant. The conditional densities of μ_i and β are $f_1(\mu_i|\alpha, \beta) = Ga(y_i + \alpha, \beta + t_i)$ and $f_2(\beta|\alpha, \mu_i) = Ga(b + n\alpha, c + \Sigma\mu_i)$ respectively, while that of α is

$$f_3(\alpha|\beta, \mu) \propto e^{-a\alpha} \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \right]^n \left(\prod_{i=1}^n \mu_i \right)^{\alpha-1}. \quad (11)$$

This density is nonstandard but log-concave and cannot be sampled directly. However, adaptive rejection sampling [27] may be used. By contrast, sampling from the gamma densities for μ_i and β is straightforward. For a Gibbs sampling MCMC application, we would repeatedly sample $\mu_i^{(t+1)}$ from f_1 conditional on $\alpha^{(t)}$ and $\beta^{(t)}$, then $\beta^{(t+1)}$ from f_2 conditional on $\mu_i^{(t+1)}$ and $\alpha^{(t)}$, and $\alpha^{(t+1)}$ from f_3 conditional on $\mu_i^{(t+1)}$ and $\beta^{(t+1)}$. By repeated sampling of $\mu_i^{(t)}$, $\alpha^{(t)}$, and $\beta^{(t)}$, for iterations $t = 1, \dots, T$, we approximate the marginal densities of these parameters, with the approximation improving as T increases.

Convergence and Identifiability

There are many unresolved questions around the assessment of convergence of MCMC sampling procedures [14]. It is preferable to use two or more parallel chains with diverse starting values to ensure full coverage of the sample space of the parameters, and to diminish the chance that the sampling will become trapped in a small part of the space [23]. Single long runs may be adequate for straightforward problems, or as a preliminary to obtain inputs to multiple chains. Convergence for multiple chains may be assessed using Gelman–Rubin scale reduction factors (SRF) that compare variation in the sampled parameter values within and between chains. Parameter samples from poorly identified models will show wide divergence in the sample paths between different chains and variability of sampled parameter values between chains will considerably exceed the variability within any one chain. In practice, SRFs under 1.2 are taken as indicating convergence.

Analysis of sequences of samples from an MCMC chain amounts to an application of time-series methods, in regard to problems such as assessing stationarity in an autocorrelated sequence. Autocorrelation at lags 1, 2, and so on may be assessed from the full set of sampled values $\theta^{(t)}$, $\theta^{(t+1)}$, $\theta^{(t+2)}$, $\dots$, or from subsamples K steps apart $\theta^{(t)}$, $\theta^{(t+K)}$, $\theta^{(t+2K)}$, $\dots$, and so on. If the chains are mixing satisfactorily, then the autocorrelations in the $\theta^{(t)}$ will fade to zero as the lag increases (e.g., at lag 10 or 20). High autocorrelations that do not decay mean that less information about the posterior distribution is provided by each iteration and a higher sample size T is necessary to cover the parameter space.

Problems of convergence in MCMC sampling may reflect problems in model identifiability due to overfitting or redundant parameters. Choice of diffuse priors increases the chance of poorly identified models, especially in complex hierarchical models or small samples [20]. Elicitation of more informative priors or application of parameter constraints may assist identification and convergence. Slow convergence will show in poor ‘mixing’ with high autocorrelation between successive sampled values of parameters, and trace plots that wander rather than rapidly fluctuating around a stable mean. Conversely, running multiple chains often assists in diagnosing poor identifiability of models. Correlation between parameters tends to delay convergence

and increase the dependence between successive iterations. Reparameterization to reduce correlation – such as centering predictor variables in regression – usually improves convergence [22].

Choice of Prior

Choice of the prior density is an important issue in Bayesian inference, as inferences may be sensitive to the choice of prior, though the role of the prior may diminish as sample size increases. From the viewpoint of a subject-matter specialist, the prior is the way to include existing subject-matter knowledge and ensure the analysis is ‘evidence consistent’. There may be problems in choosing appropriate priors for certain types of parameters: variance parameters in random effects models are a leading example [15].

A long running (and essentially unresolvable) debate in Bayesian statistics revolves around the choice between objective ‘off-the-shelf’ priors, as against ‘subjective’ priors that may include subject-matter knowledge. There may be a preference for off-the-shelf or reference priors that remove any subjectivity in the analysis. In practice, just proper but still diffuse priors are a popular choice (see the WINBUGS manuals for several examples), and a sensible strategy is to carry out a sensitivity analysis with a range of such priors. Just proper priors are those that still satisfy the conditions for a valid density: they integrate to 1, whereas ‘improper’ priors do not. However, they are diffuse in the sense of having a large variance that does not contain any real information about the location of the parameter.

Informative subjective priors based on elicited opinion from scientific specialists, historical studies, or the weight of established evidence, can also be justified. One may even carry out a preliminary evidence synthesis using forms of meta-analysis to set an informative prior; this may be relevant for treatment effects in clinical studies or disease prevalence rates [48]. Suppose the unknown parameter is the proportion π of parasuicide patients making another suicide attempt within a year of the index event. There is considerable evidence on this question, and following a literature review, a prior might be set with the mean recurrence rate 15%, and 95% range between 7.5 and 22.5%. This corresponds approximately to a beta prior with parameters 7.5 and 42.5. A set of previous representative studies might be used

more formally in a form of meta-analysis, though if there are limits to the applicability of previous studies to the current target population, the information from previous studies may be down-weighted in some way. For example, the precision of an estimated relative risk or prevalence rate from a previous study may be halved. A diffuse prior on π might be a $\text{Be}(1,1)$ prior, which is in fact equivalent to assuming a proportion uniformly distributed between 0 and 1.

Model Comparison and Criticism

Having chosen a prior and obtained convergence for a set of alternative models, one is faced with choosing between models (or possibly combining inferences over them) and with diagnostic checking (e.g., assessing outliers). Several methods have been proposed for model choice and diagnosis based on Bayesian principles. These include many features of classical model assessment such as penalizing complexity and requiring accurate predictions (i.e., cross-validation). To develop a Bayesian adaptation of frequentist model choice via the AIC, Spiegelhalter et al. [47] propose estimating the effective total number of parameters or model dimension, d_e . Thus, d_e is the difference between the mean $\bar{D}$ of the sampled deviances $D^{(t)}$ and the deviance $D(\bar{\theta}|y)$ at the parameter posterior mean $\bar{\theta}$. This total is generally less than the nominal number of parameters in complex hierarchical random effects models (see **Generalized Linear Mixed Models**). The deviance information criterion is then $\text{DIC} = D(\bar{\theta}|y) + 2d_e$. Congdon [12] considers repeated parallel sampling of models to obtain the density of $\Delta\text{DIC}_{jk} = \text{DIC}_j - \text{DIC}_k$ for models j and k .

Formal Bayesian model assessment is based on updating prior model probabilities $P(M_j)$ to posterior model probabilities $P(M_j|Y)$ after observing the data. Imagine a scenario to obtain probabilities of a particular model being true, given a set of data Y , and assuming that one of J models, including the model in question, is true; or possibly that truth resides in averaging over a subset of the J models. Then

$$P(M_j|Y) = \frac{[P(M_j)P(Y|M_j)]}{\sum_{j=1}^J [P(M_j)P(Y|M_j)]}, \quad (12)$$

where $P(Y|M_j)$ are known as marginal likelihoods. Approximation methods for $P(Y|M_j)$ include those

presented by Gelfand and Dey [19], Newton and Raftery [42], and Chib [8]. The Bayes factor is used for comparing one model against another under the formal approach and is the ratio of marginal likelihoods

$$B_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}. \quad (13)$$

Congdon [13] considers formal choice based on Monte Carlo estimates of $P(M_j|Y)$ without trying to approximate marginal likelihoods.

Whereas data analysis is often based on selecting a single best model and making inferences as if that model were true, such an approach neglects uncertainty about the model itself, expressed in the posterior model probabilities $P(M_j|Y)$ [30]. Such uncertainty implies that for closely competing models, inferences should be based on model-averaged estimates

$$E[g(\theta)] = \sum_j P(M_j|Y)g(\theta_j|y). \quad (14)$$

Problems with formal Bayes model choice, especially in complex models or when priors are diffuse, have led to alternative model-assessment procedures, such as those including principles of predictive cross-validation; see [18, 35].

Bayesian Applications in Psychology and Behavioral Sciences

In psychological and behavioral studies, the methods commonly applied range from clinical trial designs with individual patients as subjects to population based ecological studies. The most frequent analytic frameworks are **multivariate analysis** and **structural equation models** (SEMs), more generally, latent class analysis (LCA), diagnosis and classification studies (*see Discriminant Analysis; k-means Analysis*), and panel and **case-control studies**. Bayesian applications to psychological data analysis include SEM, LCA, and item response analysis, see [3, 9, 37, 38, 46]; and factor analysis per se [1, 4, 36, 43, 45]. Here we provide a brief overview of a Bayesian approach to missing data in panel studies, with a worked example using the WINBUGS package (code available at www.geog.qmul.ac.uk/staff/congdon.html). This package presents great analytic flexibility while leaving issues such as choice of sampling methods to an inbuilt expert system [11].

An Illustrative Application

Panel studies with data y_{it} on subjects i at times t are frequently subject to attrition (permanent loss from observation from a given point despite starting the study), or intermittently missing responses [32]. Suppose that some data are missing in a follow-up study of a particular new treatment due to early patient withdrawal. One may wonder whether this loss to observation is ‘informative’, for instance, if early exit is due to adverse consequences of the treatment. The fact that some patients do withdraw generates a new form of data, namely, binary indicators $r_{it} = 1$ when the response y_{it} is missing, and $r_{it} = 0$ when a response is present. If a subject drops out permanently at time t , they contribute to the likelihood at that point with an observation $r_{it} = 1$, but are not subsequently included in the model. If withdrawal is informative, then the probability that an exit occurs (the probability that $r = 1$) is related to y_{it} . However, this value is not observed if the patient exits, raising the problem that a model for $\Pr(r = 1)$ has to be in terms of a hypothetical value, namely, the value for y_{it} that would have been observed had the response actually happened [39]. This type of missing data mechanism is known as missingness not at random (MNAR) (*see Missing Data*).

By contrast, if missingness is random (often written as MAR, or missingness at random), then $\Pr(r_{it} = 1)$ may depend on values of observed variables, such as preceding y values ($y_{i,t-1}$, $y_{i,t-2}$, etc.), but not on the values of possibly missing variables such as y_{it} itself. Another option is missingness completely at random (abbreviated as MCAR), when none of the data collected or missing is relevant to explaining the chance of missingness (*see Missing Data*). Only in this scenario is complete case analysis valid.

In concrete modeling terms, in the MCAR model, missingness at time t is not related to any other variable, whereas in the informative model missingness could be related to any other variable, missing, or observed. Suppose we relate the probability π_{it} that $r_{it} = 1$ to current and preceding values of the response y and to a known covariate x (e.g., time in the trial). Then,

$$\text{logit}(\pi_{it}) = \eta_1 + \eta_2 y_{it} + \eta_3 y_{i,t-1} + \eta_4 x_{it} + \eta_5 x_{i,t-1} \quad (15)$$

Under the MNAR scenario, dropout at time t (causing r_{it} to be 1) may be related to the possibly

missing value of y_{it} at that time. This would mean that the current value of y influences the chance that $r = 1$ (i.e., $\eta_2 \neq 0$). Note that y_{it} is an extra unknown when $r_{it} = 1$ (i.e., the hypothetical missing value is imputed or ‘augmented’ under the model). Under a MAR scenario, by contrast, previous observed values of y , or current/preceding x may be relevant to the chance of a missing value, but the current value of y is not relevant. So η_3, η_4 , and η_5 may be nonzero but one would expect $\eta_2 = 0$. Finally, MCAR missingness would mean $\eta_2 = \eta_3 = \eta_4 = \eta_5 = 0$.

To illustrate how dropout might be modeled in panel studies, consider a longitudinal trial comparing a control group with two drug treatments for schizophrenia, haloperidol, and risperidone [17]. The response y_{it} was the Positive and Negative Symptom Scale (PANSS), with higher scores denoting more severe illness, and obtained at seven time points (selection, baseline, and at weeks 1, 2, 4, 6, and 8). Let v_t denote the number of weeks, with the baseline defined by $v_2 = 0$, and selection into the study by $v_1 = -1$. Cumulative attrition is only 0.6% at baseline but reaches 1.7%, 13.5%, 23.6%, and 39.7% in successive waves, reaching 48.5% in the final wave. The question is whether attrition is related to health status: if the dropout rate is higher for those with high PANSS scores (e.g., because they are gaining no benefit from their treatment), then observed time paths of PANSS scores are in a sense unrepresentative.

Let treatment be specified by a trichotomous indicator G_i (control, haloperidol, risperidone). We assume the data are Normal with $y_{it} \sim N(\mu_{it}, \sigma^2)$. If subject i drops out at time $t = t^*$, then $r_{it^*} = 1$ at that time but the person is excluded from the model for times $t > t^*$. The model for μ_{it} includes a grand mean M , main treatment effects $\delta_j (j = 1 \dots 3)$, linear and quadratic time effects, θ_j and γ_j , both specific to treatment j , and random terms over both subjects and individual readings:

$$\mu_{it} = M + \delta_{Gi} + \theta_{Gi}v_t + \gamma_{Gi}v_t^2 + U_i + e_{it}. \quad (16)$$

Since the model includes a grand mean, a corner constraint $\delta_1 = 0$ is needed for identifiability. The U_i are subject level indicators of unmeasured morbidity factors, sometimes called ‘permanent’ random effects. The model for e_{it} allows for dependence over time (autocorrelation), with an allowance also for the unequally spaced observations (the gap between

observations is sometimes 1 week, sometimes 2). Thus,

$$e_{it} \sim N(\rho^{[v_t - v_{t-1}]}e_{i,t-1}, \sigma_e^2). \quad (17)$$

We take ρ , the correlation between errors one week apart, to be between 0 and 1. Note that the model for y also implicitly includes an uncorrelated error term with variance σ^2 .

The dropout models considered are more basic than the one discussed above. Thus, two options are considered. The first option is a noninformative (MAR) model allowing dependence on preceding (and observed) $y_{i,t-1}$ but not on the current, possibly missing, y_{it} . So

$$\text{logit}(\pi_{it}) = \eta_{11} + \eta_{12}y_{i,t-1}. \quad (18)$$

The second adopts the alternative informative missingness scenario (MNAR), namely,

$$\text{logit}(\pi_{it}) = \eta_{21} + \eta_{22}y_{it} \quad (19)$$

since this allows dependence on (possibly missing) contemporaneous scores y_{it} . We adopt relatively diffuse priors on all the model parameters (those defining μ_{it} and π_{it}), including $N(0, 10)$ priors on δ_j and θ_j and $N(0, 1)$ priors on γ_j .

With the noninformative dropout model, similar results to those cited by Diggle [17, p. 221] are obtained (see Table 1). Dropout increases with PANSS score under the first dropout model (the coefficient η_{12} is clearly positive with 95% CI restricted to positive values), so those remaining in the trial are increasingly ‘healthier’ than the true average, and so increasingly unrepresentative. The main treatment effect for risperidone, namely δ_3 , has a negative mean (in line with the new drug producing lower average PANSS scores over all observation points) but its 95% interval is not quite conclusive. The estimates for the treatment-specific linear trends in PANSS scores (θ_1, θ_2 , and θ_3) do, however, conclusively show a fall under risperidone, and provide evidence for the effectiveness of the new drug. The quadratic effect for risperidone reflects a slowing in PANSS decline for later observations. Note that the results presented in Table 1 constitute ‘posterior summaries’ in a very simplified form.

Introducing the current PANSS score y_{it} into the model for response r_{it} makes the dropout model informative. The fit improves slightly under the predictive criterion [35] based on comparing replicate

Table 1 PANSS model parameters, alternative dropout models

	Noninformative			Informative		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Response model						
η_{11}	-5.32	-5.84	-4.81			
η_{12}	0.034	0.028	0.039			
η_{21}				-5.58	-6.28	-4.80
η_{22}				0.034	0.026	0.041
Observation model						
Main treatment effect						
δ_2	1.43	-1.95	5.02	3.32	-0.95	7.69
δ_3	-1.72	-4.29	1.49	-4.66	-8.63	-0.66
Autocorrelation parameter						
ρ	0.94	0.91	0.97	0.96	0.93	0.99
Linear time effects						
θ_1	0.04	-1.29	1.48	0.06	-0.61	0.68
θ_2	1.03	-0.11	2.25	1.15	0.38	1.88
θ_3	-1.86	-2.40	-1.16	-0.81	-1.16	-0.42
Quadratic time effects						
γ_1	-0.066	-0.258	0.117	-0.017	-0.186	0.158
γ_2	-0.066	-0.232	0.095	0.046	-0.135	0.240
γ_3	0.104	0.023	0.175	0.108	0.023	0.199

data z_{it} to actual data y_{it} . The parameter η_{22} has a 95% CI confined to positive values and suggests that missingness may be informative; to conclusively establish this, one would consider more elaborate dropout models including both y_{it} , earlier y values, $y_{i,t-1}$, $y_{i,t-2}$, and so on, and possibly treatment group too. In terms of inferences on treatment effectiveness, both main treatment and linear time effects for risperidone treatment are significantly negative under the informative model, though the time slope is less acute.

Note that maximum likelihood analysis is rather difficult for this type of problem. In addition to the fact that model estimation depends on imputed y_{it} responses when $r_{it} = 1$, it can be seen that the main effect treatment parameters under the first dropout model have somewhat skewed densities. This compromises classical significance testing and derivation of confidence limits in a maximum likelihood (ML) analysis. To carry out analysis via classical (e.g., ML) methods for such a model requires technical knowledge and programming skills beyond those of the average psychological researcher. However, using a Bayes approach, it is possible to apply this method for a fairly unsophisticated user of WINBUGS. Note that the original presentation of the informative analysis via classical methods

does not include parameter standard errors [17, Table 9.5], whereas obtaining full parameter summaries (which are anyway more extensive than simply means and standard errors) is unproblematic via MCMC sampling.

As an example of a relatively complex hypothesis test that is simple under MCMC sampling, consider the probability that $\delta_3 < \min(\delta_1, \delta_2)$, namely, that the main treatment effect (in terms of reducing the PANSS score) under risperidone is greater than either of the other two treatments. This involves inserting a single line in WINBUGS, namely,

```
test.del <- step(min(delta[1],
                      delta[2]) - delta[3])
```

and monitoring the proportion of iterations where the condition $\delta_3 < \min(\delta_1, \delta_2)$ holds. Under the informative dropout model, we find this probability to be .993 (based on the second half of a two-chain run of 10 000 iterations), whereas for the noninformative dropout model it is .86 (confirming the inconclusive nature of inference on the parameter δ_3 under this model). Such flexibility in hypothesis testing is one feature of Bayesian sampling estimation via MCMC.

As a postscript, we can say that the current ‘state of play’ on simulation-based estimation leaves much

scope for development. We can see that MCMC has brought a revolution to the implementation of Bayesian inference and this is likely to continue as computing speeds continue to improve. Our analysis has demonstrated the ease of determining posterior distributions in complex applications, and of developing significance tests that allow for skewness and other non-Normality. These, among other benefits of MCMC techniques, are discussed in a number of accessible books, for example [3, 28].

Note

1. Suppose a chain is defined on a space S . A chain is irreducible if for any pair of states $(s_i, s_j) \in S$ there is a nonzero probability that the chain can move from s_i to s_j in a finite number of steps. A state is positive recurrent if the number of steps the chain needs to revisit the state has a finite mean. If all its states are positive recurrent, then the chain itself is positive recurrent. A state has period k if it can only be revisited after a number of steps that is a multiple of k . Otherwise, the state is aperiodic. If all its states are aperiodic, then the chain itself is aperiodic. Positive recurrence and aperiodicity together constitute ergodicity.

References

- [1] Aitkin, M. & Aitkin, I. (2005). Bayesian inference for factor scores, *Contemporary Advances in Psychometrics*, Lawrence Erlbaum (In press).
- [2] Albert, J. & Chib, S. (1996). Computation in Bayesian econometrics: an introduction to Markov Chain Monte Carlo, in *Advances in Econometrics*, Vol. 11A, T. Fomby & R. Hill, eds, JAI Press, pp. 3–24.
- [3] Albert, J. & Johnson, V. (1999). *Ordinal Data Modeling*, Springer-Verlag, New York.
- [4] Arminger, G. & Muthén, B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm, *Psychometrika* **63**, 271–300.
- [5] Bernardo, J. & Smith, A. (1994). *Bayesian Theory*, Wiley, Chichester.
- [6] Best, N., Spiegelhalter, D., Thomas, A. & Brayne, C. (1996). Bayesian analysis of realistically complex models, *Journal of the Royal Statistical Society, Series A* **159**, 323–342.
- [7] Bockenholt, U., Kamakura, W. & Wedel, M. (2003). The structure of self-reported emotional experiences: a mixed-effects poisson factor model, *British Journal of Mathematical and Statistical Psychology* **56**, 215–229.
- [8] Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**, 1313–1321.
- [9] Coleman, M., Cook, S., Matthyse, S., Barnard, J., Lo, Y., Levy, D., Rubin, D. & Holzman, P. (2002). Spatial and object working memory impairments in schizophrenia patients: a Bayesian item-response theory analysis, *Journal of Abnormal Psychology* **111**, 425–435.
- [10] Congdon, P. (2001). Predicting adverse infant health outcomes using routine screening variables: modelling the impact of interdependent risk factors, *Journal Applied Statistics* **28**, 183–197.
- [11] Congdon, P. (2003). *Applied Bayesian Models*, John Wiley.
- [12] Congdon, P. (2005a). Bayesian predictive model comparison via parallel sampling, *Computational Statistics and Data Analysis* **48**(4), 735–753.
- [13] Congdon, P. (2005b). *Bayesian Model Choice Based on Monte Carlo Estimates of Posterior Model Probabilities*, Computational Statistics and Data Analysis, (In press).
- [14] Cowles, M. & Carlin, B. (1996). Markov Chain Monte-Carlo convergence diagnostics: a comparative study, *Journal of the American Statistical Association* **91**, 883–904.
- [15] Daniels, M. (1999). A prior for the variance in hierarchical models, *Canadian Journal of Statistics* **27**, 567–578.
- [16] Department of Health (2002). *Learning from Bristol: The Department of Health's Response to the Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary 1984–1995*, (Available from <http://www.doh.gov.uk/bristolinquryresponse/index.htm>).
- [17] Diggle, P. (1998). Dealing with missing values in longitudinal studies, in *Statistical Analysis of Medical Data: New Developments*, B. Everitt & G. Dunn, eds, Hodder Arnold, London.
- [18] Geisser, S. & Eddy, W. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **74**, 153–160.
- [19] Gelfand, A. & Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society, Series B* **56**, 501–514.
- [20] Gelfand, A. & Sahu, S. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models, *Journal of the American Statistical Association* **94**, 247–253.
- [21] Gelfand, A. & Smith, A. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.
- [22] Gelfand, A., Sahu, S. & Carlin, B. (1995). Efficient parameterizations for normal linear mixed models, *Biometrika* **82**, 479–488.
- [23] Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457–511.
- [24] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

- [25] George, E., Makov, U. & Smith, A. (1993). Conjugate likelihood distributions, *Scandinavian Journal of Statistics* **20**, 147–156.
- [26] Gilks, W. (1996). Full conditional distributions, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman and Hall, London, pp. 75–88.
- [27] Gilks, W. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics* **41**, 337–348.
- [28] Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). Introducing Markov chain Monte Carlo, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman and Hall, London, pp. 1–20.
- [29] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.
- [30] Hoeting, J., Madigan, D., Raftery, A. & Volinsky, C. (1999). Bayesian model averaging: a tutorial, *Statistical Science* **14**, 382–401.
- [31] Hoijtink, H., Klugkist, I. (2003). *Using the Bayesian Approach in Psychological Research*, Manuscript, Department of Methodology and Statistics, Utrecht University, The Netherlands.
- [32] Ibrahim, J., Chen, M. & Lipsitz, S. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable, *Biometrika* **88**, 551–564.
- [33] Joseph, L., Gyorkos, T. & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard, *American Journal of Epidemiology* **141**, 263–272.
- [34] Kass, R., Tierney, L., Kadane, J. (1988). Asymptotics in Bayesian computation, in *Bayesian Statistics 3*, J. Bernardo, DeGroot, M., Lindley, D., Smith, A., eds, Oxford University Press, 261–278.
- [35] Laud, P. & Ibrahim, J. (1995). Predictive model selection, *Journal of the Royal Statistical Society (Series B)* **57**, 247–262.
- [36] Lee, S., Press, S. (1998). Robustness of Bayesian factor analysis estimates, *Communications in Statistics – Theory and Methods* **27**, 1871–1893.
- [37] Lee, S. & Song, X. (2004). Bayesian model selection for mixtures of structural equation models with an unknown number of components, *British Journal of Mathematical and Statistical Psychology* **56**, 145–165.
- [38] Lewis, C. (1983). Bayesian inference for latent abilities, in *On Educational Testing*, S.B. Anderson, S. Helmick & J., eds, San Francisco, Jossey-Bass, pp. 224–251.
- [39] Little, R. & Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd edition, John Wiley, New York.
- [40] Marshall, C., Best, N., Bottle, A. & Aylin, P. (2004). Statistical issues in the prospective monitoring of health outcomes across multiple units, *Journal of the Royal Statistical Society, Series A* **167**, 541–559.
- [41] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.
- [42] Newton, D. & Raftery, A. (1994). Approximate Bayesian inference by the weighted bootstrap, *Journal of the Royal Statistical Society, Series B* **56**, 3–48.
- [43] Press, S., Shigemasa, K. (1999). A note on choosing the number of factors, *Communications in Statistics – Theory and Methods* **28**, 1653–1670.
- [44] Roberts, C. (1996). Markov chain concepts related to sampling algorithms, in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson & D. Spiegelhalter, eds, Chapman and Hall, London, pp. 45–59.
- [45] Rowe, D. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*, CRC Press, Boca Raton.
- [46] Scheines, R., Hoijtink, H. & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models, *Psychometrika* **64**, 37–52.
- [47] Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society (Series B)* **64**, 583–639.
- [48] Spiegelhalter, D., Abrams, K. & Myles, J. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley.
- [49] Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.

(See also **Markov Chain Monte Carlo Item Response Theory Estimation**)

PETER CONGDON

Markov Chain Monte Carlo Item Response Theory Estimation

LISA A. KELLER

Volume 3, pp. 1143–1148

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Markov Chain Monte Carlo Item Response Theory Estimation

Introduction

Item response theory (IRT) (*see* **Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data**) postulates the probability of a correct response to an item, given a person's underlying ability level, θ . There are many types of item response models, all of which are designed to model different testing situations. However, all of these models contain parameters that characterize the items, and parameters that characterize the examinee. For these models to be useful, the item and person parameters need to be estimated from item response data.

Traditionally, joint **maximum likelihood** (JML) and marginal maximum likelihood (MML) procedures have been used to estimate item parameters (*see* **Maximum Likelihood Estimation**). MML has been favored over JML as the JML estimates are not consistent [8]. The MML estimates proved to be very robust, and have performed well for the traditional models. However, the need for new models is growing, as testing practices become more sophisticated. As model complexity increases, so does the estimation of the model parameters. Therefore, a need for different estimation methods is necessary if model development is to continue, since the MML procedure is not feasible for estimating all models. Fortunately, with the advancements in technology, different estimation methods have become possible. As a result, several new models have been developed to meet the growing needs of measurement. Among the more popular models are the testlet response model [2, 17, 18] and the hierarchical rater model [12, 15]. Both of these models have employed Markov Chain Monte Carlo (MCMC) methods for estimating item parameters (*see* **Markov Chain Monte Carlo and Bayesian Statistics**).

MCMC procedures are becoming more popular among psychometricians primarily due to its flexibility. As noted above, new problems that were previously too difficult, or even impossible, to solve with the traditional methods can be solved. In concept, the

idea of MCMC estimation is to simulate a sample from the posterior distribution of the parameters, and to use the sample to obtain point estimates of the parameters. Since some (most) posterior distributions are difficult to sample from directly, modifications can be made to the process to allow for sampling from a more convenient distribution (e.g., normal), yet selecting the values from that distribution that could realistically have come from the posterior distribution of interest. Any distribution may be used provided that the range of that distribution is chosen appropriately [4]. The drawback to MCMC techniques is the time it often takes to provide a solution. The amount of time it takes depends on many factors, including the choice of the distribution, and some analyses can take weeks of computer time! However, as computing power continues to grow, so will the ease of implementing MCMC methods. As such, it is a technique that will likely gain even more popularity than it enjoys today.

Initial research into MCMC for estimating IRT parameters began with the traditional models, for which MML estimates were available. First, dichotomous models were studied [1, 10, 11, 13], and then this methodology was extended to polytomous models [9, 14, 19]. By doing so, comparisons between estimates obtained with MML and MCMC could be compared, and the MCMC procedures appear to produce estimates similar to those found using MML. As these models are relatively simple, MCMC for IRT parameters will be discussed using these models. The general principles can be applied to estimating parameters of more complex models.

Markov Chain Monte Carlo

Before describing the details of Markov Chain Monte Carlo estimation, a brief description of Markov chains and the introduction of some terminology are necessary.

Markov Chains

A **Markov chain** can be thought of as a system that evolves through various stages, and, at any given stage, exists in a certain state. A Markov chain is a system where the state that the system exists at any stage is determined solely based on probabilities. Further, if a Markov chain is in a particular state

2 Markov Chain Monte Carlo IRT Estimation

at a given stage, the probability that in the next stage it will be at a given state depends only on the present state, and not on past states. That is, at stage k , the probability that it exists in a given state in stage $k+1$ depends only on the state of the chain at stage k , and not stages $1 \dots k-1$. The probabilities of going to the various states are termed the *transition probabilities* and are assumed known. Given that there are typically multiple states that the Markov chain can inhabit at a given stage, the *transition matrix* provides the transition probabilities of going from one state to any other state. That is, the p_{ij} element of the matrix is the probability of going from state i to state j . A distinguishing feature of a Markov chain is that these probabilities do not change over time.

Notation and Terminology

Let $s_i^{(m)}$ denote the probability that the Markov chain is in state i after m transitions. Supposing that there are n different states, there are n such probabilities. These probabilities can be organized into an $n \times 1$ vector, which will be referred to as the *state vector* of the Markov chain. After each transition, a new state vector is formed. As the chain must be in *some* state at any given stage, the sum of the elements of each state vector must be one.

As a Markov chain progresses, the state vectors may tend to converge to a vector of probabilities, say S . Therefore, for a ‘suitably large’ number of transitions, it is reasonable to assume $S_m = S$. This vector S is referred to as the *steady state vector*, or *stationary vector*, for the Markov chain. That is, the vector S provides the long-term probability that the chain will be in each of the various states. In order for the Markov chain to converge to a unique steady-state vector, some regularity conditions must be met: The transition matrix, P , must be a stochastic matrix, that is, the sum of any row is one, and for some integer m , P^m has every entry in the matrix positive.

Markov Chain Monte Carlo Estimation

The basic idea behind MCMC estimation is to create a Markov chain consisting of states $M_0, M_1, M_2, \dots$, where $M_k = (\theta^k, \beta^k)$, and θ, β are the unknown IRT parameters. The chain is created so that the stationary distribution, $\pi(\theta, \beta)$, to which the chain

converges, is the posterior distribution, $p(\theta, \beta|X)$, where X is the matrix of item response data. States (observations) from the Markov chain are simulated and inferences about the parameters are made based on these observations.

The behavior of the Markov chain is determined by its transition kernel $t[(\theta^0, \beta^0), (\theta^1, \beta^1)] = P[M_{k+1} = (\theta^1, \beta^1) | M_k = (\theta^0, \beta^0)]$, which is the probability of moving to a new state, (θ^1, β^1) , given that the current state is (θ^0, β^0) [13]. The stationary distribution $\pi(\theta, \beta)$ satisfies

$$\int_{\theta, \beta} t[(\theta^0, \beta^0), (\theta^1, \beta^1)] \pi(\theta^0, \beta^0) d(\theta^0, \beta^0) = \pi(\theta^1, \beta^1) [13] \quad (1)$$

If we can define the transition kernel such that the stationary distribution, $\pi(\theta, \beta)$, is equal to the posterior distribution, $p(\theta, \beta|X)$, then, after a suitably large number of transitions, m , the remaining observations of the Markov chain behave as if from the posterior distribution. The first m observations are discarded, and are often referred to as the ‘burn-in’ [13]. The observations from the stationary distribution can then be considered a sample from the posterior distribution, and estimates of parameters can be obtained using sample statistics. Given this simple outline of Markov chain techniques, methods for creating the appropriate Markov chain are needed. There are several methods for generating the Markov chains. The most widely used include Gibbs sampling and the Metropolis–Hastings (M–H) algorithm within Gibbs. Each of these methods will be discussed next.

Gibbs Sampling. It was shown by Geman and Geman [6] that the transition kernel

$$t_G[(\theta^0, \beta^0), (\theta^1, \beta^1)] = p(\theta^1 | \beta^0, X) p(\beta^1 | \theta^1, X) \quad (2)$$

has stationary distribution $\pi(\theta, \beta) = p(\theta, \beta|X)$. A Markov chain with a transition kernel constructed in this manner is called a *Gibbs sampler*; the factors $p(\theta^1 | \beta^0, X)$, and $p(\beta^1 | \theta^1, X)$ are the complete conditional distributions of the model. Observations from the Gibbs sampler (θ^k, β^k) are simulated by repeated sampling from the complete conditional distributions. Therefore, to go from $(\theta^{k-1}, \beta^{k-1})$ to (θ^k, β^k) , two transition steps are required:

1. Draw $\theta^k \sim p(\theta|X, \beta^{k-1})$
2. Draw $\beta^k \sim p(\beta|X, \theta^k)$

Hence, the Gibbs sampler employs the standard IRT technique: Estimate one set of parameters holding the other as fixed, and known.

It can be noted that both $p(\theta|X, \beta)$ and $p(\beta|X, \theta)$ are proportional to the joint distribution $p(X, \theta, \beta) = p(X|\theta, \beta)p(\theta, \beta)$:

$$p(\theta|X, \beta) = \frac{p(X|\theta, \beta)p(\theta, \beta)}{\int p(X|\theta, \beta)p(\theta, \beta) d\theta}$$

and

$$p(\beta|X, \theta) = \frac{p(X|\theta, \beta)p(\theta, \beta)}{\int p(X|\theta, \beta)p(\theta, \beta) d\beta} \quad (3)$$

Assuming independence of θ, β , we can write this as $p(\theta|X, \beta) \propto p(X|\theta, \beta)p(\theta)$ and $p(\beta|X, \theta) \propto p(X|\theta, \beta)p(\beta)$. Setting up a Gibbs sampler requires computing the normalizing constants, $\int p(X|\theta, \beta)p(\theta, \beta)d\theta$, $\int p(X|\theta, \beta)p(\theta, \beta)d\beta$, which can be computationally complex. However, there are techniques available in MCMC to circumvent these calculations. One popular choice is to use the Metropolis–Hastings algorithm within Gibbs sampling.

Metropolis–Hastings within Gibbs. As discussed above, both Gibbs samplings yield Markov chains having the desired stationary distribution, $p(\theta, \beta|X)$. However, since it is not trivial to sample from any general distribution, the Gibbs sampler is modified by using the Metropolis–Hastings algorithm, which allows the user to specify a proposal distribution, which is easier to sample from than the complete conditional distributions, and to employ a rejection sampling technique that discards draws from the proposal distribution, which are unlikely to also be draws from the desired conditional distribution. By doing so, a chain with the same stationary distribution can be produced [16]. The Gibbs sampler is modified by choosing proposal distributions for each transition step instead of directly sampling from the complete conditional distributions. The M–H within Gibbs sampling uses different proposal distributions, $g_\theta(\theta^0, \theta^1)$ and $g_\beta(\beta^0, \beta^1)$ for each transition step:

1. Attempt to draw $\theta^k \sim p(\theta|\beta^{k-1}, X)$:
 - (a) Draw $\theta^* \sim g_\theta(\theta^{k-1}, \theta)$.
 - (b) Accept $\theta^k = \theta^*$ with probability

$$\alpha(\theta^{k-1}, \theta^*) = \min \left\{ \frac{p(X|\theta^*, \beta^{k-1})p(\theta^*, \beta^{k-1}) \times g_\theta(\theta^*, \theta^{k-1})}{p(X|\theta^{k-1}, \beta^{k-1})p(\theta^{k-1}, \beta^{k-1}) \times g_\theta(\theta^{k-1}, \theta^*)}, 1 \right\} \quad (4)$$

otherwise set $\theta^k = \theta^{k-1}$

2. Attempt to draw $\beta^k \sim p(\beta|\theta^k, X)$:
 - (a) Draw $\beta^* \sim g_\beta(\beta^{k-1}, \beta)$.
 - (b) Accept $\beta^k = \beta^*$ with probability

$$\alpha(\beta^{k-1}, \beta^*) = \min \left\{ \frac{p(X|\theta^k, \beta^*)p(\theta^k, \beta^*) \times g_\beta(\beta^*, \beta^{k-1})}{p(X|\theta^k, \beta^{k-1})p(\theta^k, \beta^{k-1}) \times g_\beta(\beta^{k-1}, \beta^*)}, 1 \right\} \quad (5)$$

otherwise set $\beta^k = \beta^{k-1}$

The resulting Markov chain has the desired stationary distribution $\pi(\theta, \beta) = p(\theta, \beta|X)$. It should be noted that if either g_θ or g_β is symmetric, then it cancels out in the probability of acceptance ratio. While it may be desirable to use a symmetric distribution to eliminate the proposal distribution, the chain will converge to the stationary distribution more quickly if asymmetric jumping rules are used [4].

Example

Patz and Junker [13] used the following proposal distributions for estimating parameters of the two-parameter logistic model:

$$\begin{aligned} g_B(\log(\beta_{1j}^{k-1}), \log(\beta_{1j})) &= N(\beta_{1j}^{k-1}, 0.3) \\ g_B(\beta_{2j}^{k-1}, \beta_{1j}) &= N(\beta_{2j}^{k-1}, 1.1) \end{aligned} \quad (6)$$

where β_{1j} is the a -parameter for item j , and β_{2j} is the b -parameter for item j .

The above algorithm provides a means to estimate both item parameters and person parameters jointly. However, in traditional estimation, MML has been favored over JML due to the consistency of the marginal estimates that is lacking in the joint

4 Markov Chain Monte Carlo IRT Estimation

estimates. Yao, Patz, and Hanson [20] developed an M–H within Gibbs algorithm that provides marginal estimates of the item parameters. Once the item parameter estimates are obtained, person parameters can be estimated using ML.

The marginal distribution of the parameters is given by:

$$P(\beta|X) \propto P(\beta) \int P(X|\theta, \beta) P(\theta) d\theta$$

This can be written as:

$$P(\beta) \prod_{i=1}^N \int_{\theta} \prod_{j=1}^J P(X_{ij}|\theta_i, \beta_j) P(\theta) d\theta$$

which can be approximated using quadrature points, resulting in:

$$P(\beta|X) \cong P(\beta) \prod_{i=1}^N h(\beta, X_i) \quad (7)$$

where

$$h(\beta, X_i) = \sum_{l=1}^n \prod_{j=1}^J P_{ij}(X_{ij}|\theta^l, \beta_j) A^l$$

where $\theta^1, \dots, \theta^n$ are quadrature points for the examinee population distribution, and $A^1, \dots, A^n$ are the weights at these points [20].

The algorithm used to draw samples from this marginal distribution of the item parameters is given below. Assuming that there are J items to be calibrated, the index j will refer to the item, whereas the index i will refer to one of the N examinees.

- (a) Draw $\beta_j^* \sim g_m(\beta_j|\beta_j^{m-1})$ where $g_m(\beta_j|\beta_j^{m-1})$ is a transition kernel from β_j to β_j^{m-1} .
- (b) For each item, J , calculate the acceptance probability of the parameters:

$$\alpha_j^*(\beta^{k-1}, \beta^*) = \min \left\{ \frac{P(\beta_j^*|\beta_{<j}^m, \beta_{>j}^{m-1}, X) \times g_m(\beta_j^{m-1}|\beta_j^*)}{P(\beta_j^{m-1}|\beta_{<j}^m, \beta_{>j}^{m-1}, X) \times g_m(\beta_j^*|\beta_j^{m-1})}, 1 \right\} \quad (8)$$

where

$$\begin{aligned} \beta_{<j}^m &= (\beta_i^m, i = 1, \dots, j-1) \\ \beta_{>j}^m &= (\beta_i^m, i = j+1, \dots, J) \end{aligned}$$

- (c) Accept each $\beta_j^m = \beta_j^*$ with probability α_j^* , otherwise let $\beta_j^m = \beta_j^{m-1}$

$$\alpha_j^*(\beta^{k-1}, \beta^*) = \min \left\{ \frac{P(\beta_j^*|\beta_{<j}^m, \beta_{>j}^{m-1}, X) \times g_m(\beta_j^{m-1}|\beta_j^*)}{P(\beta_j^{m-1}|\beta_{<j}^m, \beta_{>j}^{m-1}, X) \times g_m(\beta_j^*|\beta_j^{m-1})}, 1 \right\} \quad (9)$$

where

$$\begin{aligned} \beta_{<j}^m &= (\beta_i^m, i = 1, \dots, j-1) \\ \beta_{>j}^m &= (\beta_i^m, i = j+1, \dots, J) \end{aligned}$$

Example

There are many examples of using MCMC techniques to estimate parameters in IRT in the literature, as noted above. The example of Yao, Patz, and Hanson [20], where the marginal distribution was used, will be used here to illustrate the types of proposal distributions used for estimating parameters of the three-parameter logistic model. Since the 3PL is the model used, there are three-item parameters to be estimated for each item: a_j , b_j , and c_j . These parameters will be represented as elements of a parameter vector, β , and as such, the notation will be $a_j = \beta_{1j}$, $b_j = \beta_{2j}$, $c_j = \beta_{3j}$.

Proposal Distributions. Example proposal distributions for the three parameters are:

$$\begin{aligned} g_m(\beta_{1j}|\beta_{1j}^{m-1}) &\sim N(\beta_{1j}^{m-1}, 0.17^2) \\ g_m(\log(\beta_{2j})|\log(\beta_{2j}^{m-1})) &\sim N(\beta_{2j}^{m-1}, 0.17^2) \\ g_m(\beta_{3j}|\beta_{3j}^{m-1}) &= \begin{cases} 0.10 & \text{if } \beta_{3j} \in (\beta_{3j}^{m-1} - 0.05, \beta_{3j}^{m-1} + 0.05) \\ 0.00 & \text{otherwise} \end{cases} \end{aligned} \quad (10)$$

Convergence of the Markov Chain

The parameter estimates resulting from a Markov chain cannot be trusted unless there is evidence that the chain has converged to the posterior distribution. Cowles and Carlin [3] provide a description of several methods used to assess convergence, as well as software to implement the techniques. Another

approach is to create several different chains by choosing different starting values. If all the chains result in the same parameter estimates, evidence is provided that they have converged to the same distribution, and it must be the same distribution [5]. A common tool used to assess convergence is to use Geweke's [7] convergence diagnostic. The diagnostic is based on the equality of means of different parts of the Markov chains. Specifically, comparing the means from the first and last part of the chain, the two samples should have the same mean, if the draws came from the stationary distribution. A z-test is used to test the equality of means, where the asymptotic standard error of the difference is used as the standard error in the test statistic. Details are provided in [7].

References

- [1] Baker, F. (1998). An investigation of parameter recovery characteristics of a Gibbs sampling procedure, *Applied Psychological Measurement* **22**(2), 153–169.
- [2] Bradlow, E.T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets, *Psychometrika* **64**(2), 153–168.
- [3] Cowles, K. & Carlin, B. (1995). *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review (University of Minnesota Technical Report)*, University of Minnesota, Minneapolis.
- [4] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, New York.
- [5] Gelman, A. & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**, 457–472.
- [6] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- [7] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments, in *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds, Clarendon Press, Oxford.
- [8] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff Publishing, Boston.
- [9] Hoijtink, H. & Molenaar, I.W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks, *Psychometrika* **67**, 171–189.
- [10] Keller, L.A. (2002). A comparisons of MML and MCMC estimation for the two-parameter logistic model, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, New Orleans.
- [11] Kim, S.-H. & Cohen, A.S. (1999). Accuracy of parameter estimation in Gibbs sampling under the two-parameter logistic model, in *Paper presented at the meeting of the American Educational Research Association*, Quebec.
- [12] Patz, R.J. (1996). *Markov Chain Monte Carlo Methods for Item Response Theory Models with Applications for the National Assessment of Educational Progress*, Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.
- [13] Patz, R.J. & Junker, B.W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics* **24**(2), 146–178.
- [14] Patz, R.J. & Junker, B.W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics* **24**(4), 342–366.
- [15] Patz, R.J., Junker, B.W., Johnson, M.S. & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational data, *Journal of Educational and Behavioral Statistics* **27**(4), 341–384.
- [16] Tierney, L. (1994). Exploring posterior distributions with Markov chains, *Annals of Statistics* **22**, 1701–1762.
- [17] Wainer, H., Bradlow, E.T. & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in adaptive testing, in *Computerized Adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer, Boston, pp. 245–270.
- [18] Wang, X., Bradlow, E.T. & Wainer, H. (2002). A general Bayesian model for testlets: theory and applications, *Applied Psychological Measurement* **26**, 109–128.
- [19] Wollack, J.A., Bolt, D.M., Cohen, A.S., Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: a comparison of marginal likelihood estimation and Markov Chain Monte Carlo estimation, *Applied Psychological Measurement*, **26**(3), 339–352.
- [20] Yao, L., Patz, R.J. & Hanson, B.A. (2002). More efficient Markov Chain Monte Carlo estimation in IRT using marginal posteriors, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, New Orleans.

Further Reading

- Chib, S. & Greenberg, E. (1995). Understanding the Metropolis-Hasting algorithm, *The American Statistician* **49**, 327–335.
- Gelman, A., Roberts, G.O. & Gilks, W.R. (1996). Efficient metropolis jumping rules, in *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith, eds, Oxford University Press, New York, pp. 599–608.

LISA A. KELLER

Markov, Andrei Andreevich

ADRIAN BANKS

Volume 3, pp. 1148–1149

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Markov, Andrei Andreevich

Born: June 14, 1856, in Ryazan, Russia.

Died: July 20, 1922, in Petrograd (now St Petersburg), Russia.

Markov was a Russian mathematician who developed the theory of sequences of dependent random variables in order to extend the weak **law of large numbers** and **central limit theorem**. In the process, he introduced the concept of a chain of dependent variables, now known as a **Markov chains**. It is this idea that has been applied and developed in a number of areas of the behavioral sciences.

Markov's family moved to St Petersburg when he was a boy. At school, he did not excel in most subjects, with the exception of mathematics in which he showed precocious talent, developing what he believed to be a novel method for the solution of linear differential equations (it transpired that it was not new). On graduation, he entered the Faculty of Mechanics and Mathematics at St Petersburg University where he would stay for the remainder of his career. He defended his Ph.D. thesis in 1884 'On certain applications of the algebraic continuous fractions'. Markov was politically very active, for example, he opposed honorary membership of the Academy of Sciences for members of the royal family as he felt they had not earned the honor. His research was partially fueled by a similarly vehement dispute with Nekrasov who claimed that independence is required for the law of large numbers. Markov's study of what became known as Markov chains ultimately disproved Nekrasov's ideas. In 1913, Markov included in his book the first application of a Markov

chain. He studied the sequence of 20 000 letters in A. S. Pushkin's poem 'Eugeny Onegin' and established the probability of transitions between the vowels and consonants, which form a Markov chain. He died in 1922 as a result of sepsis that had developed after one of several operations that he underwent in the course of his life and that which stemmed from a congenital knee deformity and later complications.

A Markov chain consists of multiple states of a process and the probabilities of moving from one state to another over time, known as transition probabilities. A first-order Markov chain has the property that the subsequent state of the process is dependent on the current state but not on the preceding states. Higher-order Markov chains are dependent on previous states. The chief advantage that this method holds is the ability to analyze sequences of data. Markov chains have been used to model a wide range of behavioral data and cognitive processes, especially learning. Hidden Markov models are an extension in which the observable states are the product of unobservable states, which form a Markov chain. More recently, Markov chains have been used to generate sequences of data for use in Monte Carlo studies.

Further Reading

- Basharin, G.P., Langville, A.N. & Naumov, V.A. (2004). The life and work of A.A. Markov, *Linear Algebra and its Applications* **386**, 3–26.
- Markov, A.A. (1971). *Extension of the limit theorems of probability theory to a sum of variables connected in a chain*, Reprinted in Howard R., ed., *Dynamic Probabilistic Systems*, Vol. 1, Markov Chains, Wiley, New York.
- Wickens, T.D. (1989). *Models for Behavior: Stochastic Processes in Psychology*, W.H. Freeman, San Francisco.

ADRIAN BANKS

Markov Chains

DONALD LAMING

Volume 3, pp. 1149–1152

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Markov Chains

Suppose you are playing blackjack. You are first dealt a '6' and then a '9', to give you a miserable total of 15. What are you going to do? If you choose to 'hit' (draw an additional card), your total will rise to

Total after drawing one additional card

p		12	13	14	15	16	17	18	19	20	21	>21
r	11	1/13	1/13	1/13	1/13	1/13	1/13	1/13	1/13	1/13	4/13	
e	12	0	1/13	1/13	1/13	1/13	1/13	1/13	1/13	1/13	1/13	4/13
s	13	0	0	1/13	1/13	1/13	1/13	1/13	1/13	1/13	1/13	5/13
e	14	0	0	0	1/13	1/13	1/13	1/13	1/13	1/13	1/13	6/13
n	15	0	0	0	0	1/13	1/13	1/13	1/13	1/13	1/13	7/13
t	16	0	0	0	0	0	1/13	1/13	1/13	1/13	1/13	8/13
T	17	0	0	0	0	0	0	1/13	1/13	1/13	1/13	9/13
o	18	0	0	0	0	0	0	0	1/13	1/13	1/13	10/13
t	19	0	0	0	0	0	0	0	0	1/13	1/13	11/13
a	20	0	0	0	0	0	0	0	0	0	1/13	12/13
l												

(1)

16, 17, 18, 19, 20 or 21, each with a probability of 1/13 (ignoring the complications of sampling without replacement), and with probability 7/13 you will 'go bust'. Those probabilities are independent of whether you received the '6' before the '9', or received the '9' first. They are independent of whether your 15 was made up of a '6' and a '9', or a '5' and a '10', or a '7' and an '8'. Your probabilities of future success depend solely on your present total, and owe nothing to how that total was reached. *That* is the *Markov* property.

The matrix above sets out the probabilities of different outcomes following a 'hit' with different totals (again ignoring the complications of sampling without replacement, and I emphasize that this is only a part of the probabilities associated with blackjack, but it will suffice for illustration). The future course of play depends only on the present total and this constitutes a *state* of the Markov chain. Each row of the matrix sums to unity and is therefore a proper multinomial distribution of outcomes, a different distribution for each state. If the total after one 'hit' still falls short of 21, it is possible to 'hit' again, and the transformation defined by the matrix can be

applied more than once. More generally, a matrix of this kind (with all rows summing to unity) can be used to calculate the possible evolution of many systems that similarly inhabit a set of discrete states. The essential condition for such calculation is that the entire future of the system depends only on its present state and is entirely independent of its past history.

Everyday examples of Markov systems tend to evolve in continuous time. The duration of telephone calls tends to an exponential distribution to a surprising degree of accuracy, and exhibits the Markov property. Suppose you telephone a call center to enquire about an insurance claim. 'Please hold; one of our consultants will be with you shortly.' About ten minutes later, still waiting for a 'consultant', you are losing patience. The unpalatable fact is that the time you must now *expect* to wait, after already waiting for ten minutes, is just as long as when you first started. The Markov property means that the future (how long you still have to wait) is entirely independent of the past (how long you have waited already). Light bulbs, on the other hand, are not Markov devices. The probability of failure as you switch the light on increases with the age of the bulb. Likewise, if the probabilities in (1) were recalculated on the basis of sampling *without* replacement, they would be found to depend on how the total had been reached – a '6' and a '9' versus a '5' and a '10' versus a '7' and an '8'.

The chief use of Markov chains in psychology has been in the formulation of models for learning. Bower [1] asked his subjects to learn a list of 10

2 Markov Chains

paired-associates. The stimuli were pairs of consonants and five were paired with the digit '1', five with '2'. Subjects were required to guess, if they did not know the correct response; after each response they were told the correct answer. Bower proposed the following model, comprised of two Markov states (L & U) and an initial distribution of pairs between them.

$$\begin{array}{c} \text{Trial } n \\ \begin{array}{c} L \\ U \end{array} \end{array} \begin{array}{c} \text{Trial } n+1 \\ \begin{array}{cc} L & U \end{array} \end{array} \begin{array}{c} \text{Probability} \\ \text{correct} \end{array} \begin{array}{c} \text{Initial} \\ \text{distribution} \end{array} \begin{array}{c} \left(\begin{array}{cc} 1 & 0 \\ c & 1-c \end{array} \right) \\ \frac{1}{2} \\ 0 \\ 1 \end{array} \quad (2)$$

$$\begin{array}{c} \text{Trial } n \\ \begin{array}{c} P \\ T \\ F \\ N \end{array} \end{array} \begin{array}{c} \text{Trial } n+1 \\ \begin{array}{ccc} P & T & F \end{array} \end{array} \begin{array}{c} \text{Probability} \\ \text{of} \\ \text{avoidance} \end{array} \begin{array}{c} \text{Initial} \\ \text{distribution} \end{array} \begin{array}{c} \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & (1-q) & q \\ e & (1-e)(1-q) & (1-e)q \\ cd & c(1-d)(1-q) & c(1-d)q \end{array} \right) \\ 0 \\ 0 \\ 0 \\ (1-c) \end{array} \begin{array}{c} 1 \\ 1 \\ 0 \\ 0 \end{array} \begin{array}{c} 0 \\ 0 \\ 0 \\ 1 \end{array} \quad (4)$$

This model supposes that on each trial each hitherto unlearned pairing is learnt with probability c and, until a pairing is learnt, the subject is guessing (correct with probability $1/2$). Once a pairing is learnt (state L), subsequent responses are always correct. There is no exit from state L , which is therefore *absorbing*. State U (unlearned), on the other hand, is *transient*. (More precisely, a state is said to be transient if return to that state is less than certain.) Bower's data fitted his model like a glove: but it also needs to be said that any more complicated experiment poses problems not seen here.

This is not the only way that Bower's idea can be formulated. The three-state model

$$\begin{array}{c} \text{Trial } n \\ \begin{array}{c} A \\ C \\ E \end{array} \end{array} \begin{array}{c} \text{Trial } n+1 \\ \begin{array}{cc} C & E \end{array} \end{array} \begin{array}{c} \text{Probability} \\ \text{correct} \end{array} \begin{array}{c} \text{Initial} \\ \text{distribution} \end{array} \begin{array}{c} \left(\begin{array}{cc} 1 & 0 \\ 0 & \frac{1}{2} \end{array} \right) \\ \frac{1}{2}(1-d) \\ \frac{1}{2} \end{array} \begin{array}{c} 1 \\ 1 \\ 0 \end{array} \begin{array}{c} \frac{1}{2}d \\ \frac{1}{2}(1-d) \\ \frac{1}{2} \end{array} \quad (3)$$

with $d = 2c/(1+c)$ is equivalent to the previous two-state model (2), in the sense that the probability of any and every particular set of data is the same

whichever model is used for the calculation [4, pp. 312–314]. The data for each paired-associate in Bower's experiment consists of a sequence of guesses followed by a criterion sequence of correct responses. The errors all occur in State E , correct guesses prior to the last error in State C , and all the responses following learning (plus, possibly some correct guesses immediately prior to learning), in State A . These three states are all *identifiable*, in the sense that (provided there is a sufficiently long terminal sequence of correct responses to allow the inference that the pair has been learned) it can be inferred uniquely from the sequence of responses which state the system occupied at any given trial in the sequence. Models (2) and (3) are *equivalent*.

A more elaborate model of this kind was proposed by Theios and Brelsford [6] to describe avoidance learning by rats. The rats start off naïve (State N) and are shocked. At that point, they learn how to escape with probability c and exit State N . They also learn, for sure, the connection between the warning signal (90 dB noise) and the ensuing shock, but that connection is formed only temporarily with probability $(1-d)$ (i.e., exit to States T or F) and may be forgotten with probability q (State F , rather than T) before the next trial. A trial in State F (meaning of warning signal forgotten) means that the rat will be shocked, whence the connection of the warning signal to the following shock may be acquired permanently with probability e . Here again all the states are identifiable from the observed sequence of responses. The accuracy of the model was well demonstrated in a series of experimental manipulations by Brelsford [2].

A rather different use of a Markov chain is illustrated by **Shannon's** [5] approximations to English text. English text consists of strings of letters and spaces (ignoring the punctuation) conforming to various high-level sequential constraints. Shannon approximated those constraints to varying degrees

with a Markov chain. A *zeroth* order approximation was produced by selecting succeeding characters independently and at random, each with probability $1/27$. A first-order approximation consisted of a similar sequence of letters, but now selected in proportion to their frequencies of occurrence in English text. A second-order approximation was constructed by matching successive pairs of letters to their natural frequency of occurrence. In this approximation, not only did the overall letter frequencies match English text, but the probabilities of selection depended also on the letter preceding. A third order approximation

IN NO IST LAT WHEY CRATICT FROURE BIRS
GROCID PONDEOME OF DEMONSTURES OF
THE REPTAGIN IS REGOACTIONA OF CRE.

meant that each sequence of three successive letters was matched in frequency to its occurrence in English text. The corresponding Markov chain now has 27^2 distinct states. There was, however, no need to construct the transition matrix (a hideous task); it was sufficient to take the last two letters of the approximation and open a book at random, reading on until those same two letters were discovered in correct sequence. The approximation continued with whatever letter followed the two in the book.

The 27×27 matrix for constructing a second-order approximation illustrates some further properties of Markov chains. Since all 27 characters will recur, sooner or later, in English text, all states of the chain (the letters A to Z and the space) can be reached from all other states. The chain is said to be *irreducible*. In comparison, in example (4), return to State N is not possible after it has been left and thereafter the chain can be reduced to just three states. If a Markov chain is irreducible, it follows that all states are certain to recur in due course. A *periodic* state is one that can be entered only every n steps, after which the process moves on to the other states. States that are not subject to such a restriction are *aperiodic*, and if return is also certain, they are called *ergodic*. Since the 27×27 matrix in question is irreducible, all its states are ergodic. It follows that, however the approximation to English text starts off, after a sufficient length has been generated, the frequency of different states (letters and spaces) will tend to a stationary distribution, which, in this case, will correspond to the long-run frequency of the characters in English text.

To sum-up, the matrix (5) exemplifies several important features of Markov chains. The asterisks denote any arbitrary entry between 0 and 1; the row sums, of course, are all unity.

$$\begin{array}{c}
 S_1 \quad S_2 \quad S_3 \quad S_4 \quad S_5 \quad S_6 \quad S_7 \quad S_8 \quad S_9 \\
 \begin{array}{c}
 S_1 \\
 S_2 \\
 S_3 \\
 S_4 \\
 S_5 \\
 S_6 \\
 S_7 \\
 S_8 \\
 S_9
 \end{array}
 \begin{pmatrix}
 * & * & * & * & * & * & * & * & * \\
 * & * & * & * & * & * & * & * & * \\
 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 0 & 0 & & 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\
 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\
 0 & 0 & 0 & 0 & 0 & 0 & * & * & *
 \end{pmatrix}
 \end{array}
 \quad (5)$$

1. It is possible to access all other states from S_1 and S_2 , but return (except from S_1 to S_2 and vice-versa) is not possible. S_1 and S_2 are *transient*.
2. Once S_3 is entered, there is no exit. S_3 is *absorbing*.
3. S_4 , S_5 , and S_6 exit only to each other. They form a closed (absorbing) subset; the subchain comprised of these three states is irreducible. Moreover, once this set is entered, S_4 exits only to S_5 , S_5 only to S_6 , and S_6 only to S_4 , so that thereafter each of these states is visited every three steps. These states are *periodic*, with period 3.
4. S_7 , S_8 , and S_9 likewise exit only to each other, and also form a closed irreducible subchain. But each of these three states can exit to each of the other two and the subchain is *aperiodic*. Since return to each of these three states is certain, once the subset has been entered, the subchain is *ergodic*.

For further properties of Markov chains the reader should consult [3, Chs. XV & XVI].

References

- [1] Bower, G.H. (1961). Application of a model to paired-associate learning, *Psychometrika* **26**, 255–280.
- [2] Breisford Jr, J.W. (1967). Experimental manipulation of state occupancy in a Markov model for avoidance conditioning, *Journal of Mathematical Psychology* **4**, 21–47.

4 Markov Chains

- [3] Feller, W. (1957). *An Introduction to Probability Theory and its Applications*, Vol. 1, 2nd Edition, Wiley, New York.
- [4] Greeno, J.G. & Steiner, T.E. (1964). Markovian processes with identifiable states: general considerations and application to all-or-none learning, *Psychometrika* **29**, 309–333.
- [5] Shannon, C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal* **27**, 379–423.
- [6] Theios, J. & Brelsford Jr, J.W. (1966). Theoretical interpretations of a Markov model for avoidance conditioning, *Journal of Mathematical Psychology* **3**, 140–162.

(See also **Markov, Andrei Andreevich**)

DONALD LAMING

Martingales

DONALD LAMING

Volume 3, pp. 1152–1154

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Martingales

A simple example of a martingale would be one's cumulative winnings in an 'absolutely fair' game. For example, Edith and William play a regular game of bridge, once a week, against Norman and Sarah for £10 per 100 points. This is a very satisfactory arrangement because these two pairs are exactly matched in their card skills. Coupled with a truly random assignment of cards to the four players, this match ensures that their *expected* winnings on each hand are zero; the game is *absolutely fair*. This does not, however, preclude one pair or the other building up a useful profit over the course of time.

Norman and Sarah's cumulative winnings are

$$S_n = X_1 + X_2 + \cdots + X_n, \quad (1)$$

where X_i is the amount they won on the i th hand. This equation describes an additive process, and there are two elementary conditions on the X_i that permit simple and powerful inferences about the behavior of the sums S_n .

1. If the X_n are independent and identically distributed, (1) describes a **random walk** *qv*. In a random walk, successive steps, X_n , are independent; this means that the distributions of successive S_n can be calculated by repeated convolution of the X_n .
2. If, instead, the expectations of successive X_n are zero, the sequence $\{S_n\}$ is a *martingale*. I emphasize that X_n need not be independent of the preceding steps $\{X_i, i = 1, \dots, n-1\}$, merely that its expectation is always zero, irrespective of the values that those preceding steps happen to have had. Formally, $E\{X_{n+1}|X_1, X_2, \dots, X_n\} = 0$. This has the consequence that $E\{S_{n+1}|X_1, X_2, \dots, X_n\} = S_n$ (which is the usual definition of a martingale).

Continuing the preceding example, the most important rewards in bridge come from making 'game', that is, 100 points or more from contracts bid and made. If Norman and Sarah already have a part-score toward 'game' from a previous hand, their bidding strategy will be different in consequence. Once a partnership has made 'game', most of their penalties (for failing to make a contract) are doubled; so bidding strategy changes again. None of

these relationships abrogate the martingale property; the concept is broad.

The martingale property means that the variances of the sums S_n are monotonically increasing, whatever the relationship of X_{n+1} to the preceding X_n . The variance of S_{n+1} is

$$\begin{aligned} E\{(S_n + X_{n+1})^2\} &= E\{S_n^2 + 2S_n X_{n+1} + X_{n+1}^2\} \\ &= E\{S_n^2 + X_{n+1}^2\}, \end{aligned} \quad (2)$$

because $E\{X_{n+1}\} = 0$. From this, it follows that if the variances of the sums S_n are bounded, then S_n tends to a limiting distribution [2, p. 236]. Martingales are important, not so much as models of behavior in their own right, but as a concept that simplifies the analysis of more complicated models, as the following example illustrates.

Yellott [9] reported a long binary prediction experiment in which observers had to guess which of two lights, A and B, would light up on the next trial. A fundamental question was whether observers modified their patterns of guessing on every trial or only when they found they had guessed wrongly. The first possibility may be represented by a linear model. Let p_n be the probability of choosing light A on trial n . If Light A does indeed light up, put

$$p_{n+1} = \alpha p_n + (1 - \alpha); \quad (3a)$$

otherwise

$$p_{n+1} = \alpha p_n. \quad (3b)$$

Following each trial p_n in (3) is replaced by a weighted average of p_n and the light actually observed (counting Light A as 1 and Light B as 0). If Light A comes on with probability a , fixed over successive trials, then p_n tends asymptotically to a – that is, 'probability matching', a result that is commonly observed in binary prediction. This finding provides important motivation for the linear model, except that the same result can be derived from other models [1, esp. pp. 179–181]. So, for the last 50 trials of his experiment, Yellott switched on whichever light the subject selected (i.e., noncontingent success reinforcement with probability 1). If observers modified their patterns of guessing only when they had guessed wrong, noncontingent success reinforcement should effect no change at all in the pattern of guessing. But for the linear model on

2 Martingales

trial $n + 1$,

$$\begin{aligned} E\{p_{n+1}\} &= p_n[\alpha p_n + (1 - \alpha)] + (1 - p_n)[\alpha p_n] \\ &= p_n, \end{aligned} \quad (4)$$

so the sequence $\{p_n\}$ is a martingale. The variance of p_n therefore increases monotonically and, for any one sequence of trials, p_n should tend either to 0 or to 1. Yellott found no systematic changes of that kind.

However, a formally similar experiment by Howarth and Bulmer [4] yielded a different result. The observers in this experiment were asked to report a faint flash of light that had been adjusted to permit about 50% detections. The intensity of the light was constant over successive trials, but there was no knowledge of results. So the response, detection or failure to detect, fulfilled a role analogous to noncontingent success reinforcement. Successive responses were statistically related in a manner consistent with the linear model (3). Moreover, the authors reported ‘The experiment was stopped on two occasions after the probability of seeing had dropped almost to zero’ [4, p. 164]; that is, for two observers, p_n decreased to 0. There was a similar decrease to zero in the proportion of positive diagnoses by a consultant pathologist engaged in screening cervical smears [3]. Her frame of judgment shifted to the point that she was passing as ‘OK’ virtually every smear presented to her for expert examination.

The tasks of predicting the onset of one of two lights or of detecting faint flashes of light do not make sense unless both alternative responses are appropriate on different trials. This generates a prior expectation that some particular proportion of each kind of response will be required. A prior expectation can be incorporated into the linear model by replacing the reinforcement in (3) (Light A = 1, Light B = 0) with the response actually uttered (there was no reinforcement in Howarth and Bulmer’s experiment) and then replacing the response (1 for a detection, 0 for a miss) with a weighted average of the prior expectation and the response. This gives

$$p_{n+1} = \alpha p_n + (1 - \alpha)[b\pi_\infty + (1 - b)]; \quad (5a)$$

$$p_{n+1} = \alpha p_n + (1 - \alpha)b\pi_\infty. \quad (5b)$$

Here, π_∞ is the prior expectation of what proportion of trials will present a flash of light for detection,

and the expression in square brackets applies the linear model of (3) to the effect of that prior expectation. The probability of detection tends asymptotically to π_∞ . But, notwithstanding that the mean no longer satisfies the martingale condition, the variance still increases to a limiting value, greater than would be obtained from independent binomial trials [5, p. 464]. This increased variability shows up in the proportions of positive smears reported by different pathologists [6], in the proportions of decisions (prescription, laboratory investigation, referral to consultant, follow-up appointment) by family doctors [8, p. 158] and in the precision of the ‘quantal’ experiment. The ‘quantal’ experiment was a procedure introduced by Stevens and Volkman [7] for measuring detection thresholds. Increments of a fixed magnitude were presented repeatedly for 25 successive trials. The detection probability for that magnitude was estimated by the proportion of those increments reported by the observer. The precision of 25 successive observations in the ‘quantal’ experiment is equivalent to about five independent binomial trials [5] (*see Catalogue of Probability Density Functions*).

References

- [1] Atkinson, R.C. & Estes, W.K. (1963). Stimulus sampling theory, in *Handbook of Mathematical Psychology*, Vol. 2, R.D. Luce, R.R. Bush & E. Galanter, eds, Wiley, New York, pp. 121–268.
- [2] Feller, W. (1957). *An Introduction to Probability Theory and its Applications*, Vol. 2, Wiley, New York.
- [3] Fitzpatrick, J., Utting, J., Kenyon, E. & Burston, J. (1987). *Internal Review into the Laboratory at the Women’s Hospital, Liverpool*, Liverpool Health Authority, 8th September.
- [4] Howarth, C.I. & Bulmer, M.G. (1956). Non-random sequences in visual threshold experiments, *Quarterly Journal of Experimental Psychology* **8**, 163–171.
- [5] Laming, D. (1974). The sequential structure of the quantal experiment, *Journal of Mathematical Psychology* **11**, 453–472.
- [6] Laming, D. (1995). Screening cervical smears, *British Journal of Psychology* **86**, 507–516.
- [7] Stevens, S.S. & Volkman, J. (1940). The quantum of sensory discrimination, *Science* **92**, 583–585.
- [8] Wilkin, D., Hallam, L., Leavey, R. & Metcalfe D. (1987). *Anatomy of Urban General Practice, 1987*, Tavistock Publications, London.
- [9] Yellott Jr, J.I. (1969). Probability learning with noncontingent success, *Journal of Mathematical Psychology* **6**, 541–575.

DONALD LAMING

Matching

VANCE W. BERGER, LI LIU AND CHAU THACH

Volume 3, pp. 1154–1158

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Matching

Models for Matched Pairs

Evaluation is generally inherently comparative. For example, a treatment is generally neither ‘good’ nor ‘bad’ in absolute terms, but rather better than or worse than another treatment at bringing about the desired result or outcome. The studies that allow for such comparative evaluations must then also be comparative. It would not do, for example, to study only one treatment and then declare it to be the winner. To isolate the effects of the treatments under study, avoid **confounding** treatment effects with unit effects, and possibly increase the **power** of the study to detect treatment differences, the units under study must be carefully matched across treatment groups. To appreciate the need to control such confounding, consider comparing the survival times of patients in a cancer clinic to those of patients in a sports injury clinic. It might be expected that the patients with sports injuries would live longer, but this is not generally accepted as proof that treatment (for either sports injuries or for cancer) is better at a sports injury clinic than it is at a cancer center. The difference in survival times across the ‘treatment’ groups is attributable not to the treatments themselves (cancer center vs. sports injury clinic) but rather to underlying differences in the patients who choose one or the other.

While the aforementioned example is an obvious example of confounding, and one that is not very likely to lead to confusion, there are less obvious instances of confounding that can lead to false conclusions. For example, Marcus [6] evaluated a randomized study of a culturally sensitive AIDS education program [13]. At baseline, the treatment group had significantly lower AIDS knowledge scores (39.89 vs. 36.72 on a 52-question test, $p = 0.005$), so an unadjusted comparison would be confounded.

The key to avoiding confounding is in ensuring the comparability of the comparison groups in every way other than the treatments under study. This way, by the process of elimination, any observed difference can be attributed to differences in the effects of the treatments under study. Ideally, the control group for any given patient or set of patients would be the same patient or set of patients, under identical conditions. This is the idea behind **crossover trials**,

in which patients are randomized not to treatment conditions, but rather to sequences of treatment conditions, so that each patient experiences each treatment condition, in some order.

But crossovers are not the ideal solution, because while each patient is exposed to each treatment condition, this exposure is not under identical conditions. Time must necessarily elapse between the sequential exposures, and patients do not remain the same over time, even if left untreated. The irony is that the very nature of the crossover design, in which the patient is treated initially with one treatment in the hope of improving some condition (i.e., changing the patient), interferes with homogeneity over time. Under some conditions, **carryover effects** (the effect of treatment during one period on outcomes measured during a subsequent period when a different treatment is being administered) may be minimal, but, in general, this is a serious concern.

Besides crossover designs, there are other methods to ensure matching. The matching of the comparison groups may be at the group level, at the individual unit level, or sometimes at an intermediate level. For example, if unrestricted randomization is used to create the comparison groups, then the hope is that the groups are comparable with respect to both observed and unobserved covariates, but there is no such hope for any particular subset of the comparison groups. If, however, the randomization is stratified, say by gender, then the hope is that the females in one treatment group will be comparable to the females in each other treatment group, and that the males in one treatment group will be comparable to the males in each other treatment group (*see Stratification*). When **randomization** is impractical or impossible, the matching may be undertaken in manual mode. There are three types of matching – simple matching, symmetrical matching, and split sample [5]. In simple matching, a given set of covariates, such as age, gender, and residential zip code, may be specified, and each unit (subject) may be matched to one (resulting in paired data) or several other subjects on the basis of these covariates. In most situations, simple matching is also done to increase the power of the study to detect differences between the treatment groups due to the reduction in the variability between the treatment groups. Pauling [8] presented a data set of simple matching in his Table 33.2. This table presents the time until cancer patients were determined to be beyond treatment.

2 Matching

There are 11 cancer patients who were treated with ascorbic acid (vitamin C), and ‘for each treated patient, 10 controls were found of the same sex, within five years of the same age, and who had suffered from cancer of the same primary organ and histological tumor type. These 1000 cancer patients (overall there were more than 11 cases, but only these 11 cases are presented in Table 33.2) comprise the control group. The controls received the same treatment as the ascorbic-treated patients except for the ascorbate’ [8]. Of note, Pauling [8] went on to state that ‘It is believed that the ascorbate-treated patients represent a random selection of all of the terminal patients in the hospital, even though no formal randomization process was used’ [8]. It would seem that what is meant here is that there are no systematic differences to distinguish the cases from the controls, and this may be true, but it does not, in any way, equate to proper randomization [1].

The matching, though not constituting randomization, may provide somewhat of a basis for inference, because if the matching were without any hidden bias, then the time to ‘untreatability’ for the case would have the same chance to exceed or to be exceeded by that of any of the controls matched to that case. Under this assumption, one may proceed with an analysis based on the ranking of the case among its controls. For example, Case #28 had a time of 113 days, which is larger than that of five controls and smaller than that of the other five controls. So, Case #28 is quite typical of its matched controls. Of course, any inference would be based on all 11 cases, and not just Case #11.

There is **missing data** for one of the controls for Case #35, so we exclude this case for ease of illustration (a better analysis, albeit a more complicated one, would make use of all cases, including this one), and proceed with the remaining 10 cases, as in Table 1. Cases #37 and #99 had times of 0 days, and each had at least one control with the same time, giving rise to the ties in Table 1. These ties also complicate the analysis somewhat, but for our purposes, it will not matter if they are resolved so that the case had the longer time or the shorter time, as the ultimate conclusion will be that there was no significant difference at the 0.05 level. To illustrate this, we take a conservative approach (actually, it is a liberal approach, as it enhances the differences between the groups, but this is conservative relative to the claim that there is no significant difference) as follows. First, we note that

Table 1 Data from Pauling [8]

Case	Controls with shorter times	Ties	Controls with longer times
28	5	0	5
33	1	0	9
34	6	0	4
36	9	0	1
37	0	2	8
38	2	0	8
84	6	0	4
90	4	0	6
92	2	0	8
99	0	1	9
Sum	35	3	62

there were 62 shorter case times versus only 35 longer case times (this could form the basis for a U-test [10, 11], ignoring the ties), and so the best chance to find a difference would be with a one-sided test to detect shorter case times. We then help this hypothesis along by resolving the ties so that the case times are always shorter. This results in combining the last two columns of Table 1, so that it is 35 versus 65.

As noted, the totals, be it 35 versus 62 or 35 versus 65, are sufficient to enable us to conduct a U-test (*see Wilcoxon–Mann–Whitney Test*) [10, 11]. However, we propose a different analysis because the U-test treats as interchangeable all comparisons; we would like to make more use of the matched sets. Consider, then, any given rank out of the 11 (one case matched to 10 controls). For example, consider the second largest. Under the null hypothesis, the probability of being at this rank or higher is 2/11, so this can form the basis for a binomial test. Of course, one can conduct a binomial test for any of the rank positions. The data summary that supports such analyses is presented in Table 2. The rank of the case is one more than the number of control times exceeded by the case time, or one more than the second column of Table 1. It is known that artificially small P values result from selection of the test based upon the data at hand, in particular, when dealing with maximized differences among cut-points of an ordered contingency table, or a Lancaster decomposition (*see Kolmogorov–Smirnov Tests*) [2, 9]. Nevertheless, we select the binomial test that yields the most significant (smallest) one-sided P value, which is at rank 7.

We see that 9 of the 10 cases had a rank less than seven. The point probability of this outcome is, by

Table 2 Data from Table 1 put into a form amenable for binomial analysis

Rank	Sets in which the case has this rank	Cumulative
1	2 ^a	2
2	1	3
3	2	5
4	0	5
5	1	6
6	1	7
7	2	9
8	0	9
9	0	9
10	1	10
11	0	10
Total	10	

^aCases #37 and #99 have the lowest rank by the tie-breaking convention.

the binomial formula, $(10)(7/11)^9(4/11) = 0.0622$. The P value is this plus the null probability of all other more extreme outcomes. In this case, the only more extreme outcome is to have all 10 cases have rank less than seven, and this has null probability $(7/11)^{10} = 0.0109$, so the one-sided binomial P value is the sum of these two null probabilities, or 0.0731, which, with every advantage given toward a finding of statistical significance, still came up short. This is not surprising with so small a sample size. The method, of course, can be applied with larger sample sizes as well.

The second type of matching is symmetrical matching, where the effect of two different treatments are tested on opposite sides of the body. Kohnen et al. [4] compared the difference between standard and a modified (zero-compression) Hansatome micro-keratome head in the incidence of epithelial defects. Ninety-three patients (186 eyes) were enrolled in the study. To avoid confounding, each patient's two eyes were matched. In one eye, the flaps were created using the standard Hansatome head and in the other eye, the flaps were created using a modified design (zero-compression head). Epithelial defects occurred in 21 eyes in which the standard head was used and in 2 eyes (2.1%) in which the zero-compression head was used. Two patients who had an epithelial defect using the zero-compression head also had an epithelial defect in the eye in which the standard head was used. McNemar's test is appropriate to analyze such data from matched pairs of subjects

with a dichotomous (yes–no) response. The P value based on McNemar's test [7] is less than 0.001, and suggests that the modified Hansatome head significantly reduced the occurrence of the epithelial defects. To compare matched pairs with continuous response, the paired t Test [12] or Wilcoxon signed-ranks test [3] is proper. Shea et al. [14] studied the microarchitectural bone adaptations of the concave and convex spinal facets in idiopathic scoliosis. Biopsy specimens of facet pairs at matched anatomic levels were obtained from eight patients. The concave and convex facets were analyzed for bone porosity. The mean porosity (and standard deviation) for the concave and convex facets was 16.5% $\pm$ 5.8% and 24.1% $\pm$ 6.2%, respectively. The P value based on the paired t Test is less than 0.03, and suggests that the facets on the convex side were significantly more porous than those on the concave side.

The third type of matching is the split sample design, in which each individual is divided into two parts and one treatment is randomized to one part, while the other part gets the second treatment. For example, a piece of fabric could be cut into two pieces and each piece is tested with one of two detergents. Symmetrical matching and split samples both result in paired data. With these designs, the study usually has more power to detect differences between the treatment groups, compared to unmatched designs, because the individual variation is reduced.

Matching bears some resemblance to randomized trials that stratify by time of entry onto the trial, as cohorts of patients form blocks, and randomization occurs within the block. Often, each block will have the same number of cases (called *treated patients*) and controls (patients treated with the control group), with obvious modification for more than two treatment groups. The similarity in the previous sentence applies both within blocks and across blocks; that is, often any given block has the same number of patients randomized to each treatment condition (1:1 allocation within each block), and the size of each block will be the same (constant block size). To the extent that time trends render patients within a block more comparable to each other than to patients in other blocks, we have a structure similar to that of the Pauling [8] matched study, except that each matched set, or block, may have more than one case, or treated patient.

References

- [1] Berger, V.W. & Bears, J. (2003). When can a clinical trial be called 'randomized'? *Vaccine* **21**, 468–472.
- [2] Berger, V.W. & Ivanova, A. (2002). Adaptive tests for ordered categorical data, *Journal of Modern Applied Statistical Methods* **1**(2), 269–280.
- [3] Box, G., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, New York.
- [4] Kohnen, T., Terzi, E., Mirshahi, A. & Bühren, J. (2004). Intraindividual comparison of epithelial defects during laser in situ keratomileusis using standard and zero-compression Hansatome microkeratome heads, *Journal of Cataract and Refractive Surgery* **30**(1), 123–126.
- [5] Langley, R. (1971). *Practical Statistics Simply Explained*, Dover Publications, New York.
- [6] Marcus, S.M. (2001). A sensitivity analysis for subverting randomization in controlled trials, *Statistics in Medicine* **20**, 545–555.
- [7] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157.
- [8] Pauling, L. (1985). Supplemental ascorbate, vitamin C, in the supportive treatment of cancer, in *Data*, D.F. Andrews & A.M. Herzberg, eds, Springer-Verlag, New York, pp. 203–207.
- [9] Permutt, T. & Berger, V.W. (2000). A new look at rank tests in ordered 2xk contingency tables, *Communications in Statistics – Theory and Methods* **29**(5 and 6), 989–1003.
- [10] Pettitt, A.N. (1985). Mann-Whitney-Wilcoxon statistic, in *The Encyclopedia of Statistical Sciences*, Vol. 5, S. Kotz & N.L. Johnson, eds, Wiley, New York.
- [11] Serfling, R.J. (1988). U-statistics, in *The Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz & N.L. Johnson, eds, Wiley, New York.
- [12] Shea, K.G., Ford, T., Bloebaum, R.D., D'Astous, J. & King, H. (2004). A comparison of the microarchitectural bone adaptations of the concave and convex thoracic spinal facets in idiopathic scoliosis, *The Journal of Bone and Joint Surgery-American Volume* **86-A**(5), 1000–1006.
- [13] Stevenson, H.C. & Davis, G. (1994). Impact of culturally sensitive AIDS video education on the AIDS risk knowledge of African American adolescents, *AIDS Education and Prevention* **6**, 40–52.
- [14] Wayne, W.D. (1978). *Applied Nonparametric Statistics*, Houghton Mifflin Company.

VANCE W. BERGER, LI LIU AND CHAU THACH

Mathematical Psychology

RICHARD A. CHECHILE

Volume 3, pp. 1158–1164

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mathematical Psychology

Mathematical psychology deals with the use of mathematical and computational modeling methods to measure and explain psychological processes. Although probabilistic models dominate the work in this area, mathematical psychology is distinctly different from the statistical analysis of data. While there is a strong interest in mathematical psychology with the measurement of psychological processes, mathematical psychology does not deal with multiple-item tests such as intelligence tests. The study of multi-item tests and questionnaires is a central focus of psychometrics (see **History of Psychometrics; Psychophysical Scaling**). Psychometrics has a close linkage with education and clinical psychology—fields concerned with psychological assessment. In contradistinction to psychometrics, mathematical psychology is concerned with measuring and describing the behavioral patterns demonstrated in experimental research. Thus, mathematical psychology has a close tie to experimental psychology—a linkage that is analogous to the connection between theoretical and experimental physics.

The Development of Mathematical Psychology

One of the earliest examples of experimental psychology is an 1860 book by **Fechner** [10] on psychophysics – the science of relating physical dimensions to perceived psychological dimensions. This treatise is also a pioneering example of mathematical psychology. While Fechner's work was followed by a steady stream of experimental research, there was not a corresponding development for mathematical psychology. Mathematical psychology was virtually nonexistent until after World War II. By that time psychological science had progressed to the point where (a) statistical tools for data analysis were common, (b) rich databases were developed in most areas of psychology that demonstrated regular behavioral patterns, and (c) a wide range of theories had been proposed that were largely verbal descriptions. During World War II, a number of psychologists worked with engineers, physicists, and mathematicians. This collaboration stimulated

the development of a more rigorous theoretical psychology that employed mathematical methods. That time period also saw the rapid development of new branches of applied mathematics such as control theory, cybernetics, information theory, system theory, game theory, and automata theory. These new mathematical developments would all prove useful in the modeling of psychological processes.

By the 1950s, a number of papers with mathematical models were being published regularly in the journal, *Psychological Review*. Books and monographs about mathematical psychology also began appearing [22]. Regular summer workshops on mathematical behavioral sciences were being conducted at Stanford University. Early in the 1960s, there was an explosion of edited books with high quality papers on mathematical psychology. Eventually, textbooks about mathematical psychology were published [7, 14, 15, 27, 33]. These texts helped to introduce the subject of mathematical psychology into the graduate training programs at a number of universities. In 1964, the *Journal of Mathematical Psychology* was launched with an editorial board of Richard Atkinson, **Robert Bush**, **Clyde Coombs**, William Estes, Duncan Luce, William McGill, George Miller, and Patrick Suppes. The first conference on mathematical psychology was held in 1968. The next eight meetings were informally organized by the interested parties, but in 1976, an official professional society was created – the Society for Mathematical Psychology. The officers of this organization (a) decide the location and support the arrangements for an annual conference, (b) select the editor and recommend policies for the *Journal of Mathematical Psychology*, (c) recognize new researchers in the field by offering a 'young investigator' award, and (d) recognize especially important new work by awarding an 'outstanding paper' prize.

Outside North America, the subfield of mathematical psychology developed as well. In 1965, the *British Journal of Statistical Psychology* changed its name to the *British Journal of Mathematical and Statistical Psychology* in order to include mathematical psychology papers along with the more traditional test theory and psychometric papers. In 1967, the journal *Mathematical Social Sciences* was created with an editorial board containing many European social scientists as well as North Americans. In 1971, a group of European mathematical psychologists held a conference

in Paris. The group has come to be called the *European Mathematical Psychology Group*. This society also has an annual meeting. A number of edited books have emerged from the papers presented at this conference series. In 1989, Australian and Asian mathematical psychologists created their own series of research meetings (i.e., the Australasian Mathematical Psychology Conference).

Research in Mathematical Psychology

Mathematical psychology has come to be a broad topic reflecting the utilization of mathematical models in any area of psychology. Because of this breadth, it has not been possible to capture the content of mathematical psychology with any single volume. Consequently, mathematical psychology research is perhaps best explained with selected examples. The following examples were chosen to reflect a range of modeling techniques and research areas.

Signal Detection Theory

Signal detection theory (SDT) is a well-known model that has become a standard tool used by experimental psychologist. Excellent sources of information about this approach can be found in [11, 34]. The initial problem that motivated the development of SDT was the measurement of the strength of a sensory stimulus and the control of decision biases that affect detection processes. On each trial of a typical experiment, the subject is presented with either a faint stimulus or no stimulus. If the subject adopts a lenient standard for claiming that the stimulus is present, then the subject is likely to be correct when the stimulus is present. However, such a lenient standard for decision making means that the subject is also likely to have many false alarms, that is, claiming a stimulus is present when it is absent. Conversely, if a subject adopts a strict standard for a detection decision, then there will be few false alarms, but at the cost of failing on many occasions to detect the stimulus. Although SDT developed in the context of sensory psychophysics, the model has also been extensively used in memory and information processing research because the procedure of providing either stimulus-present or stimulus-absent trials is common in many research areas.

In what might be called the *standard form of SDT*, it is assumed that there are two underlying normal distributions on a psychological strength axis. One distribution represents the case when the stimulus is absent, and the other distribution is for the case when the stimulus is present. The stimulus-present distribution is shifted to the right on the strength axis relative to the stimulus-absence distribution. Given data from at least two experimental conditions where different decision criteria are used, it is possible to estimate the parameter d' , which is defined as the separation between the two normal distributions divided by the standard deviation of the stimulus-absent distribution. It is also possible to estimate the ratio of the standard deviations of the two distributions.

Multinomial Process Tree Models

The data collected in a signal detection experiment are usually categorical, that is, the hit rates and false alarm rates under conditions that have different decision criteria. In fact, many other experimental tasks also consist of measuring proportions in various response categories. A number of mathematical psychologists prefer to model these proportions directly in terms of underlying psychological processes. The categorical information is referred to as multinomial data. In this area of research, the mathematical psychologist generates a detailed probability description as to how underlying psychological processes result in the various proportions for the observed multinomial data. Most researchers develop a probability tree representation where the branches of the tree correspond to the probabilities of the latent psychological processes. The leaves of the tree are the observed categories in an experiment. These models have thus come to be called *multinomial processing tree (MPT)* models.

An extensive review of many of MPT models in psychological research can be found in [3]. Given experimental data, that is, observed proportions for the various categories for responding, the latent psychological parameters of interest can be estimated. General methods for estimating parameters for this class of models are specified in [5, 12]. The goal of MPT modeling is to obtain measures of the latent psychological parameters. For example, a method was developed to use MPT models to obtain separate

measures of memory storage and retrieval for a task that has an initial free recall test that is followed by a series of forced-choice recognition tests [5, 6]. MPT models are tightly linked to a specific experimental task. The mathematical psychologist often must invent an experimental task that provides a means of measuring the underlying psychological processes of interest.

Information Process and Reaction Time Models

A number of mathematical psychologists model processes in the areas of psychophysics, information processing (*see* **Information Theory**), and cognition; see [9, 16, 19, 30]. These researchers have typically been interested in explaining properties of dependent measures from experiments, such as the participants' response time, percentage correct, or the trade-off of time with task accuracy. For example, many experiments require the person to make a choice response, that is, the individual must decide if a stimulus is the same or different from some standard. The statistical properties of the time distribution of 'same' response are not equivalent to the distribution of the 'different' response. Mathematical psychologists have been interested in accounting for the entire response-time distribution for each type of response. One successful model for response time is the relative judgment, **random walk model**, [17]. Random walk models are represented by a state variable (a real-value number that reflects the current state of evidence accumulation). The state variable is contained between two fixed boundaries. The state variable for the next time increment is the same as the prior value plus or minus a random step amount. Eventually, the random walk of the state variable terminates when the variable reaches either one of the two boundaries. The random walk model results in a host of predictions for the same-different response times.

Axiomatic Measurement Theories and Functional Equations

Research in this area does not deal with specific stochastic models of psychological processes but rather focuses on the necessary and sufficient conditions for a general type of measurement scale or a form of a functional relationship among variables.

Typically, a minimal set of principles or axioms is considered, and the consequences of the assumed axioms are derived. If the resulting theory is not supported empirically, then at least one of the axioms must be in error. Furthermore, this approach provides experimental psychologists with an opportunity to test the theory by assessing the validity of essential axioms.

A well-known example of the axiomatic approach is the choice axiom [18]. To illustrate the axiom, let us consider an example of establishing a preference ordering of a set of candidates. For each pair of candidates, there will be choice probability P_{ij} that denotes the probability that candidate i is preferred over candidate j . The choice axiom states that the probability that any candidate is preferred is statistically independent of the removal of one of the candidates. In essence, this principle is the same as a long-held principle in voting theory called the *independence* of irrelevant alternatives, that is, the relative strength between two candidates should not change if another candidate is disqualified. Luce shows that this axiom implies that there must exist numbers v_i associated with each alternative, and that the choice probability is $v_i/(v_i + v_j)$. The result of this formal analysis also establishes that we cannot characterize the choice probabilities in terms of a ratio of underlying strengths if we can show empirically that for some pair of candidates the relative preference is altered by the removal of some other candidate.

A similar approach to research in mathematical psychology can be found in the utilization of functional equation theory. D'Alembert, an eighteenth century mathematician, initiated this branch of mathematics, and it has proved to be important for psychological theorizing. Perhaps the most famous illustration of functional equation analysis from mathematics is the Cauchy equation. It can be proved that the only continuous function that satisfies the constraint that $f(x + y) = f(x) + f(y)$ where x and y are nonnegative is $f(x) = cx$, where c is a constant. Functional equation analysis can be used to determine the form of a mathematical relationship between critical variables. Theorems are developed that demonstrate that only one mathematical function satisfies a given set of requirements without the necessity of curve fitting or statistical analysis. To illustrate this approach, let us consider the analysis provided by Luce [20] in the

area of psychophysics. Consider two stimuli of intensities I and I' , respectively, and let $f(I)$ and $f(I')$ be the corresponding psychological perception of the stimuli. It can be established that if $I/I' = c$, where c is a constant, and if $f(I)/f(I') = g(c)$, then $f(I)$ must be a power function, that is, $f(I) = AI^\beta$, where A and β are positive constants that are independent of I .

Judgment and Decision-making Models

Economics and mathematics have developed rational models for decision making. However, there is considerable evidence from experimental psychology that people often do not behave according to these 'normative' rational models [13]. For example, let us consider the psychological worth of a gamble. Normative economic theory posits that a risky commodity (e.g., a gamble) is worth the expected utility of the gamble. Utility theory was first formulated by the mathematician Daniel **Bernoulli** in 1738 as a solution to a gambling paradox. Prior to that time, the worth of a gamble was the expected value for the gamble, for example, given a gamble that has a 0.7 probability for a gain of \$10 and a probability of 0.3 for a loss of \$10, the expected value is $0.7(\$10) + (0.3)(-\$10) = \$4$. However, Bernoulli considered a complex gamble based on a bet doubling system and showed that the gamble had infinite expected value. Yet, individuals did not perceive that gamble as having infinite value. To resolve this discrepancy, Bernoulli replaced the monetary values with subjective worth numbers for the monetary outcomes – these numbers were called *utility values*. Thus, for the above example, the subjective utility is $0.7[U(\$10)] + (0.3)[U(-\$10)]$, where the u function is nonlinear, monotonic increasing function of dollar value. General axioms for expected **utility theory** have been developed [31], and that framework has become a central theoretical perspective in economics. However, experimental psychologists provided numerous demonstrations that this theory is not an accurate descriptive theory. For example, the Allais paradox illustrates a problem with expected utility theory [7, 32]. Subjects greatly prefer a certain \$2400 to a gamble where there is a 0.33 probability for \$2500, a 0.66 probability for \$2400, and a 0.01 chance for \$0. Thus, from expected utility theory, it follows that $U(\$2400) > 0.33[U(\$2500)] +$

$0.66[U(\$2400)] + 0.01[U(\$0)]$, which is equivalent to $0.34[U(\$2400)] > 0.33[U(\$2500)] + 0.01[U(\$0)]$. However, these same subjects prefer a lottery with a 0.33 chance for \$2500 to a lottery that has a 0.34 chance for \$2400. This second preference implies according to expected utility theory that $0.33[U(\$2500)] + 0.67[U(\$0)] > 0.34[U(\$2400)] + 0.66[U(\$0)]$, which is equivalent to $0.34[U(\$2400)] < 0.33[U(\$2500)] + 0.01[U(\$0)]$. Notice that preferences for the gambles are inconsistent, that is, from the first preference ordering we deduced that $0.34[U(\$2400)] > 0.33[U(\$2500)] + 0.01[U(\$0)]$, but from the second preference ordering we obtained that $0.34[U(\$2400)] < 0.33[U(\$2500)] + 0.01[U(\$0)]$. Clearly, there is a violation of a prediction from expected utility theory. In an effort to achieve more realistic theories for the perception of the utility of gambles, mathematical psychologists and theoretical economists formulated alternative models. For example, Luce [21] extensively explores alternative utility models in an effort to find a model that more accurately describes the behavior of individuals.

Models of Memory

The modeling of memory has been an active area of research; see [26]. One of the best-known memory models is the Atkinson and Shiffrin multiple-store model [2]. This model deals with both short-term and long-term memory with a system of three memory stores that can be used in a wide variety of ways. The memory stores are a very short-term sensory register, a short-term store, and a permanent memory store. This model stimulated considerable experimental and theoretical research.

The Atkinson and Shiffrin model did not carefully distinguish between recall and recognition behavior measures. These measures are quite different. In subsequent memory research, there has been more attention given to the similarities and differences in recall and recognition measures. In particular, a considerable amount of interest has centered on a class of recognition memory models that has come to be called 'global matching models'. Models in this class differ widely as to the basic representation of information. One type of representation takes a position like that of the Estes array model [8]. For this model, each memory is considered as a separate N -dimensional vector of attributes. A recognition

memory probe activates all of the existing items in the entire memory system by an amount that is dependent on the similarity between the probe and the memory representation. In the Estes array model for classification and recognition, the similarity function between two N -dimensional memory vectors is defined by the expression ts^{N-k} , where t represents the similarity value for the match of k attributes and s is a value between 0 and 1 that is associated with the reduced activation caused by a feature mismatch. The recognition decision is a function of the total similarity produced by the recognition memory probe with all the items in the memory system.

Another model in the global matching class is the Murdock TODAM model [24, 25]. This model uses a distributed representation of information in the memory system. It is assumed that memory is composed of a single vector. The TODAM model does not have separate vectors for various items in memory. Recognition is based on a vector function that depends only on the recognition probe and current state of the memory vector. As more items are added to memory, previous item information may be lost. For TODAM and many other memory models, **Monte Carlo** (i.e., random sampling) methods are used to generate the model predictions for various experimental conditions. In general, these models have a number of parameters – often more parameters than are identifiable in a single condition of an experiment. However, given values for the model parameters, it would be possible to account for the data obtained in that condition as well as many other experimental conditions. In fact, a successful model is usually applied across a wide range of experiments and conditions without major changes in the values for the parameters.

Neural Network Models

Many of the mathematical models of psychological processes fall into the category of **neural networks**. There is also considerable interest in neural network models outside of psychology. For example, physicists modeling materials known as spin glasses have found that neural network models, which were originally developed to explain learning, can also describe the behavior of these substances [4].

In general, there is a distinction between real and artificial neural networks. Real neural networks deal

with brain structures such as the hippocampus or the visual cortex [28]. For real neural networks, the researcher focuses on the functioning of a set of cells or a brain structure as opposed to the behavior of the whole animal. However, artificial neural networks are algorithms mainly for learning and memory, and are more likely to be linked to observable behavior. Artificial neural networks are distributed computing systems that pass information throughout a network (see [1, 23, 29]). As the network ‘experiences’ new input patterns and receives feedback from its ‘behavior’, there are changes in the properties of links and nodes in the network, which in turn affect the behavior of the network. Artificial neural networks have been valuable as models for learning and pattern recognition. There are many possible arrangements for artificial neural network.

Summary

Mathematical psychology is the branch of theoretical psychology that uses mathematics to measure and describe the regularities that are obtained from psychological experiments. Some research in this area (i.e., signal detection theory and multinomial processing tree models) is focused on extracting measures of latent psychological processes from a host of observable measures. Other models in mathematical psychology are focused on the description of the changes in psychological processes across experimental conditions. Examples of this approach would include models of information processing, reaction time, learning, memory, and decision making. The mathematical tools used in mathematical psychology are diverse and reflect the wide range of psychological problems under investigation.

References

- [1] Anderson, J.A. (1995). *Practical Neural Modeling*, MIT Press, Cambridge.
- [2] Atkinson, R.C. & Shiffrin, R. (1968). Human memory: a proposed system and its control processes, in *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, K.W. Spence & J.T. Spence, eds, Academic Press, New York, pp. 89–195.
- [3] Batchelder, W.H. & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling, *Psychonomic Bulletin & Review* 6, 57–86.

- [4] Bovier, A. & Picco, P., eds (1998). *Mathematical Aspects of Spin Glasses and Neural Networks*, Birkhäuser, Boston.
- [5] Chechile, R.A. (1998). A new method for estimating model parameters for multinomial data, *Journal of Mathematical Psychology* **42**, 432–471.
- [6] Chechile, R.A. & Soraci, S.A. (1999). Evidence for a multiple-process account of the generation effect, *Memory* **7**, 483–508.
- [7] Coombs, C.H., Dawes, R.M. & Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction*, Prentice-Hall, Englewood Cliffs.
- [8] Estes, W.K. (1994). *Classification and Cognition*, Oxford University Press, New York.
- [9] Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*, Oxford University Press, Oxford.
- [10] Fechner, G.T. (1860). *Elemente der Psychophysik*, Breitkopf & Härtel, Leipzig.
- [11] Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, Wiley, New York.
- [12] Hu, X. & Phillips, G.A. (1999). GPT.EXE: a powerful tool for the visualization and analysis of general processing tree models, *Behavior Research Methods Instruments & Computers* **31**, 220–234.
- [13] Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
- [14] Laming, D. (1973). *Mathematical Psychology*, Academic Press, London.
- [15] Levine, G. & Burke, C.J. (1972). *Mathematical Model Techniques for Learning Theories*, Academic Press, New York.
- [16] Link, S.W. (1992). *The Wave Theory of Difference and Similarity*, Erlbaum, Hillsdale.
- [17] Link, S.W. & Heath, R.A. (1975). A sequential theory of psychological discrimination, *Psychometrika* **40**, 77–105.
- [18] Luce, R.D. (1959). *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York.
- [19] Luce, R.D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*, Oxford University Press, New York.
- [20] Luce, R.D. (1993). *Sound & Hearing: A Conceptual Introduction*, Erlbaum, Hillsdale.
- [21] Luce, R.D. (2000). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*, Erlbaum, London.
- [22] Luce, R.D., Bush, R.R. & Galanter, E., eds (1963;1965). *Handbook of Mathematical Psychology*, Vol. I, II and III, Wiley, New York.
- [23] McClelland, J.L. & Rumelhart, D.E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, MIT Press, Cambridge.
- [24] Murdock, B.B. (1982). A theory for the storage and retrieval of item and associative information, *Psychological Review* **89**, 609–626.
- [25] Murdock, B.B. (1993). TODEM2: a model for the storage and retrieval of item, associative, and serial-order information, *Psychological Review* **100**, 183–203.
- [26] Raaijmakers, J.G.W. & Shiffrin, R.M. (2002). Models of memory, in *Stevens' Handbook of Experimental Psychology*, Vol. 2, H. Pashler & D. Medin, eds, Wiley, New York, pp. 43–76.
- [27] Restle, F. & Greeno, J.G. (1970). *Introduction to Mathematical Psychology*, Addison-Wesley, Reading.
- [28] Rolls, E.T. & Treves, A. (1998). *Neural Networks and Brain Function*, Oxford University Press, Oxford.
- [29] Rumelhart, D.E. & McClelland, J.L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT Press, Cambridge.
- [30] Townsend, J.T. & Ashby, F.G. (1983). *Stochastic Modeling of Elementary Psychological Processes*, Cambridge University Press, Cambridge.
- [31] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.
- [32] von Winterfeld, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.
- [33] Wickens, T.D. (1982). *Models for Behavior: Stochastic Processes in Psychology*, W. H. Freeman, San Francisco.
- [34] Wickens, T.D. (2002). *Elementary Signal Detection Theory*, Oxford University Press, Oxford.

RICHARD A. CHECHILE

Maximum Likelihood Estimation

CRAIG K. ENDERS

Volume 3, pp. 1164–1170

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Maximum Likelihood Estimation

Many advanced statistical models (e.g., **structural equation models**) rely on maximum likelihood (ML) estimation. In this entry, we will explore the basic principles of ML estimation using a small data set consisting of 10 scores from the Beck Depression Inventory (BDI) and a measure of perceived social support (PSS). ML parameter estimates are desirable because they are both *consistent* (i.e., the estimate approaches the population parameter as sample size increases) and *efficient* (i.e., have the lowest possible variance, or sampling fluctuation), but these characteristics are asymptotic (i.e., true in large samples). Thus, although these data are useful for pedagogical purposes, it would normally be unwise to use ML with such a small N . The data vectors are as follows.

$$\begin{aligned} \text{BDI} &= [5, 33, 17, 21, 13, 17, 5, 13, 17, 13], \\ \text{PSS} &= [17, 13, 13, 5, 17, 17, 19, 15, 3, 11]. \end{aligned} \quad (1)$$

The goal of ML estimation is to identify the population parameters (e.g., a mean, regression coefficient, etc.) most likely to have produced the sample data (see **Catalogue of Probability Density Functions**). This is accomplished by computing a value called the *likelihood* that summarizes the fit of the data to a particular parameter estimate. Likelihood is conceptually similar to probability, although strictly speaking they are not the same. To determine how ‘likely’ the sample data are, we must first make an assumption about the population score distribution. The normal distribution is frequently used for this purpose.

The shape of a normal curve is described by a complicated formula called a *probability density function* (PDF). The PDF describes the relationship between a set of scores (on the horizontal axis) and the relative probability of observing a given score (on the vertical axis). Thus, the height of the normal curve at a given point along the x -axis provides information about the relative frequency of that score in a normally distributed population with a given mean and variance, μ and σ^2 .

The univariate normal PDF is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}, \quad (2)$$

and is composed of two parts: the term in the exponent is the squared standardized distance from the mean, also known as **Mahalanobis distance**, and the term preceding the exponent is a scaling factor that makes the area under the curve equal to one.

The normal PDF plays a key role in computing the likelihood value for a sample. To illustrate, suppose it was known that the BDI had a mean of 20 and standard deviation of 5 in a particular population. The likelihood associated with a BDI score of 21 is computed by substituting $\mu = 20$, $\sigma = 5$, and $x_i = 21$ into (2) as follows.

$$\begin{aligned} L_i &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \\ &= \frac{1}{\sqrt{2(3.14)(5^2)}} e^{-[(21-20)/5]^2/2} = .078. \end{aligned} \quad (3)$$

The resulting value, .078, can be interpreted as the relative probability of $x_i = 21$ in a normally distributed population with $\mu = 20$ and $\sigma = 5$. Graphically, .078 represents the height of this normal distribution at a value of 21, as seen in Figure 1.

Extending this concept, the likelihood for every case can be computed in a similar fashion, the results of which are displayed in Table 1. For comparative purposes, notice that the likelihood associated with a BDI score of 5 is approximately .001. This tells us that the relative probability of $x_i = 5$ is much lower than that of $x_i = 21$ (.001 versus .078, respectively), owing to the fact that the latter score is more likely to have occurred from a normally distributed population with $\mu = 20$. By extension, this also illustrates that the likelihood associated with a given x_i will change as the population parameters, μ and σ , change.

Table 1 Likelihood and log likelihood values for the hypothetical sample

Case	BDI	Likelihood (L_i)	Log L_i
1	5	.001	-7.028
2	33	.003	-5.908
3	17	.067	-2.708
4	21	.078	-2.548
5	13	.030	-3.508
6	17	.067	-2.708
7	5	.001	-7.028
8	13	.030	-3.508
9	17	.067	-2.708
10	13	.030	-3.508

2 Maximum Likelihood Estimation

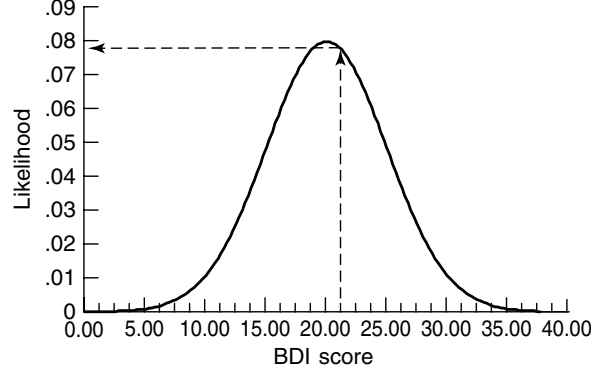


Figure 1 Graph of the univariate normal PDF. The height of the curve at $x_i = 21$ represents the relative probability (i.e., likelihood) of this score in a normally distributed population with $\mu = 20$ and $\sigma = 5$

Having established the likelihood for individual x_i values, the likelihood for the sample is obtained via multiplication. From probability theory, the joint probability for a set of independent events is obtained by multiplying individual probabilities (e.g., the probability of jointly observing two heads from independent coin tosses is $(.50)(.50) = .25$). Strictly speaking, likelihood values are not probabilities; the probability associated with any single score from a continuous PDF is zero. Nevertheless, the likelihood value for the entire sample is defined as the product of individual likelihood values as follows:

$$L = \prod_{i=1}^N \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x_i - \mu)/\sigma]^2/2} \right\}, \quad (4)$$

where Π is the product operator. Carrying out this computation using the individual L_i values from Table 1, the likelihood value for the sample is approximately .000000000000000001327 – a very small number!

Because L becomes exceedingly small as sample size increases, the logarithm of (4) can be used to make the problem more computationally attractive. Recall that one of the rules of logarithms is $\log(ab) = \log(a) + \log(b)$. Applying logarithms to (4) gives the *log likelihood*, which is an additive, rather than multiplicative, model.

$$\log L = \sum_{i=1}^N \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x_i - \mu)/\sigma]^2/2} \right\}. \quad (5)$$

Individual $\log L_i$ values are shown in Table 1, and the log likelihood value for the sample is the sum

of individual $\log L_i$, approximately -41.16 . The log likelihood value summarizes the likelihood that this sample of 10 cases originated from a normally distributed population with parameters $\mu = 20$ and $\sigma = 5$. As can be inferred from the individual L_i values in the table, higher values (i.e., values closer to zero) are indicative of a higher relative probability. As we shall see, the value of the log likelihood can be used to ascertain the ‘fit’ of a sample to a particular set of population parameters.

Thus far, we have worked under a scenario where population parameters were known. More typically, population parameters are unknown quantities that must be estimated from the data. The process of identifying the unknown population quantities involves ‘trying out’ different parameter values and calculating the log likelihood for each. The final *maximum likelihood estimate* (MLE) is the parameter value that produced the highest log likelihood value.

To illustrate, suppose we were to estimate the BDI population mean from the data. For simplicity, assume the population standard deviation is $\sigma = 5$. The sample log likelihood for μ values ranging between 10 and 20 are given in Table 2. Beginning with $\mu = 10$, the log likelihood steadily increases (i.e., improves) until μ reaches a value of 15, after which the log likelihood decreases (i.e., gets worse). Thus, it appears that the MLE of μ is near a value of 15.

The relationship between the population parameters and the log likelihood value, known as the *log likelihood function*, can also be depicted graphically. As seen in Figure 2, the height of the log likelihood

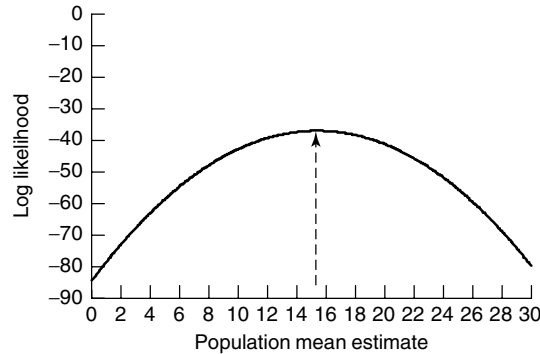


Figure 2 The maximum of the log likelihood function is found at $\mu = 15.4$

Table 2 Log likelihood values for different estimates of the population mean

μ estimate	Log likelihood
10	-42.764
11	-40.804
12	-39.244
13	-38.084
14	-37.324
15	-36.964
16	-37.004
17	-37.444
18	-38.284
19	-39.524
20	-41.164

function reaches a maximum between $\mu = 15$ and 16. More precisely, the function is maximized at $\mu = 15.40$, where the log likelihood takes on a value of -36.9317645 . To confirm this value is indeed the maximum, note that the log likelihood values for $\mu = 15.399$ and 15.401 are -36.9317647 and -36.9617647 , respectively, both of which are smaller (i.e., worse) than the log likelihood produced by $\mu = 15.40$. Thus, the MLE of the BDI population mean is 15.40 , as it is the value that maximizes the log likelihood function. Said another way, $\mu = 15.4$ is the population parameter most likely to have produced this sample of 10 cases.

In practice, the process of ‘trying out’ different parameter values en route to maximizing the log likelihood function is aided by calculus. In calculus, a slope, or rate of change, of a function at a fixed point is known as a *derivative*. To illustrate, tangent lines are displayed at three points on the log likelihood

function in Figure 3. The slope of these lines is the first derivative of the function with respect to μ . Notice that the derivative (i.e., slope) of the function at $\mu = 6$ is positive, while the derivative at $\mu = 24$ is negative. You have probably already surmised that the tangent line has a slope of zero at the value of μ associated with the function’s maximum. Thus, the MLE can be identified via calculus by setting the first derivative to zero and solving for corresponding value of μ on the horizontal axis.

Identifying the point on the log likelihood function where the first derivative equals zero does not ensure that we have located a maximum. For example, imagine a U-shaped log likelihood function where the first derivative is zero at ‘the bottom of the valley’ rather than ‘at the top of the hill’. Fortunately, verifying that the log likelihood function is at its maximum, rather than minimum, can be accomplished by checking the sign of the second derivative. Because second derivatives also play an important role in estimating the variance of an MLE, a brief digression into calculus is warranted.

Suppose we were to compute the first derivative for every point on the log likelihood function. For $\mu < 15.4$, the first derivatives are positive, but decrease in value as μ approaches 15.4 (i.e., the slopes of the tangent lines become increasingly flat close to the maximum). For $\mu > 15.4$, the derivatives become increasingly negative (i.e., more steep) as you move away from the maximum. Now imagine creating a new graph that displays the value of the first derivative (on the vertical axis) for every estimate of μ on the horizontal axis. Such a graph is called a *derivative function*, and second derivatives are defined as the slope of a line tangent to the

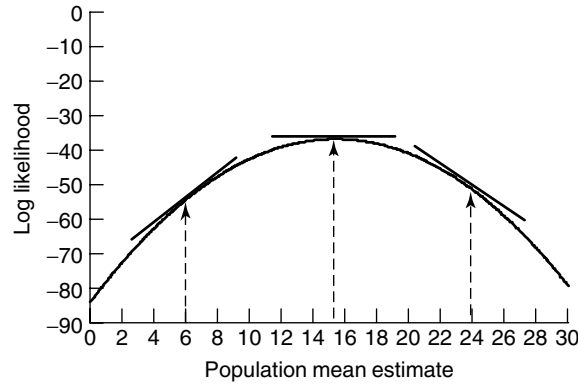


Figure 3 First derivatives are slopes of lines tangent to the log likelihood function. The derivative associated with $\mu = 6$ is positive, while the derivative associated with $\mu = 24$ is negative. The derivative is zero at the function's maximum, $\mu = 15.4$

derivative function (i.e., the derivative of a derivative, or rate of change in the slopes). A log likelihood function that is concave down (e.g., Figure 3) would produce a derivative function that begins in the upper left quadrant of the coordinate system (the derivatives are large positive numbers at $\mu < 15.4$), crosses the horizontal axis at the function's maximum value, and continues with a downward slope into the lower right quadrant of the coordinate system (the derivatives become increasingly negative at $\mu > 15.4$). Further, the second derivative (i.e., the slope of a tangent line) at any point along such a function would be negative. In contrast, the derivative function for a U-shaped log likelihood would stretch from the upper right quadrant to the lower left quadrant, and would produce a positive second derivative. Thus, a negative second derivative verifies that the log likelihood function is at a maximum.

Identifying the MLE is important, but we also want to know how much uncertainty is associated with an estimate – this is accomplished by examining the parameter's standard error. The variance of an MLE, the square root of which is its standard error, is also a function of the second derivatives. In contrast to Figure 2, imagine a log likelihood function that is very flat. The first derivatives, or slopes, would change very little from one value of μ to the next. One way to think about second derivatives is that they quantify the rate of change in the first derivatives. A high rate of change in the first derivatives (e.g., Figure 2) reflects greater certainty about the parameter estimate (and thus a lower standard error), while

derivatives that change very little (e.g., a relatively flat log likelihood function) would produce a larger standard error. When estimating multiple parameters, the matrix of second derivatives, called the **information matrix**, is used to produce a covariance matrix (see **Correlation and Covariance Matrices**) of the parameters (different from a covariance matrix of scores), the diagonal of which reflects the variance of the MLEs. The standard errors are found by taking the square root of the diagonal elements in the parameter covariance matrix.

Having established some basic concepts, let us examine a slightly more complicated scenario involving multiple parameters. To illustrate, consider the regression of BDI scores on PSS. There are now two parameters of interest, a regression intercept, β_0 , and slope, β_1 (see **Multiple Linear Regression**). The log likelihood given in (5) is altered by replacing μ with the conditional mean from the regression equation (i.e., $\mu = \beta_0 + \beta_1 x_{1i}$).

$$\log L = \sum_{i=1}^N \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(y_i - \beta_0 - \beta_1 x_{1i})/\sigma]^2/2} \right\}. \quad (6)$$

Because there are now two unknowns, estimation involves 'trying out' different values for both β_0 and β_1 . Because the log likelihood changes as a function of two parameters, the log likelihood appears as a three-dimensional surface. In Figure 4, the values of β_0 and β_1 are displayed on separate axes, and the vertical axis gives the value of the log likelihood for every combination of β_0 and β_1 . A precise solution

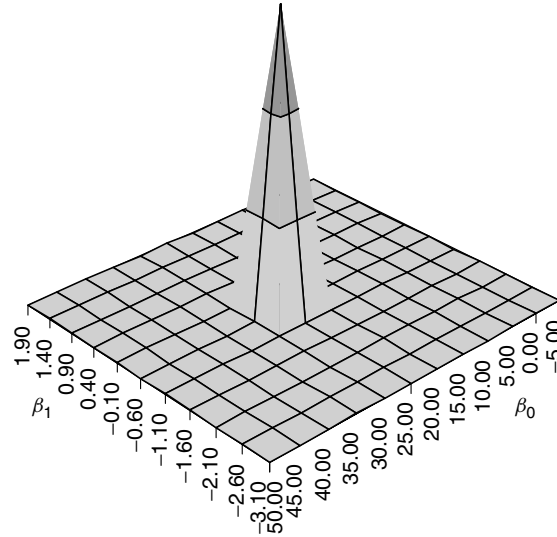


Figure 4 Log likelihood surface for simple regression estimation. ML estimates of β_0 and β_1 are 23.93 and $-.66$, respectively

for β_0 and β_1 relies on the calculus concepts outlined previously, but an approximate maximum can be identified at the intersection of $\beta_0 = 25$ and $\beta_1 = -.60$ (the ML estimates obtained from a model-fitting program are $\beta_0 = 23.93$ and $\beta_1 = -.66$).

Standard errors for the MLEs are obtained as a function of the second derivatives, and can be used to construct single-parameter significance tests of the null hypothesis, $\beta_j = 0$. Because the asymptotic sampling distribution for an MLE is itself a normal distribution, a z test is constructed by dividing the parameter estimate by its *asymptotic standard error* (ASE), the square root of the appropriate diagonal element in the parameter covariance matrix.

$$z = \frac{\hat{\beta} - 0}{ASE(\hat{\beta})}. \quad (7)$$

To illustrate, the ASE for β_1 is .429, and the corresponding z ratio is $-.656/.429 = -1.528$. This z ratio can subsequently be compared to a critical value obtained from the unit normal table. For example, the critical value for a two-tailed test using $\alpha = .05$ is ± 1.96 , so β_1 is not significantly different from zero. The ASEs can also be used to construct confidence intervals around the MLEs in the usual fashion.

A more general strategy for conducting significance tests involves comparing log likelihood values

from two nested models. Models are said to be nested if the parameters of one model (i.e., the ‘reduced’ model) are a subset of the parameters from a second model (i.e., the ‘full’ model). For example, consider a multiple regression analysis with five predictor variables, x_1 through x_5 . After obtaining MLEs of the regression coefficients, suppose we tested a reduced model involving only x_1 and x_2 . If the log likelihood changed very little after removing x_3 through x_5 , we might reasonably conclude that this subset of predictors is doing little to improve the fit of the model – conceptually, this is similar to testing a set of predictors using an F ratio.

More formally, the likelihood ratio test is defined as the difference between $-2 \log L$ values (sometimes called the *deviance*) for the full and reduced model. This difference is asymptotically distributed as a chi-square statistic with degrees of freedom equal to the difference in the number of estimated parameters between the two models. A nonsignificant likelihood ratio test indicates that the fit of the reduced model is not significantly worse than that of the full model. It should be noted that the likelihood ratio test is only appropriate for use with nested models, and could not be used to compare models with different sets of predictors, for example. Likelihood-based indices such as the Bayesian Information Criterion (BIC) can

instead be used for this purpose (*see Bayesian Statistics*).

The examples provided thus far have relied on the univariate normal PDF. However, many common statistical models (e.g., structural equation models) are derived under the multivariate normal PDF, a generalization of the univariate normal PDF to $k \geq 2$ dimensions. The multivariate normal distribution (*see Catalogue of Probability Density Functions*) can only be displayed graphically in three dimensions when there are two variables, and in this case the surface of the bivariate normal distribution looks something like a bell-shaped ‘mound’ (e.g., see [2, p. 152]).

The multivariate normal PDF is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})/2}. \quad (8)$$

In (8), $\mathbf{x}$ and $\boldsymbol{\mu}$ are vectors of scores and means, respectively, and $\boldsymbol{\Sigma}$ is the covariance matrix of the scores. Although (8) is expressed using vectors and matrices, the major components of the PDF remain the same; the term in the exponent is a multivariate extension of Mahalanobis distance that takes into account both the variances and covariances of the scores, and the term preceding the exponent is a scaling factor that makes the volume under the density function equal to one (in the multivariate case, probability values are expressed as volumes under the surface).

Consistent with the previous discussion, estimation proceeds by ‘trying out’ different values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in search for estimates that maximize the log likelihood. Note that in some cases the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may not be the ultimate parameters of interest but may be functions of model parameters (e.g., $\boldsymbol{\Sigma}$ may contain covariances predicted by the parameters from a structural equation model, or $\boldsymbol{\mu}$ may be a function of regression coefficients from a **hierarchical model**). In any case, an individual’s contribution to the likelihood is obtained by substituting her vector of scores, $\mathbf{x}$, into (8) and solving, given the estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. As before, the sample likelihood is the product of (8) over the N cases, and the log likelihood is obtained by summing the logarithm of (8) over the N cases.

The estimation of multiple parameters generally requires the use of iterative numerical optimization techniques to maximize the log likelihood. One such

technique, the estimation maximization (EM) algorithm, is discussed in detail in [11], and these algorithms are discussed in more detail elsewhere [1]. Conceptually, the estimation process is much like climbing to the top of a hill (i.e., the log likelihood surface), similar to that shown in Figure 4. The first step of the iterative process involves the specification of initial values for the parameters of interest. These *starting values* may, in some cases, be provided by the analyst, and in other cases are arbitrary values provided automatically by the model-fitting program. The choice of different starting values is tantamount to starting the climb from different locations on the topography of the log likelihood surface, so a good set of starting values may reduce the number of ‘steps’ required to reach the maximum. In any case, these initial values are substituted into the multivariate PDF to obtain an initial log likelihood value. In subsequent iterations, adjustments to the parameter estimates are chosen such that the log likelihood consistently improves (this is accomplished using derivatives), eventually reaching its maximum. As the solution approaches its maximum, the log likelihood value will change very little between subsequent iterations, and the process is said to have converged if this change falls below some threshold, or *convergence criterion*.

Before closing, it is important to distinguish between full maximum likelihood (FML) and restricted maximum likelihood (RML), both of which are commonly implemented in model-fitting programs. Notice that the log likelihood obtained from the multivariate normal PDF includes a mean vector, $\boldsymbol{\mu}$, and a covariance matrix, $\boldsymbol{\Sigma}$. FML estimates these two sets of parameters simultaneously, but the elements in $\boldsymbol{\Sigma}$ are not adjusted for the uncertainty associated with the estimation of $\boldsymbol{\mu}$ – this is tantamount to computing the variance using N rather than $N - 1$ in the denominator. As such, FML variance estimates will exhibit some degree of negative bias (i.e., they will tend to be too small), particularly in small samples. RML corrects this problem by removing the parameter estimates associated with $\boldsymbol{\mu}$ (e.g., regression coefficients) from the likelihood. Thus, maximizing the RML log likelihood only involves the estimation of parameters associated with $\boldsymbol{\Sigma}$ (in some contexts referred to as **variance components**). Point estimates for $\boldsymbol{\mu}$ are still produced when using RML, but these parameters are

estimated in a separate step using the RML estimates of the variances and covariances.

In practice, parameter estimates obtained from FML and RML tend to be quite similar, perhaps trivially different in many cases. Nevertheless, the distinction between these two methods has important implications for hypothesis testing using the likelihood ratio test. Because parameters associated with μ (e.g., regression coefficients) do not appear in the RML likelihood, the likelihood ratio can only be used to test hypotheses involving variances and covariances. From a substantive standpoint, these tests are often of secondary interest, as our primary hypotheses typically involve means, regression coefficients, etc. Thus, despite the theoretical advantages associated with RML, FML may be preferred in many applied research settings, particularly given that the two approaches tend to produce similar variance estimates.

In closing, the intent of this manuscript was to provide the reader with a brief overview of the basic principles of ML estimation. Suffice to say, we have only scratched the surface in the short space allotted here, and interested readers are encouraged to consult more detailed sources on the topic [1].

References

- [1] Eliason, S.R. (1993). *Maximum Likelihood Estimation: Logic and Practice*, Sage, Newbury Park.
- [2] Johnson, R.A. & Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, 5th Edition, Prentice Hall, Upper Saddle River.

(See also **Direct Maximum Likelihood Estimation; Optimization Methods**)

CRAIG K. ENDERS

Maximum Likelihood Item Response Theory Estimation

ANDRÉ A. RUPP

Volume 3, pp. 1170–1175

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Maximum Likelihood Item Response Theory Estimation

Owing to a variety of theoretical and computational advances in the last 15 years, measurement literature has seen an upsurge in the theoretical development of modeling approaches and the practical development of routines for estimating their constituent parameters [10, 16, 19]. Despite the methodological and technical variety of these approaches, however, several common principles and approaches underlie the parameter estimation processes for them. This entry reviews these principles and approaches but, for more in-depth readings, a consultation of the comprehensive book by Baker [1], the didactic on the expectation maximization (EM)-algorithm by Harwell, Baker, and Zwarts [8] and its extension to Bayesian estimation (*see* **Bayesian Statistics; Bayesian Item Response Theory Estimation**) by Harwell and Baker [7], the technical articles on estimation with the EM-algorithm by Bock and Aitkin [2], Bock and Lieberman [3], Dempster, Laird, and Rubin [5], and Mislevy [11, 12], as well as the review article by Rupp [20], are recommended.

Conceptual Foundations for Parameter Estimation

The process of calibrating statistical models is generally concerned with two different sets of parameters, one set belonging to the assessment items and one set belonging to the examinees that interact with the items. While it is common to assign labels to item parameters such as ‘difficulty’ and ‘guessing’, no such meaning is strictly implied by the statistical models and philosophical considerations regarding parameter interpretation, albeit important, are not the focus here.

In this entry, examinees are denoted by $i = 1, \dots, I$, items are denoted by $j = 1, \dots, J$, the latent predictor variable (*see* **Latent Variable**) is unidimensional and denoted by θ , and the response probability for a correct response to a given item or item category is denoted by P or, more specifically, by $P_j(X_{ij} = x_{ij}|\theta)$. The following descriptions will

be general and, therefore, will leave the specific functional form that separates different latent variable models unspecified. However, as an example, one may want to consider the unidimensional three-parameter logistic model from item response theory (IRT) (*see* **Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data**), with the functional form

$$P_j(X_{ij} = x_{ij}|\theta) = \gamma_j + (1 - \gamma_j) \frac{\exp[\alpha_j(\theta - \beta_j)]}{1 + \exp[\alpha_j(\theta - \beta_j)]}. \quad (1)$$

To better understand common estimation approaches, it is necessary to understand a fundamental assumption that is made by most latent variable models, namely, that of *conditional* or *local independence*. This assumption states that the underlying data-generation mechanism for an observed data structure, as formalized by a statistical model, is of that dimensionality d that renders responses to individual items independent of one another for any given person. As a result, the *conditional* probability of observing a response vector $\mathbf{x}_i$ for a given examinee can then be expressed as

$$P(\mathbf{X}_i = \mathbf{x}_i|\theta) = \prod_{j=1}^J P_j(X_{ij} = x_{ij}|\theta). \quad (2)$$

Given that the responses of the examinees are independent of one another, because examinees are typically viewed as randomly and independently sampled from a population of examinees with latent variable distribution $g(\theta)$ [9], the *conditional* probability of observing all response patterns (i.e., the *conditional* probability of observing the data) can, therefore, be expressed as the double-product

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^I \prod_{j=1}^J P_j(X_{ij} = x_{ij}|\theta). \quad (3)$$

This probability, if thought of as a function of θ , is also known as the *likelihood* for the data, $L(\theta|\mathbf{X} = \mathbf{x})$. Under the assumption of θ as a random effect, one can further integrate out θ to obtain the *unconditional* or *marginal probability* of observing

2 Maximum Likelihood IRT Estimation

the data,

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^I \left\{ \int_{\Theta} \prod_{j=1}^J P_j(X_{ij} = x_{ij} | \theta_i) g(\theta) d\theta \right\}, \quad (4)$$

also known as the *marginal likelihood* for the data. Here, $g(\theta)$ denotes the probability distribution of θ in the population, which is often assumed to be standard normal but which can, technically, be of any form as long as it has interval-scale support on the real numbers.

However, while (3) and (4) are useful for a conceptual understanding of the estimation routines, it is numerically easier to work with their logarithmic counterparts. Hence, one obtains, on the basis of (3), the *log-likelihood* of the data ($\log -L$),

$$\log -L = \sum_{i=1}^I \sum_{j=1}^J \log [P_j(X_{ij} = x_{ij} | \theta)] \quad (5)$$

and, based on (4), the *marginal log-likelihood* of the data ($\log -L_M$),

$$\log -L_M = \sum_{i=1}^I \log \left\{ \int_{\Theta} \prod_{j=1}^J P_j(X_{ij} = x_{ij} | \theta) g(\theta) d\theta \right\}, \quad (6)$$

which are theoretical expressions that include latent θ values that need to be estimated for practical implementation.

Estimation of θ Values

Estimating latent θ values amounts to replacing them by manifest values from a finite subset of the real numbers. This requires the establishment of a latent variable metric and a common latent variable metric is one with a mean of 0 and a standard deviation of 1. Consequently, one obtains the estimated counterparts to (3 and 5), which are the *estimated likelihood*,

$$\tilde{P}(\mathbf{X} = \mathbf{x} | \theta) \cong \prod_{i=1}^I \prod_{j=1}^J \tilde{P}_j(X_{ij} = x_{ij} | T) \quad (7)$$

and the *estimated log-likelihood*,

$$\log -\tilde{L} \cong \sum_{i=1}^I \sum_{j=1}^J \log [\tilde{P}_j(X_{ij} = x_{ij} | T)] \quad (8)$$

where the letter T is used to denote that manifest values are used and $\tilde{P}$ indicates that this results in an estimated probability. Similarly, (4 and 6) require the estimation of the distribution $g(\theta)$ where, again, a suitable subset of the real numbers is selected (e.g., the interval from -4 to 4 if one assumes that $\theta \sim N(0, 1)$), along with a number of K evaluation points that are typically selected to be equally spaced in that interval. For each evaluation point, the approximate value of the selected density function is then computed, which can be done for a theoretically selected distribution (e.g., a standard normal distribution) or for an empirically estimated distribution (i.e., one that is estimated from the data). For example, in BILOG-MG [24] and MULTILOG [22], initial density weights, also known as *prior density weights*, are chosen to start the estimation routine, which then become adjusted or replaced at each iteration cycle by empirically estimated weights, also known as *posterior density weights* [11].

If one denotes the k th density weight by $A(T_k)$, one thus obtains, as a counterpart to (4), the *estimated marginal likelihood*,

$$\tilde{P}(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^I \left[\sum_{k=1}^K \left\{ \prod_{j=1}^J \tilde{P}_j(X_{ij} = x_{ij} | T_k) \right\} A(T_k) \right] \quad (9)$$

and, as a counterpart to (6), the *estimated marginal log-likelihood*,

$$\begin{aligned} \log -\tilde{L}_M &= \sum_{i=1}^I \log \left[\sum_{k=1}^K \left\{ \prod_{j=1}^J P_j(X_{ij} = x_{ij} | T_k) \right\} A(T_k) \right] \end{aligned} \quad (10)$$

where, again, the letter T is used to denote that manifest values are used and $\tilde{P}$ indicates that this results in an estimated probability. With these equations at hand we are now ready to discuss the three most common estimation approaches, *Joint Maximum*

Likelihood (JML), Conditional Maximum Likelihood (CML), and Marginal Maximum Likelihood (MML).

JML and CML Estimation

Historically, JML was the approach of choice for estimating item and examinee parameters. It is based on an *iterative estimation scheme* that cyclically estimates item and examinee parameters by computing the partial derivatives of the $\log -L$ (see (6 and 8)) with respect to these parameters, setting them equal to 0, and solving for the desired parameter values. At each step, provisional parameter estimates from the previous step are used to obtain updated parameter estimates for the current step. While this approach is both intuitively appealing and practically relatively easy to implement, it suffers from theoretical drawbacks. Primarily, it can be shown that the resulting parameter estimates are *not consistent*, which means that, asymptotically, they do *not* become arbitrarily close to their population targets in probability, which would be desirable [4, p. 323]. This is understandable, however, because, with each additional examinee, an additional incidental θ parameter is added to the pool of parameters to be estimated and, with each additional item, a certain number of structural item parameters is added as well, so that increasing either the number of examinees or the number of items does not improve asymptotic properties of the joint set of estimators.

An alternative estimation approach is CML, which circumvents this problem by replacing the unknown θ values directly by values of manifest statistics such as the *total score*. For this approach to behave properly, however, the manifest statistics have to be *sufficient for θ* , which means that they have to contain all available information in the data about θ and have to be directly computable. In other words, the use of CML is restricted to classes of models with *sufficient statistics* such as the Rasch model in IRT [6, 23]. Since the application of CML is restricted to subclasses of models and JML does not possess optimal estimation properties, a different approach is needed, which is the niche that MML fills.

MML Estimation

The basic idea in MML is similar to CML, namely, to overcome the theoretical inconsistency

of item parameter estimators and to resolve the iterative dependency of previous parameter estimates in JML. This is accomplished by first integrating out θ (see (4 and 6)) and then maximizing the $\log -L_M$ (see (9 and (10)) to obtain the MML estimates of the item parameters using first- and second-order derivatives and a numerical algorithm such as Newton–Raphson for solving the equations involved. The practical maximization process of (10) is done via a modification of a missing-data algorithm, the EM algorithm [2, 3, 5]. This algorithm uses the *expected* number of examinees at each evaluation point, $\bar{n}_{jk}$, and the *expected* number of correct responses at each evaluation point, $\bar{r}_{jk}$, as ‘artificial’ data and then maximizes the $\log -L_M$ at each iteration.

Specifically, the EM algorithm employs *Bayes Theorem* (see **Bayesian Belief Networks**). In general, the theorem expresses the *posterior probability* of an event, after observing the data, as a function of the *likelihood* for the data and the *prior probability* of the event, before observing the data. In order to implement the EM algorithm, the kernel of the $\log -L_M$ is expressed with respect to the *posterior probability for θ* , which is

$$P_i(\theta|\mathbf{X}_i = \mathbf{x}_i) = \frac{L(\mathbf{X}_i = \mathbf{x}_i|\theta)h(\theta)}{\int_{\Theta} L(\mathbf{X}_i = \mathbf{x}_i|\theta)h(\theta)d\theta} \quad (11)$$

where $h(\theta)$ is a *prior distribution* for θ . Using this theorem, the MML estimation process within the EM algorithm is comprised of three steps, which are repeated until *convergence* of the item parameter estimates is achieved.

First, the posterior probability of θ for each examinee i at each evaluation point k is computed via

$$P_{ik}(T_k|\mathbf{X}_i) = \frac{\left\{ \prod_{j=1}^J P_j(X_{ij} = x_{ij}|T_k) \right\} A(T_k)}{\sum_{s=1}^K \left\{ \prod_{j=1}^J P_j(X_{ij} = x_{ij}|T_s) \right\} A(T_s)} \quad (12)$$

as an approximation to (11) at evaluation point k . This is accomplished by using *provisional* item parameter estimates from the previous iteration to

compute $P_j(X_{ij} = x_{ij}|T_k)$ for a chosen model. Second, using these posterior probabilities, the artificial data for each item j at each evaluation point k are generated using

$$\begin{aligned}\bar{n}_{jk} &= \sum_{i=1}^I P_{ik}(T_k|\mathbf{X}_i) \\ \bar{r}_{jk} &= \sum_{i=1}^I X_{ij} P_{ik}(T_k|\mathbf{X}_i).\end{aligned}\quad (13)$$

Third, the first-order derivatives of the estimated $\log -L_M$ function in (10) with respect to the item parameters are set to 0 and are solved for the item parameters. In addition, the information matrix at these point estimates is computed using the Newton–Gauss/Fisher scoring algorithm to estimate their precision. For that purpose, (10) and its derivatives are rewritten using the artificial data; the entire process is then repeated until convergence of parameter estimates has been achieved. Instead of just performing the above steps, however, programs such as BILOG-MG and MULTILOG allow for a fully Bayesian estimation process. In that framework, additional prior distributions can be specified for *all* item parameters and *all* examinee distribution parameters, which are then incorporated into the $\log -L_M$ and its derivatives where they basically contribute additive terms. Finally, it should be noted that it is common to group examinees by observed response patterns and to use the observed frequencies of each response pattern to reduce the number of computations but that step is not reproduced in this exposition to preserve notational clarity.

Thus, rather than obtaining individual θ estimates, in MML one obtains item parameter estimates and only the distribution parameters of $g(\theta)$. Nevertheless, if subsequently desired, the item parameter estimates can be used as ‘known’ values to obtain estimates of the examinee parameters using *maximum likelihood (ML)*, Bayesian *expected a posteriori (EAP)* or Bayesian *maximum a posteriori (MAP)* estimation and their precision can be estimated using the *estimated information matrices* at the point estimates for the ML approach, the *estimated standard deviation of the posterior distributions* for the EAP approach, or the *estimated posterior information matrices* for the MAP approach.

Alternative Estimation Approaches

While JML, CML, and, specifically, MML within a fully Bayesian framework, are the most flexible estimation approaches, extensions of these approaches and other parameter estimation techniques exist. For example, it is possible to incorporate collateral information about examinees into the MML estimation process to improve its precision [13]. Nonparametric models sometimes require techniques such as principal components estimation [15] or kernel-smoothing [18]. More complex psychometric models, as commonly used in cognitively diagnostic assessment, for example, sometimes require specialized software routines altogether [14, 17, 21]. Almost all estimation approaches are based on the foundational principles presented herein, which substantially unify modeling approaches from an estimation perspective.

References

- [1] Baker, F.B. (1992). *Item Response Theory: Parameter Estimation Techniques*, Assessment Systems Corporation, St. Paul.
- [2] Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika* **40**, 443–459.
- [3] Bock, R.D. & Lieberman, M. (1970). Fitting a response model to n dichotomously scored items, *Psychometrika* **35**, 179–197.
- [4] Casella, G. & Berger, R.L. (1990). *Statistical Inference*, Duxbury, Belmont.
- [5] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society. Series B* **39**, 1–38.
- [6] Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*, Erlbaum, Mahwah.
- [7] Harwell, M.R. & Baker, F.B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: a didactic, *Applied Psychological Measurement* **15**, 375–389.
- [8] Harwell, M.R., Baker, F.B. & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: a didactic, *Journal of Educational Statistics* **13**, 243–271.
- [9] Holland, P.W. (1990). On the sampling theory foundations of item response theory models, *Psychometrika* **55**, 577–602.
- [10] McDonald, R.P. (1999). *Test Theory: A Unified Treatment*, Erlbaum, Mahwah.
- [11] Mislevy, R.J. (1984). Estimating latent distributions, *Psychometrika* **49**, 359–381.

-
- [12] Mislevy, R.J. (1986). Bayes modal estimation in item response models, *Psychometrika* **51**, 177–195.
- [13] Mislevy, R.J. & Sheehan, K.M. (1989). The role of collateral information about examinees in item parameter estimation, *Psychometrika* **54**, 661–679.
- [14] Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G. & Johnson, L. (2002). Making sense of data from complex assessments, *Applied Measurement in Education* **15**, 363–389.
- [15] Mokken, R.J. (1997). Nonparametric models for dichotomous responses, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer-Verlag, New York, pp. 351–368.
- [16] Muthén, B.O. (2002). Beyond SEM: general latent variable modeling, *Behaviormetrika* **20**, 81–117.
- [17] Nichols, P.D., Chipman, S.F. & Brennan, R.L., eds (1995). *Cognitively Diagnostic Assessment*, Erlbaum, Hillsdale.
- [18] Ramsay, J.O. (1997). A functional approach to modeling test data, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer-Verlag, New York, pp. 381–394.
- [19] Rupp, A.A. (2002). Feature selection for choosing and assembling measurement models: a building-block based organization, *International Journal of Testing* **2**, 311–360.
- [20] Rupp, A.A. (2003). Item response modeling with BILOG-MG and MULTILOG for windows, *International Journal of Testing* **4**, 365–384.
- [21] Tatsuoaka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach, in *Cognitively Diagnostic Assessment*, P.D. Nichols, S.F. Chipman & R.L. Brennan, eds, Erlbaum, Hillsdale, pp. 327–360.
- [22] Thissen, D. (1991). *MULTILOG: Multiple Category Item Analysis and Test Scoring Using Item Response Theory [Computer software]*, Scientific Software International, Chicago.
- [23] van der Linden, W.J. & Hambleton, R.K., eds. (1997). *Handbook of Modern Item Response Theory*, Springer-Verlag, New York.
- [24] Zimowski, M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT Analysis and Test Maintenance for Binary Items [Computer Software]*, Scientific Software International, Chicago.

ANDRÉ A. RUPP

Maxwell, Albert Ernest

DAVID J. BARTHOLOMEW

Volume 3, pp. 1175–1176

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Maxwell, Albert Ernest

Born: July 7, 1916, in County Cavan.

Died: 1996, in Leeds.

Albert Ernest Maxwell, always known as ‘Max’, is remembered primarily in the research community for his collaboration with D. N. Lawley in producing the monograph *Factor Analysis as a Statistical Method*. The first edition was published as one of Butterworth’s mathematical texts in 1963, with a reprint four years later [3] (see also [2]). The second edition followed in 1971 with publication in the United States by American Elsevier. The second edition was larger in all senses of the word and included new work by Jöreskog and others on fitting the factor model by maximum likelihood. It also contained new results on the sampling behavior of the estimates. The title of the book is significant. Up to its appearance, **factor analysis** (which had been invented by **Spearman** in 1904) had been largely developed within the psychological community – although it had points of contact with **principal component analysis** introduced by **Harold Hotelling** in 1933. Factor analysis had been largely ignored by statisticians, though **M. G. Kendall** and M. S. Bartlett were notable exceptions. More than anyone else, Lawley and Maxwell brought factor analysis onto the statistical stage and provided a definitive formulation of the factor model and its statistical properties. This is still found, essentially unchanged, in many contemporary texts. It would still be many years, however, before the prejudices of many statisticians would be overcome and the technique would obtain its rightful place in the statistical tool kit. That it has done so is due in no small measure to Maxwell’s powers of exposition as a writer and teacher. Maxwell, himself, once privately expressed a pessimistic view of the reception of this book by the statistical community but its long life and frequent citation belie that judgment.

Less widely known, perhaps, was his role as a teacher of statistics in the behavioral sciences. This was focused on his work at the Institute of Psychiatry in the University of London where he taught and advised from 1952 until his retirement in 1978. The distillation of these efforts is contained in another small but influential monograph *Multivariate Analysis in Behavioural Research* published in 1977 [5].

This ranged much more widely than factor analysis and principal components analysis, and included, for example, a chapter on the **analysis of variance** in matrix notation, which was more of a novelty when it was published in 1977 than it would be now. The book was clear and comprehensive but, perhaps, a little too concise for the average student. But, backed by good teaching, such as that Maxwell provided, it must have made a major impact on generations of research workers. Unusually perhaps, space was found for a chapter on the analysis of **contingency tables**. This summarized a topic that had been the subject of another of his earlier, characteristically brief, monographs, this time *Analysing Qualitative Data* from 1964 [4]. A final chapter by his colleague, Brian Everitt, introduced **cluster analysis**. The publication of Everitt’s monograph on that subject, initially on behalf of the long defunct UK Social Science Research Council (SSRC), was strongly supported by Maxwell [1].

Much of the best work by academics is done in the course of advisory work, supervision, refereeing, committee work, and so forth. This leaves little mark on the pages of history, but Maxwell’s role at the Institute and beyond made full use of his talents in that direction. Outside the Institute, he did a stint on the Statistics Committee of the SSRC where his profound knowledge of psychological statistics did much to set the course of funded research in his field as well as, occasionally, enlightening his fellow committee members.

Maxwell’s career did not follow the standard academic model. He developed an early interest in psychology and mathematics at Trinity College, Dublin. This was followed by a spell as a teacher at St. Patrick’s Cathedral School in Dublin of which he became the headmaster at the age of 25. His conversion to full-time academic work was made possible by the award of a Ph.D. from the University of Edinburgh in 1950. Two years later, he left teaching and took up a post as lecturer in statistics at the Institute of Psychiatry, where he remained for the rest of his working life. Latterly, he was head of the Biometrics unit at the Institute.

References

- [1] Everitt, B.S. (1974). *Cluster Analysis*, Heinemann, London.

2 Maxwell, Albert Ernest

- [2] Maxwell, A.E. (1961). Recent trends in factor analysis, *Journal of the Royal Statistical Society. Series A* **124**, 49–59.
- [3] Maxwell, A.E. & Lawley, D.N. (1963). *Factor Analysis as a Statistical Method*, 2nd Edition, 1971, Butterworths, London.
- [4] Maxwell, A.E. (1964). *Analysing Qualitative Data*, Wiley, London.
- [5] Maxwell, A.E. (1977). *Multivariate Analysis in Behavioural Research*, Chapman & Hall, London, (*Monographs on applied probability and statistics*).

DAVID J. BARTHOLOMEW

Measurement: Overview

JOEL MICHELL

Volume 3, pp. 1176–1183

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Measurement: Overview

Behavioral science admits a diverse class of practices under the heading of measurement. These extend beyond the measurement of physical and biological attributes and include transformations of frequencies (e.g., in psychometrics), summated ratings (e.g., in social psychology), and direct numerical estimates of subjective magnitudes (e.g., in psychophysics). These are used in the hope of measuring specifically psychological attributes and are important sources of data for statistical analyses in behavioral science.

Since the Second World War, the consensus within this science has been that *measurement* is ‘the assignment of numerals to objects or events according to rules’ [38 p. 667], a definition of measurement unique to social and behavioral science. Earlier definitions, such as the one suggested by the founder of quantitative behavioral science, **G. T. Fechner** [7], that ‘the measurement of a quantity consists of ascertaining how often a unit quantity of the same kind is contained in it’ (p. 38), reflect traditional quantitative science. More recent definitions, for example that ‘measurement is (or should be) a process of assigning numbers to objects in such a way that interesting qualitative empirical relations among the objects are reflected in the numbers themselves as well as in important properties of the number system’ [43 p. 394] are shaped by modern measurement theory. Recent dictionary entries, like Colman’s [5], defining measurement as the ‘systematic assignment of numbers to represent quantitative attributes of objects or events’ (p. 433), also reveal a shift from the earlier consensus. Tensions between the range of practices behavioral scientists call *measurement*, the way measurement is defined in traditional quantitative science, and various concepts of measurement within philosophy of science are responsible for this definitional diversity.

The Traditional Concept of Measurement

The traditional concept of measurement derives from Euclid’s *Elements* [10]. The category of *quantity* is central. An attribute is quantitative if its levels sustain ratios. Take length, for example. Any pair of lengths, l_1 and l_2 , interrelate additively, in the sense that there exist whole numbers, n and m , such that $nl_1 > l_2$ and

$ml_2 > l_1$. Prior to Euclid, a magnitude of a quantity, such as a specific length, was said to be *measured* by the number of units which, when added together, equaled it exactly. Such a concept of measurement does not accommodate all pairs of magnitudes, in particular, it does not accommodate incommensurable pairs. It was known that there are some pairs of lengths, for example, for which there are no whole numbers n and m such that n times the first exactly spans m times the second (e.g., the side and diagonal of a square) and that, therefore, for such pairs, the measure of the first relative to the second does not equal a numerical ratio (i.e., a ratio of two whole numbers). It was a major conceptual breakthrough to recognize (as Book V of Euclid’s *Elements* suggests) that the ratio of any pair of magnitudes, such as lengths, l_1 and l_2 (including incommensurable pairs) always falls between two infinite classes of numerical ratios as follows:

$$\{\text{the class of ratios of } n \text{ to } m\} < \text{the ratio of } l_1 \text{ to } l_2 \\ < \{\text{the class of ratios of } p \text{ to } q\}$$

(where n , m , p , and q range over all whole numbers such that $nl_2 < ml_1$ and $ql_1 < pl_2$). This meant that the measure of l_1 relative to l_2 could be understood as the positive *real number* (as it became known, much later, toward the end of the nineteenth century [6]) falling between these two classes.

This breakthrough not only explained the role of numbers in measurement, it also provided a guide to the internal structure of quantitative attributes: they are attributes in which ratios between any two levels equal positive real numbers. The German mathematician, Otto Hölder [12] (see also [27] and [28]) was the first to characterize the structure of unbounded, continuous, quantitative attributes and his seven axioms of quantity are similar to the following set [26] applying to lengths. Letting a , b , c , $\dots$, be any specific lengths and letting $a + b = c$ denote the relation holding between lengths a , b , and c when c is entirely composed of discrete parts, a and b :

1. For every pair of lengths, a and b , one and only one of the following is true:
 - (i) $a = b$;
 - (ii) there exists another length, c , such that $a = b + c$;
 - (iii) there exists another length, d , such that $b = a + d$.

2 Measurement: Overview

2. For any lengths a and b , $a + b > a$.
3. For any lengths a and b , $a + b = b + a$.
4. For any lengths a, b , and c , $a + (b + c) = (a + b) + c$.
5. For any length a , there is another length, b , such that $b < a$.
6. For any lengths a and b there is another length, c such that $c = a + b$.
7. For every nonempty class of lengths having an upper bound, there is a least upper bound.

(Note that an upper bound of a class of lengths is any length not less than any member of the class and that a least upper bound is an upper bound not greater than any other upper bound). The first four of these conditions state what it means for lengths to be additive. The remaining three ensure that the characterization excludes no lengths (i.e., there is no greatest or least length, nor gaps in the ordered series of them). All measurable, unbounded, continuous, quantitative attributes of physics (e.g., length, mass, time, etc.) are taken to possess this kind of structure. Hölder proved that ratios between levels of any attribute having this kind of structure possess the structure of the positive real numbers. This is a necessary condition for the identification of such ratios by real numbers.

If quantitative attributes are taken to have this kind of structure, then the meaning of measurement is explicit: *measurement is the estimation of the ratio between a magnitude of a quantitative attribute and a unit belonging to the same attribute*. This is the way measurement is understood in physics. For example,

Quantities are abstract concepts possessing two main properties: they can be measured, that means that the ratio of two quantities of the same kind, a pure number, can be established by experiment; and they can enter into a mathematical scheme expressing their definitions or the laws of physics. A unit for a kind of quantity is a sample of that quantity chosen by convention to have the value 1. So that, as already stated by Clerk Maxwell,

$$\text{physical quantity} = \text{pure number} \times \text{unit}. \quad (1)$$

This equation means that the ratio of the quantitative abstract concept to the unit is a pure number [41 pp. 765–766].

Measurement in physics is the fixed star relative to which measurement in other sciences steer. There always have been those who believe that behavioral

scientists must attempt to understand their quantitative practices within the framework of the traditional concept of measurement (see [25] and [26]).

However, from the time of **Fechner, Gustav T.**, it was questioned (e.g., see [45]) whether the attributes thought to be measurable within behavioral science also possess quantitative structure. As posed relative to the traditional concept of measurement, this question raised an issue of evidence: on what scientific grounds is it reasonable to conclude that an attribute possesses quantitative structure? In an important paper [11], the German scientist, Hermann von Helmholtz, made the case that this issue of evidence is solved for physical, quantitative attributes.

Helmholtz argued that evidence supporting the conclusion that the attributes measured in physical science are quantitative comes in two forms, namely, those later designated *fundamental* and *derived* measurement by Campbell [2]. In the case of fundamental measurement, evidence for quantitative structure is gained via an empirical operation for concatenating objects having different levels of the attribute, which results in those levels adding together. For example, combining rigid, straight rods linearly, end to end adds their respective lengths; placing marbles together in the same pan of a beam balance adds their respective weights; and starting a second process contiguous with the cessation of a first adds their respective durations. In this regard, the seven conditions given above describing the structure of a quantitative attribute are not causal laws. In and of themselves, they say nothing about how objects possessing levels of the attribute will behave under different conditions. These seven conditions are of the kind that J. S. Mill [29] called *uniformities of coexistence*. They specify the internal structure of an attribute, not the behavior of objects or events manifesting magnitudes of the attribute. This means that there is no logical necessity that fundamental measurement be possible for any attribute. That it is for a number of attributes (e.g., the geometric attributes, weight, and time), albeit in each case only for a very limited range of levels, is a happy accident of nature.

In the case of derived measurement, evidence for quantitative structure is indirect, depending upon relations between attributes already known to be quantitative and able to be measured. That a theoretical attribute is quantitative is indicated via derived measurement when objects believed to be equivalent with respect to the attribute also manifest a

constancy in some quantitative function of other, measurable attributes, as, for example, the ratio of mass to volume for different objects, all composed of the same kind of substance, is always a numerical constant. Then it seems reasonable to infer that the theoretical attribute is quantitative and measured by the relevant quantitative function. Since physical scientists have generally restricted the claim that attributes are quantitative to those to which either fundamental or derived measurement apply, there is little controversy, although some attributes, such as temperature, remained controversial longer than others (see, e.g., [22]). However, regarding the nonphysical attributes of behavioral science, the question of evidence for quantitative structure remained permanently controversial.

Nonetheless, and despite the arguments of Fechner's critics, the traditional concept of measurement provided no basis to reject the hypothesis that psychological attributes are quantitative. At most, it indicated a gap in the available evidence for those wishing to accept this hypothesis as an already established truth.

The Representational Concept of Measurement

By the close of the nineteenth century, the traditional concept of measurement was losing support within the philosophy of science and early in the twentieth century, the representational concept became dominant within that discipline. This was because thinking had drifted away from the traditional view of number (i.e., that they are ratios of quantities [32]) and toward the view that the concept of number is logically independent of the concept of quantity [35]. If these concepts are logically unrelated, then an alternative to the traditional account of measurement is required. Bertrand Russell suggested the representational concept, according to which measurement depends upon the existence of isomorphisms between the internal structure of attributes and the structure of subsystems of the real numbers. Since Hölder's paper [12] proves an isomorphism between quantitative attributes and positive real numbers, the representational concept fits all instances of physical measurement as easily as the traditional concept. Where the representational concept has an edge over the traditional concept is

in its capacity to accommodate the numerical representation of nonquantitative attributes. Chief amongst these are so-called *intensive magnitudes*.

While the concept of intensive magnitude had been widely discussed in the later middle ages, by the eighteenth century, its meaning had altered (see [14]). For medieval scholars, an intensive magnitude was an attribute capable of increase or decrease (i.e., what would now be called an *ordinal attribute*) understood by analogy with quantitative structure and, thereby, hypothesized to be both quantitative and measurable. Nineteenth century scholars, likewise, thought of an intensive magnitude as an attribute capable of increase or decrease, but it was thought of as one in which each degree is an indivisible unity and not able to be conceptualized as composed of parts and, so, it was one that could not be quantitative. While some intensive attributes, such as temperature was then believed to be, were associated with numbers, the association was thought of as 'correct only as to the more or less, not as to the how much' as Russell [34 p. 55] put it.

Many behavioral scientists, retreating in the face of Fechner's critics, had, by the beginning of the twentieth century, agreed that different levels of sensation were only intensive magnitudes and not fully quantitative in structure. Titchener [42 p. 48] summarized this consensus (using *S* to stand for sensation):

Now it is clear that, in a certain sense, the *S* may properly be termed a magnitude (*Grösse, grandeur*): in the sense, namely, that we speak of a 'more' and 'less' of *S*-intensity. Our second cup of coffee is sweeter than the first; the water today is colder than it was yesterday; *A*'s voice carries farther than *B*'s. On the other hand, the *S* is not, in any sense, a quantity (*messbare Grösse, Quantität, quantité*).

The representational concept of measurement provides a rationale for the conclusion that psychophysical methods enable measurement because it admits the numerical representation of intensive magnitudes.

While some advocates of the representational concept of measurement (e.g., [9] and [8]) were critical of the claims of behavioral scientists, others (e.g., [4]) argued that a range of procedures for the measurement of psychological attributes were instances of the measurement of intensive attributes and the representational concept of measurement began to find its way into the behavioral science literature (e.g., [13] and [39]). **S. S. Stevens** ([38] and [39]) took the representational concept further than previous advocates.

Stevens' interpretation of the representational theory was that measurement amounts to numerically modeling 'aspects of the empirical world' [39 p. 23]. The aspects modeled may differ, producing different types of **Scales of Measurement**: modeling a set of discrete, unordered classes gives a *nominal scale*; modeling an ordered attribute gives an *ordinal scale*; modeling differences between levels of an attribute gives an *interval scale*; and, on top of that, modeling ratios between levels of an attribute gives a *ratio scale*. Also, these different types of scales were said to differ with respect to the group of numerical transformations that change the specific numbers used but leave the type of measurement scale unchanged (see **Transformation**). Thus, any one-to-one transformation of the numbers used in a nominal scale map them into a new nominal scale; any order-preserving (i.e., increasing monotonic) transformation of the numbers used in an ordinal scale maps them into a new ordinal scale; any positive linear transformation (i.e., multiplication by a positive constant together with adding or subtracting a constant) maps the numbers used in an interval scale into a new interval scale; and, finally, any positive similarities transformation (i.e., multiplication by a positive constant) maps the numbers used in a ratio scale into a new ratio scale.

Stevens proposed a connection between type of scale and appropriateness of statistical operations. For example, he argued that the computation of means and variances was not permissible given either nominal or ordinal scale measures. He unsuccessfully attempted to justify his prescriptions on the basis of an alleged invariance of the relevant statistics under admissible scale transformations of the measurements involved. His doctrine of permissible statistics had been anticipated by Johnson [13] and it remains a controversial feature of Stevens' theory (see [24] and [44]). Otherwise, Stevens' theory of scales of measurement is still widely accepted within behavioral science.

The identification of these types of scales of measurement with associated classes of admissible transformations was an important contribution to the representational concept and it established a base for further development of this concept by the philosopher, Patrick Suppes, in association with the behavioral scientist, R. Duncan Luce. In a three-volume work, *Foundations of Measurement*, Luce and Suppes, with David Krantz and **Amos Tversky**, brought the representational concept of measurement to an

advanced state, carrying forward Hölder's axiomatic approach to measurement theory and using the conceptual framework of set theory. This approach to measurement involves four steps:

1. An empirical relational system is specified as a nonempty set of entities (objects or attributes of some kind) together with a finite number of distinct qualitative (i.e., nonnumerical) relations between the elements of this set. These elements and relations are the empirical primitives of the system.
2. A set of axioms is stated in terms of the empirical primitives. To the extent that these axioms are testable, they constitute a scientific theory. To the extent that they are supported by data, there is evidence favoring that theory.
3. A numerical relational system is identified such that a set of homomorphic or isomorphic mappings between the empirical and numerical systems can be proved to exist. This proof is referred to as a *representation theorem*.
4. A specification of how the elements of this set of homomorphisms or isomorphisms relate to one another is given, generally by identifying to which class of mathematical functions all transformations of any one element of this set into the other elements belong. A demonstration of this specification is referred to as a *uniqueness theorem* for the scale of measurement involved.

This approach was extended by these authors to the measurement of more than one attribute at a time, for example, to a situation in which the empirical relational system involves an ordering upon the levels of a dependent variable as it relates to two independent variables. As an illustration of this point, consider an ordering upon performances on a set of intellectual tasks (composing a psychological test, say) as related to the abilities of the people and the difficulties of the tasks. The elements of the relevant set are ordered triples (viz., a person's level of ability; a task's level of difficulty; and the level of performance of a person of that ability on a task of that difficulty). The relevant representation theorem proves that the order on the levels of performance satisfies certain conditions (what Krantz, Luce, Suppes & Tversky [16] call *double cancellation*, *solvability* and an *Archimedean condition*) if and only if it is isomorphic to a numerical system in which the elements are triples of real numbers, x , y , and z , such that $x + y = z$ and where

levels of ability map to the first number in each triple, levels of difficulty to the second, levels of performance to the third and order between different levels of performance maps to order between the corresponding values of these third numbers (the *z*s). The relevant uniqueness theorem proves that the mapping from the three attributes (in this case, abilities, difficulties, and performances) to the triples of numbers produce three interval scales of measurement, in Stevens' sense. This particular kind of example of the representational approach to measurement is known as *conjoint measurement* and while it was a relatively new concept within measurement theory, it has proved to be theoretically important. For example, it makes explicit the evidential basis of derived measurement in physical sciences in a way that earlier treatments never did.

Of the three conditions mentioned above (double cancellation, solvability, and the Archimedean condition), only the first is directly testable. As a result, applications of conjoint measurement theory make use of the fact that a finite substructure of any empirical system satisfying these three conditions, will always satisfy a finite hierarchy of cancellation conditions (single cancellation (or independence), double cancellation, triple cancellation, etc.). All of the conditions in this hierarchy are directly testable.

Following the lead of Suppes and Zinnes [40], Luce, Krantz, Suppes & Tversky [21] attempted to reinterpret Stevens' doctrine of permissible statistics using the concept of *meaningfulness*. Relations between the objects or attributes measured are said to be *meaningful* if and only if definable in terms of the primitives of the relevant empirical relational system. This concept of meaningfulness is related to the invariance, under admissible scale transformations of the measurements involved, of the truth (or falsity) of statements about the relation.

The best example of the application of the representational concept of measurement within behavioral science is Luce's [20] investigations of utility in situations involving individual decision making under conditions of risk and uncertainty. Luce's application demonstrates the way in which ideas in representational measurement theory may be combined with experimental research. Within the representational concept, the validity of any proposed method of measurement must always be underwritten by advances in experimental science.

Despite the comprehensive treatment of the representational concept within *Foundations of Measurement*, and also by other authors (e.g., [30] and [31]), the attitude of behavioral scientists to this concept is ambivalent. The range of practices currently accepted within behavioral science as providing interval or ratio scales extends beyond the class demonstrated to be scales of those kinds under the representational concept. These include psychometric methods, which account for the greater part of the practice of measurement in behavioral science. While certain restricted psychometric models may be given a representational interpretation (e.g., Rasch's [33] probabilistic item response model (see [15]), in general, the widespread practice of treating test scores (or components of test scores derived via multivariate techniques like linear factor analysis) as interval scale measures of psychological attributes has no justification under the representational concept of measurement. As a consequence, advocates of established practice within behavioral science have done little more than pay lip service to the representational concept.

The Operational Concept

As a movement in the philosophy of science, operationism was relatively short-lived, but it flourished long enough to gain a foothold within behavioral science, one that has not been dislodged by the philosophical arguments constructed by its many opponents. Operationism was initiated in 1927 by the physicist, P. W. Bridgman [1]. It had a profound influence upon Stevens, who became one of its leading advocates within behavioral science (see [36] and [37]). Operationism was part of a wider movement in philosophy of science, one that tended to think of scientific theories and concepts as reducible to directly observable terms or directly performable operations.

The best-known quotation from Bridgman's writings is his precept that 'in general we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*' ([1 p. 5], italics in original). If this is applied to the concept of *measurement*, then *measurement* is nothing more than 'a specified procedure of action which, when followed, yields a number' [17 p. 39].

The operational and representational concepts of measurement are not necessarily incompatible. The genius of Stevens' definition, that 'measurement is the assignment of numerals to objects or events according to rules' [38 p. 667] lies in the fact that it fits both the representational and the operationist concepts of measurement. If the 'rule' for making numerical assignments is understood as one that results in an isomorphism or homomorphism between empirical and numerical relational systems, then Stevens' definition clearly fits the representational concept. On the other hand, if the 'rule' is taken to be any number-generating procedure, then it fits the operational concept. Furthermore, when the components of the relevant empirical relational system are defined operationally, then the representational concept of measurement is given an operational interpretation. (This kind of interpretation was Stevens' [39] preference).

Within the psychometric tradition, Stevens' definition of measurement is widely endorsed (e.g., [19] and [23]). Any specific psychological test may be thought of as providing a precisely specified kind of procedure that yields a number (a test score) for any given person on each occasion of administration. The success of psychometrics is based upon the usefulness of psychological tests in educational and organizational applications and the operational concept of measurement has meant that these applications can be understood using a discourse that embodies a widely accepted concept of measurement. Any psychological test delivers measurements (in the operational sense), but not all such measurements are equally useful. The usefulness of test scores as measurements is generally understood relative to two attributes: *reliability* and *validity* (see **Reliability: Definitions and Estimation; Validity Theory and Applications** [19]). Within this tradition, any measure is thought to be additively composed of two components: a *true* component (defined as the expected value of the measures across an infinite number of hypothetical occasions of testing with person and test held constant) and an *error* component (the difference between the measure and the true component on each occasion). The reliability of a set of measurements provided by a test used within a population of people is understood as the proportion of the variance of the measurements that is *true*. The validity of a test is usually assessed via methods of linear regression (see **Multiple Linear Regression**), for *predictive* validity, and linear **factor analysis**, for *construct* validity, that is,

in terms of the proportion of true variance shared with an observed criterion variable or with a postulated theoretical construct.

The kinds of theoretical constructs postulated include cognitive abilities (such as general ability) and personality traits (such as extraversion). There has been considerable debate within behavioral science as to how realistically such constructs should be interpreted. While the inclusiveness of operationism has obvious advantages, it seems natural to interpret theoretical constructs as if they are real, causally explanatory attributes of people. This raises the issue of the internal structure of such attributes, that is, whether there is evidence that they are quantitative. In this way, thinking about measurement naturally veers away from operationism and toward the traditional concept.

Some behavioral scientists admit that numbers such as test scores, that are thought of as measurements of theoretical psychological attributes, 'generally reflect only ordinal relations between the amounts of the attribute' ([23] p. 446) (see [3] for an extended treatment of this topic and [18] for a discussion in the context of psychophysics). These authors maintain that at present there is no compelling evidence that psychological attributes are quantitative. Despite this, the habit of referring to such numbers as measurements is established in behavioral science and not likely to change. Thus, pressure will continue within behavioral science for a concept that admits as measurement the assignment of numbers to nonquantitative attributes.

References

- [1] Bridgman, P.W. (1927). *The Logic of Modern Physics*, Macmillan, New York.
- [2] Campbell, N.R. (1920). *Physics, the Elements*, Cambridge University Press, Cambridge.
- [3] Cliff, N. & Keats, J.A. (2003). *Ordinal Measurement in the Behavioral Sciences*, Lawrence Erlbaum Associates, Mahwah.
- [4] Cohen, M.R. & Nagel, E. (1934). *An Introduction to Logic and Scientific Method*, Routledge & Kegan Paul Ltd, London.
- [5] Colman, A.M. (2001). *A Dictionary of Psychology*, Oxford University Press, Oxford.
- [6] Dedekind, R. (1901). *Essays on the Theory of Numbers*, (Translated by W.W. Beman), Open Court, Chicago.
- [7] Fechner, G.T. (1860). *Elemente der Psychophysik*, Breitkopf and Hartel, Leipzig.

- [8] Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Craik, K.J.W., Drever, J., Guild, J., Houstoun, R.A., Irwin, J.C., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H. & Tucker, W.S. (1940). Quantitative estimates of sensory events: final report, *Advancement of Science* **1**, 331–349.
- [9] Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Drever, J., Guild, J., Houstoun, R.A., Irwin, J.C., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H. & Tucker, W.S. (1938). Quantitative estimates of sensory events: interim report, *British Association for the Advancement of Science* **108**, 277–334.
- [10] Heath, T.L. (1908). *The Thirteen Books of Euclid's Elements*, Vol. 2, Cambridge University Press, Cambridge.
- [11] Helmholtz, H.v. (1887). Zählen und Messen erkenntnistheoretisch betrachtet, *Philosophische Aufsätze Eduard Zeller zu seinem Fünfzigjährigen Doktorjubiläum* Gewindmet, Fues' Verlag, Leipzig.
- [12] Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass, *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse* **53**, 1–46.
- [13] Johnson, H.M. (1936). Pseudo-mathematics in the social sciences, *American Journal of Psychology* **48**, 342–351.
- [14] Kant, I. (1997). *Critique of Pure Reason*, Translated and edited by P. Guyer & A.E. Wood, eds, Cambridge University Press, Cambridge.
- [15] Keats, J. (1967). Test theory, *Annual Review of Psychology* **18**, 217–238.
- [16] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of Measurement*, Vol. 1, Academic Press, New York.
- [17] Lad, F. (1996). *Operational Subjective Statistical Methods: a Mathematical, Philosophical and Historical Introduction*, John Wiley & Sons, New York.
- [18] Laming, D. (1997). *The Measurement of Sensation*, Oxford University Press, Oxford.
- [19] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [20] Luce, R.D. (2000). *Utility of Gains and Losses: Measurement-theoretical and Experimental Approaches*, Lawrence Erlbaum, Mahwah.
- [21] Luce, R.D., Krantz, D.H., Suppes, P. & Tversky, A. (1990). *Foundations of Measurement*, Vol. 3, Academic Press, New York.
- [22] Mach, E. (1896). Critique of the concept of temperature; (Translated by Scott-Taggart, M.J. & Ellis, B. and reprinted in Ellis, B. (1968). *Basic Concepts of Measurement*, Cambridge University Press, Cambridge, pp. 183–196).
- [23] McDonald, R.P. (1999). *Test Theory: a Unified Treatment*, Lawrence Erlbaum, Mahwah.
- [24] Michell, J. (1986). Measurement scales and statistics: a clash of paradigms, *Psychological Bulletin* **100**, 398–407.
- [25] Michell, J. (1997). Quantitative science and the definition of measurement in psychology, *British Journal of Psychology* **88**, 355–383.
- [26] Michell, J. (1999). *Measurement in Psychology: a Critical History of a Methodological Concept*, Cambridge University Press, Cambridge.
- [27] Michell, J. & Ernst, C. (1996). The axioms of quantity and the theory of measurement, Part I, *Journal of Mathematical Psychology* **40**, 235–252.
- [28] Michell, J. & Ernst, C. (1997). The axioms of quantity and the theory of measurement, Part II, *Journal of Mathematical Psychology* **41**, 345–356.
- [29] Mill, J.S. (1843). *A System of Logic*, Parker, London.
- [30] Narens, L. (1985). *Abstract Measurement Theory*, MIT Press, Cambridge.
- [31] Narens, L. (2002). *Theories of Meaningfulness*, Lawrence Erlbaum, Mahwah.
- [32] Newton, I. (1728). Universal arithmetic: or, a treatise of arithmetical composition and resolution; (Reprinted in Whiteside, D.T. (1967). *The Mathematical Works of Isaac Newton*, Vol. 2, Johnson reprint Corporation, New York).
- [33] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danmarks Paedagogiske Institut, Copenhagen.
- [34] Russell, B. (1896). On some difficulties of continuous quantity, (Originally unpublished paper reprinted in Griffin, N. & Lewis, A.C., eds *The Collected Papers of Bertrand Russell, Volume 2: Philosophical Papers 1896–99*, Routledge, London, pp. 46–58).
- [35] Russell, B. (1903). *Principles of Mathematics*, Cambridge University Press, Cambridge.
- [36] Stevens, S.S. (1935). The operational definition of psychological terms, *Psychological Review* **42**, 517–527.
- [37] Stevens, S.S. (1936). Psychology: the propaedeutic science, *Philosophy of Science* **3**, 90–103.
- [38] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 667–680.
- [39] Stevens, S.S. (1951). Mathematics, measurement and psychophysics, in *Handbook of Experimental Psychology*, S.S. Stevens, ed., Wiley, New York, pp. 1–49.
- [40] Suppes, P. & Zinnes, J. (1963). Basic measurement theory, in *Handbook of Mathematical Psychology*, Vol. 1, R.D. Luce, R.R. Bush & E. Galanter, eds, Wiley, New York, pp. 1–76.
- [41] Terrien, J. (1980). The practical importance of systems of units; their trend parallels progress in physics, in *Proceedings of the International School of Physics 'Enrico Fermi' Course LXVIII, Metrology and Fundamental Constants*, A.F. Milone & P. Giacomo, eds, North-Holland, Amsterdam, pp. 765–769.
- [42] Titchener, E.B. (1905). *Experimental Psychology: A Manual of Laboratory Practice*, Macmillan, London.
- [43] Townsend, J.T. & Ashby, F.G. (1984). Measurement scales and statistics: the misconception misconceived, *Psychological Bulletin* **96**, 394–401.

8 Measurement: Overview

- [44] Velleman, P.F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading, *American Statistician* **47**, 65–72.
- [45] Von Kries, J. (1882). Über die Messung intensiver Grössen und über das sogenannte psychophysische Gesetz, *Vierteljahrsschrift für Wissenschaftliche Philosophie* **6**, 257–294.

(See also **Latent Variable; Scales of Measurement**)

JOEL MICHELL

Measures of Association

SCOTT L. HERSHBERGER AND DENNIS G. FISHER

Volume 3, pp. 1183–1192

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Measures of Association

Measures of association quantify the statistical dependence of two or more categorical variables. Many measures of association have been proposed; in this article, we restrict our attention to the ones that are most commonly used. We broadly divide the measures into those suitable for (a) unordered (nominal) categorical variables, and (b) ordered categorical variables. Within these two categories, further subdivisions can be defined. In the case of unordered categorical variables, measures of association can be categorized as (a) measures based on the odds ratio; (b) measures based on Pearson's χ^2 ; and (c) measures of predictive association. In the case of ordered (ordinal) categorical variables, measures of association can be categorized as (a) measures of concordance-discordance and (b) measures based on derived scores.

Measures of Association for Unordered Categorical Variables

Measures Based on the Odds Ratio

The Odds Ratio. The odds that an event will occur are computed by dividing the probability that the event will occur by the probability that the event will not occur (*see Odds and Odds Ratios*) [11]. For example, consider the **contingency table** shown in Table 1, with sex (s) as the row variable and employment status (e) as the column variable.

For males, the odds of being employed equal the probability of being employed divided by the probability of not being employed:

$$odds_{y|m} = \frac{p(y|m)}{1 - p(y|m)} = \frac{p(y|m)}{p(n|m)}$$

Table 1 Employment status of males and females

		Employment status		Total
		Employed (y)	Not employed (n)	
Sex	Male (m)	42	33	75
	Female (f)	40	125	165
	Total	82	58	40

$$= \frac{42/75}{33/75} = \frac{.56}{.44} = 1.27 \quad (1)$$

Similarly, for females:

$$\begin{aligned} odds_{y|f} &= \frac{p(y|f)}{1 - p(y|f)} = \frac{p(y|f)}{p(n|f)} \\ &= \frac{40/165}{125/165} = \frac{.242}{.758} = .319 \end{aligned} \quad (2)$$

These odds are large when employment is likely, and small when it is unlikely. The odds can be interpreted as the number of times that employment occurs for each time it does not. For example, since the odds of employment for males is 1.27, approximately 1.27 males are employed for every unemployed male, or, in round numbers, 5 males are employed for every four males that are unemployed. For females, there is approximately one employed female for every three unemployed females.

The association between sex and employment status can be expressed by the degree to which the two odds differ. This difference can be summarized by the odds ratio (α):

$$\alpha = \frac{odds_{y|f}}{odds_{y|m}} = \frac{p(y|f)/p(n|f)}{p(y|m)/p(n|m)} = \frac{1.27}{.32} = 3.98. \quad (3)$$

This odds ratio indicates that for every one employed female, there are four employed males. Although not necessary, it is customary to place the larger of the two odds in the numerator [11]. Also note that the odds ratio can be expressed as the cross-product of the cell frequencies of the contingency table:

$$\begin{aligned} \alpha &= \frac{p(y|m)p(n|f)}{p(y|f)p(n|m)} = \frac{(42/75)(125/165)}{(40/165)(33/75)} \\ &= \frac{(42)(125)}{(40)(33)} = 3.98. \end{aligned} \quad (4)$$

Thus, the odds ratio is also known as the cross-product ratio.

The odds ratio can be applied to larger than 2×2 **contingency tables**, although its interpretation becomes difficult when there are more than two columns. However, when the number of rows is greater than two, interpreting the odds ratio

2 Measures of Association

Table 2 Employment status within three regions

		Employment status		Total
		Employed (y)	Not employed (n)	
Region	West (w)	90	10	100
	South (s)	80	20	100
	East (e)	99	1	100
	Total	269	31	300

is relatively straightforward. For example, consider Table 2.

Within any of the three regions, we can determine the odds that someone in that region will be employed. This is accomplished by dividing the probability for those in a given region being employed by the probability for those in the same region not being employed. Thus, for the west, the odds of being employed are $(90/100)/(10/100) = 9$. For the south, the odds of being employed are $(80/100)/(20/100) = 4$. For the east, the odds of being employed are $(99/100)/(1/100) = 99$. From these three odds, we can compute the following three odds ratios: (a) the odds ratio of someone in the east being employed versus someone in west being employed ($\alpha = 99/9 = 11$); (b) the odds ratio of someone in the east being employed versus someone in the south being employed ($\alpha = 99/4 = 24.75$); and (c) the odds ratio of someone in the west being employed versus someone in the south being employed ($\alpha = 9/4 = 2.25$). Thus, the odds of someone in the east being employed are 11 times larger than the odds of someone in the west being employed, and 24.75 times larger than someone in the south. The odds of someone in the west being employed are 2.25 times larger than the odds of someone in the south being employed. The odds ratio is invariant under interchanges of rows and columns; switching only rows or columns changes α to $1/\alpha$ [11].

Odds ratios are not symmetric around one: An odds ratio larger than one by a given amount indicates a smaller effect than an odds ratio smaller than one by the same amount [11]. While the magnitude of an odds ratio less than one is restricted to the range between zero and one, odds ratios greater than one are not restricted, allowing the ratio to potentially take on any value. If the natural logarithm ($\ln$) of the odds ratio is taken, the

odds ratio is symmetric above and below one, with $\ln(1) = 0$. For example, to take the 2×2 example above, the odds ratio of a male being employed, compared to a female, was 3.98. If we reverse that and take the odds ratio of a female being employed compared to a male, it is $.32/1.27 = .25$. These ratios are clearly not symmetric about 1.00. However, $\ln(3.98) = 1.38$ and $\ln(.25) = -1.38$, and these ratios are symmetric.

Yule's Q Coefficient of Association. Both α and $\ln(\alpha)$ range from $-\infty$ to $+\infty$. In order to restrict the odds ratio within the interval -1 to $+1$, Yule introduced the coefficient of association, Q , for 2×2 tables. Its definition is [12]:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{\alpha - 1}{\alpha + 1}. \quad (5)$$

Therefore, if odds ratio (α) expressing the relationship between sex and employment is 3.98,

$$\begin{aligned} Q &= \frac{(42)(125) - (33)(40)}{(42)(125) + (33)(40)} \\ &= \frac{3.98 - 1}{3.98 + 1} = .59. \end{aligned} \quad (6)$$

Yule's Q is algebraically equal to the 2×2 Goodman-Kruskal γ coefficient (described below), and, thus, measures the degree of concordance or discordance between two variables.

Yule's Coefficient of Colligation. As an alternative to Q , Yule proposed the *coefficient of colligation* [12]:

$$\hat{Y} = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}} = \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1}. \quad (7)$$

The coefficient of colligation between sex and employment is

$$\begin{aligned} \hat{Y} &= \frac{\sqrt{(42)(125)} - \sqrt{(33)(40)}}{\sqrt{(42)(125)} + \sqrt{(33)(40)}} \\ &= \frac{\sqrt{3.98} - 1}{\sqrt{3.98} + 1} = .33. \end{aligned} \quad (8)$$

The coefficient of colligation has a different interpretation than the coefficient of association, and is interpreted as a **Pearson product-moment correlation coefficient r** (see below), but is not algebraically equivalent to one.

Both Q and $\hat{Y}$ are symmetric measures of association, and are invariant under changes in the ordering of rows and columns.

Measures Based on Pearson's χ^2

Pearson's Chi-square Goodness-of-fit Test Statistic. (See **Goodness of Fit for Categorical Variables**),

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{n_{ij}}, \quad (9)$$

where $\hat{n}_{ij}$ is the expected frequency in cell ij , and can be usefully transformed into several measures of association [1].

The Phi Coefficient. The *phi coefficient* is defined as [5]:

$$\Phi = \sqrt{\frac{\chi^2}{N}}. \quad (10)$$

For sex and employment,

$$\Phi = \sqrt{\frac{23.12}{240}} = 0.31. \quad (11)$$

The phi coefficient can vary between 0 and 1, and is algebraically equivalent to $|r|$. However, the lower and upper limits of Φ in a 2×2 table are dependent on two conditions. In order for Φ to equal -1 or $+1$, (a) $(n_{11} + n_{12}) = (n_{21} + n_{22})$, and (b) $(n_{11} + n_{21}) = (n_{12} + n_{22})$.

The Pearson Product-moment Correlation Coefficient. The general formula for the *Pearson product-moment correlation coefficient* is

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N s_x s_y}, \quad (12)$$

where X and Y are two continuous, interval-level variables. The categories of a dichotomous variable can be coded 0 and 1, and used in the formula for r . In a 2×2 contingency table, the calculations reduce to [10]:

$$r = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}. \quad (13)$$

For sex and employment, $r = \{(42)(125) - (33)(40) / \sqrt{[(75)(165)(82)(158)]}\} = .31$.

A symmetric measure of association also invariant to row and column order, r varies between -1 to $+1$. From the formula, it is apparent that $r = 1$ if $n_{12} = n_{21} = 0$, and $r = -1$ if $n_{11} = n_{22} = 0$. In a standardized 2×2 table, where each marginal probability = .5, $r = \hat{Y}$; otherwise $|r| < |\hat{Y}|$ except when the variables are independent or completely related.

Cramèr's Phi Coefficient. *Cramèr's phi coefficient* V is an extension of the phi coefficient to contingency tables that are larger than 2×2 [10]:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}, \quad (14)$$

where k is the number of rows or columns, whichever is smaller. Cramèr's phi coefficient is based on the fact that the maximum value chi-square can attain for a set of data is $N(k-1)$; thus, when chi-square is equal to $N(k-1)$, $V = 1$.

Cramèr's phi coefficient and the phi coefficient are equivalent for 2×2 tables. When applied to a 4×2 table for region and employment,

$$V = \sqrt{\frac{204.38}{500(2-1)}} = .64. \quad (15)$$

The Contingency Coefficient. Another variation of the phi coefficient, the *contingency coefficient* $\hat{C}$ is a measure of association that can be computed for an $r \times c$ of any size [9]:

$$\hat{C} = \sqrt{\frac{\chi^2}{\chi^2 + N}}. \quad (16)$$

Applied to the 4×2 table between region and employment,

$$\hat{C} = \sqrt{\frac{204.38}{204.38 + 500}} = .54. \quad (17)$$

Note that, unlike V , $\hat{C}$ does not equal Φ computed for 2×2 tables. For example, while V and Φ both equal .31 for the association between sex and employment,

$$\hat{C} = \sqrt{\frac{23.12}{23.12 + 240}} = .29. \quad (18)$$

4 Measures of Association

The contingency coefficient can never equal 1, since N cannot equal 0. Thus, the range of $\hat{C}$ is $0 \leq \hat{C} < +1$. Another limitation of $\hat{C}$ is its dependence on the number of rows and columns in an $r \times c$ table. The upper limit of $\hat{C}$ is

$$\hat{C}_{\max} = \sqrt{\frac{k-1}{k}}. \quad (19)$$

Therefore, for 2×2 tables, $\hat{C}$ can never exceed .71; that is, $\hat{C}_{\max} = \sqrt{[(2-1)/2]} = .71$. In fact, $\hat{C}$ will always be less than 1, even when the association between two variables is perfect [9]. In addition, values of $\hat{C}$ for two tables can only be compared when the tables have the same number of rows and the same number of columns. To counter these limitations, an adjustment can be applied to $\hat{C}$:

$$\hat{C}_{\text{adj}} = \frac{\hat{C}}{\hat{C}_{\max}}. \quad (20)$$

Thus, when the association between two variables is 1.00, $\hat{C}_{\text{adj}}$ will reflect this. For region and employment, $\hat{C}_{\text{adj}} = .54/.71 = .76$, and for sex and employment, $\hat{C}_{\text{adj}} = .29/.71 = .41$.

Measures of Predictive Association

Another class of measures of association for nominal variables is measures of prediction analogous in concept to the multiple correlation coefficient in regression analysis [12]. When there is an association between two nominal variables X and Y , then knowledge about X allows one to obtain knowledge about Y , more knowledge than would have been available without X . Let Δ_Y be the dispersion of Y and $\Delta_{Y.X}$ be the conditional dispersion of Y given X . A measure of prediction

$$\phi_{Y.X} = 1 - \frac{\Delta_{Y.X}}{\Delta_Y} \quad (21)$$

compares the conditional dispersion of Y given X to the unconditional dispersion of Y , similar to how the multiple correlation coefficient compares the conditional variance of the dependent variable to its unconditional variance [1]. When $\phi_{Y.X} = 0$, X and Y are independently distributed; when $\phi_{Y.X} = 1$, X is a perfect predictor of Y .

We describe four measures that operationalize the idea underlying $\phi_{Y.X} = 0$.

The first measure is the *asymmetric lambda coefficient* of Goodman and Kruskal [7]:

$$\begin{aligned} \lambda(Y = c|X = r) &= \frac{\sum_i \max_j p_{ij} - \max_j p_{+j}}{1 - \max_j p_{+j}}, \\ \lambda(X = r|Y = c) &= \frac{\sum_j \max_i p_{ij} - \max_i p_{+i}}{1 - \max_i p_{+i}}. \end{aligned} \quad (22)$$

The second is the *symmetric lambda coefficient* (λ) of Goodman and Kruskal [7]:

$$\lambda = \frac{\sum_i \max_j p_{ij} + \sum_j \max_i p_{ij} - \max_j p_{+j} - \max_i p_{+i}}{2 - \max_j p_{+j} - \max_i p_{+i}}. \quad (23)$$

The third is the *asymmetric uncertainty coefficient* of Theil [4]:

$$\begin{aligned} U(Y = c|X = r) &= \frac{\left(-\sum_i (p_{i+}) \ln(p_{i+}) \right) + \left(-\sum_j (p_{+j}) \ln(p_{+j}) \right) - \left(-\sum_i \sum_j (p_{ij}) \ln(p_{ij}) \right)}{-\sum_j (p_{+j}) \ln(p_{+j})}, \\ U(X = r|Y = c) &= \frac{\left(-\sum_i (p_{i+}) \ln(p_{i+}) \right) + \left(-\sum_j (p_{+j}) \ln(p_{+j}) \right) - \left(-\sum_i \sum_j (p_{ij}) \ln(p_{ij}) \right)}{-\sum_i (p_{i+}) \ln(p_{i+})}. \end{aligned} \quad (24)$$

The fourth measure of predictive association is U , Theil's *symmetric uncertainty coefficient* [4]:

$$U = \frac{2 \left[\left(-\sum_i (p_{i+}) \ln(p_{i+}) \right) + \left(-\sum_j (p_{+j}) \ln(p_{+j}) \right) - \left(-\sum_i \sum_j (p_{ij}) \ln(p_{ij}) \right) \right]}{\left(-\sum_i (p_{i+}) \ln(p_{i+}) \right) + \left(-\sum_j (p_{+j}) \ln(p_{+j}) \right)}. \quad (25)$$

We use the contingency table shown in Table 3 to illustrate, in turn, the computation of the four measures of predictive association.

The first measure, asymmetric $\lambda(Y|X)$, is interpreted as the relative decrease in the probability of incorrectly predicting the column variable Y between not knowing X and knowing X . As such, lambda can also be interpreted as a proportional reduction in variation measure – how much of the variance in one variable is accounted for by the other? In the equation for $\lambda(Y|X)$, $1 - \max_j p_{+j}$ is the minimum probability of error from the prediction that Y is a function of X , and $1 - \max_i p_{i+}$ is the minimum probability of error from a prediction that Y is a constant over X . For the data we obtain:

$$\begin{aligned} \lambda(Y|X) &= \frac{((49 + 44 + 63)/432) - (119/432)}{1 - (119/432)} \\ &= \frac{.36 - .28}{1 - .28} = .11. \end{aligned} \quad (26)$$

Thus, there is an 11% improvement in predicting the column variable Y given the knowledge of the row variable X . Asymmetric λ has the range $0 \leq \lambda(Y|X) \leq 1$.

Table 3 The computation of the four measures of predictive association

	Y_1	Y_2	Y_3	Y_4	Total
X_1	28	32	49	36	145
X_2	44	21	17	12	94
X_3	46	31	53	63	193
Total	118	84	119	111	

In general, $\lambda(Y|X) \neq \lambda(X|Y)$; $\lambda(X|Y)$ is the relative decrease in the probability of incorrectly predicting the row variable X between not knowing Y and knowing Y – hence, the term ‘asymmetric’. For these data,

$$\begin{aligned} \lambda(X|Y) &= \frac{((46 + 32 + 53 + 63)/432) - (193/432)}{1 - (193/432)} \\ &= \frac{.45 - .44}{1 - .44} = .02. \end{aligned} \quad (27)$$

Symmetric λ is the average of the two asymmetric lambdas and has the range $0 \leq \lambda \leq 1$. For our example,

$$\lambda = \frac{.36 + .45 - .28 - .44}{2 - .28 - .44} = \frac{.09}{1.28} = .07. \quad (28)$$

Theil’s asymmetric uncertainty coefficient, $U(Y|X)$, is the proportion of uncertainty in the column variable Y that is explained by the row variable X , or, alternatively, $U(X|Y)$ is the proportion of uncertainty in the row variable X that is explained by the column variable Y . The asymmetric uncertainty coefficient has the range $0 \leq U(Y|X) \leq 1$. For $U(Y|X)$, we obtain for these data:

$$U(Y|X) = \frac{(1.06) + (1.38) - (2.40)}{1.38} = .03, \quad (29)$$

and for $U(X|Y)$,

$$U(X|Y) = \frac{(1.06) + (1.38) - (2.40)}{1.06} = .04. \quad (30)$$

The symmetric U is computed as

$$U = \frac{2[(1.06) + (1.38) - (2.40)]}{(1.06) + (1.38)} = .03. \quad (31)$$

Both the asymmetric and symmetric uncertainty coefficients have the range $0 \leq U \leq 1$.

The quantities in the numerators of the equations for the asymmetric lambda and uncertainty coefficients are interpreted as measures of variation for nominal responses: In the case of lambda coefficients, the variation measure is called the *Gini concentration*, and the variation measure used in the uncertainty coefficients is called the *entropy* [12]. More specifically, in $\lambda(Y|X)$, the numerator represents the variance of the Y or column variable:

$$V(Y) = \sum_i \max_j p_{ij} - \max_j p_{+j}, \quad (32)$$

6 Measures of Association

and in $\lambda(X|Y)$, the numerator represents the variance of the X or row variable:

$$V(X) = \sum_j \max_i p_{ij} - \max_i p_{i+}. \quad (33)$$

Analogously, the numerator of $U(Y|X)$ is the variance of Y ,

$$\begin{aligned} V(Y) = & \left(- \sum_i (p_{i+}) \ln(p_{i+}) \right) \\ & + \left(- \sum_j (p_{+j}) \ln(p_{+j}) \right) \\ & - \left(- \sum_i \sum_j (p_{ij}) \ln(p_{ij}) \right); \end{aligned} \quad (34)$$

and the numerator of $U(X|Y)$ is the variance of X ,

$$\begin{aligned} V(X) = & \left(- \sum_i (p_{i+}) \ln(p_{i+}) \right) \\ & + \left(- \sum_j (p_{+j}) \ln(p_{+j}) \right) \\ & - \left(- \sum_i \sum_j (p_{ij}) \ln(p_{ij}) \right). \end{aligned} \quad (35)$$

However, in these measures, there is a positive relationship between the number of categories and the magnitude of the variation, thus introducing ambiguity in evaluating both the variance of the variables and their relationship.

Measures of Association for Ordered Categorical Variables

Measures of Concordance/Discordance

A pair of observations is *concordant* if the observation that ranks higher on X also ranks higher on Y . A pair of observations is *discordant* if the observation that ranks higher on X ranks lower on Y (see **Nonparametric Correlation (tau)**). The number of concordant pairs is

$$C = \sum_{i < i'} \sum_{j < j'} n_{ij} n_{i'j'}, \quad (36)$$

where the first summation is over all rows $i < i'$, and the second summation is over all pairs of columns $j < j'$. The number of discordant pairs is

$$D = \sum_{i < i'} \sum_{j > j'} n_{ij} n_{i'j'}, \quad (37)$$

where the first summation is over all rows $i < i'$, and the second summation is over all columns $j > j'$.

To illustrate the calculation of C and D , consider the 3×3 contingency table shown in Table 4 cross-classifying income with education.

In this table, the number of concordant pairs is

$$\begin{aligned} C = & 302(331 + 250 + 155 + 185) \\ & + 105(250 + 185) + 409(155 + 185) \\ & + 331(185) = 524\,112. \end{aligned} \quad (38)$$

Note that the process of identifying concordant pairs amounts to taking each cell frequency (e.g., 302) in turn, and multiplying that frequency by each of the cell frequencies 'southeast' of it (e.g., 331, 250, 155, 185).

The number of discordant pairs is

Table 4 Income for three levels of education

		Income				Proportion
		Low	Medium	High	Total	
Education	LT than high school	302	105	23	430	0.243
	High school	409	331	250	990	0.557
	GT than high school	15	155	185	355	0.200
	Total	726	591	458	1775	1.00
	Proportion	0.409	0.333	0.258	1.00	

$$\begin{aligned}
D &= 23(409 + 331 + 15 + 155) \\
&+ 105(409 + 15) + 250(15 + 155) \\
&+ 331(15) = 112\,915. \quad (39)
\end{aligned}$$

Discordant pairs can readily be identified by taking each cell frequency (e.g., 23) in turn, and multiplying that cell frequencies by each of the cell frequencies ‘southwest’ of it (e.g., 409, 331, 15, 155).

Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$. We can express the total number of pairs of observations as

$$\frac{n(n-1)}{2} = C + D + T_X + T_Y - T_{XY}, \quad (40)$$

where $T_X = \sum_i n_i(n_{i+} - 1)/2$ is the number of pairs tied on the row variable X , $T_Y = \sum_j n_j(n_{+j} - 1)/2$ is the number of pairs tied on the column variable Y , and $T_{XY} = \sum_i \sum_j n_{ij}(n_{ij} - 1)/2$ is the number of pairs from a common cell (tied on X and Y). In this formula for $n(n-1)/2$, T_{XY} is subtracted because pairs tied on both X and Y have been counted twice, once in T_X and once in T_Y . Therefore,

$$\begin{aligned}
T_X &= \frac{(430 \times 429) + (990 \times 989) + (355 \times 354)}{2} \\
&= 644\,625 \\
T_Y &= \frac{(726 \times 725) + (591 \times 590) + (458 \times 457)}{2} \\
&= 542\,173 \\
T_{XY} &= \frac{(302 \times 301) + (105 \times 104) + \dots + (185 \times 184)}{2} = 249\,400 \\
\frac{n(n-1)}{2} &= (524\,112) + (112\,915) + (644\,625) \\
&\quad + (542\,173) - (249\,400) \\
&= 1\,574\,425. \quad (41)
\end{aligned}$$

Kendall’s tau Statistics. Several measures of concordance/discordance are based on the difference $C - D$. Kendall’s three *tau* (τ) statistics are among the most well known of these, and can be applied to $r \times c$ contingency tables with $r \geq 2$ and $c \geq 2$. The row and column variables do not have to have the

same number of categories. The tau statistics have the range $-1 \leq 0 \leq +1$.

The three tau statistics are [6]:

$$\begin{aligned}
\tau_a &= \frac{2(C - D)}{N(N - 1)} \\
\tau_b &= \frac{(C - D)}{\sqrt{[(C + D + T_Y - T_{XY})(C + D + T_X - T_{XY})]}} \\
\tau_c &= \frac{2m(C - D)}{N^2(m - 1)}, \quad (42)
\end{aligned}$$

where m = the number of rows or column, whichever is smaller. For our data, the tau statistics are equal to

$$\begin{aligned}
\tau_a &= \frac{2(524\,112 - 112\,915)}{1775(1775 - 1)} = .26 \\
\tau_b &= \frac{(524\,112 - 112\,915)}{\sqrt{[(524\,112 + 112\,915 + 542\,173 - 249\,900) \\
&\quad (524\,112 + 112\,915 + 644\,625 - 249\,900)]}} \\
&= .42 \\
\tau_c &= \frac{3 \times 2(524\,112 - 112\,915)}{1775^2(3 - 1)} = .39. \quad (43)
\end{aligned}$$

One note of qualification is that tau-a assumes there are no tied observations; thus, strictly speaking, it is not applicable to contingency tables. It is generally more useful as a measure of rank correlation. For 2×2 tables, τ_b simplifies to the Pearson product-moment correlation obtained by assigning any scores to the rows and columns consistent with their orderings.

Goodman and Kruskal’s Gamma. Goodman and Kruskal suggested yet another measure, *gamma* (γ) based on $C - D$ [7]:

$$\gamma = \frac{C - D}{C + D}. \quad (44)$$

For our example data, $\gamma = (524\,112 - 112\,915)/(524\,112 + 112\,915) = .65$. Gamma has the range $-1 \leq 0 \leq +1$, equal to $+1$ when the data are concentrated in the upper-left to lower-right diagonal, and equal to -1 for the converse pattern. Although γ does equal zero when the two variables are independent, two variables can be completely dependent and still have a value of γ less than unity. For 2×2 tables, γ is equal to Yule’s Q .

8 Measures of Association

Somers' $\hat{d}$. Gamma and the tau statistics are symmetric measures. Somers proposed an alternative asymmetric measure [7]:

$$\hat{d}(Y = c|X = r) = \frac{(C - D)}{\left[\frac{n(n-1)}{2} - T_X \right]}, \quad (45)$$

for predicting the column variable Y from the row variable X , or

$$\hat{d}(X = r|Y = c) = \frac{(C - D)}{\left[\frac{n(n-1)}{2} - T_Y \right]}, \quad (46)$$

for predicting X from Y . For these data, we compute

$$\begin{aligned} \hat{d}(Y|X) &= \frac{(524\,112 - 112\,915)}{(1\,574\,425 - 644\,625)} = .44 \\ \hat{d}(X|Y) &= \frac{(524\,112 - 112\,915)}{(1\,574\,425 - 542\,173)} = .40. \end{aligned} \quad (47)$$

Somers' $\hat{d}$ can be interpreted as the difference between the proportions of concordant and discordant pairs, using only those pairs that are untied on the predictor variable. For 2×2 tables, Somers' $\hat{d}$ simplifies to the difference of proportions $n_{11}/n_{1+} - n_{21}/n_{2+}$ (for $\hat{d}(Y|X)$), or to the difference of proportions $n_{11}/n_{+1} - n_{12}/n_{+2}$ (for $\hat{d}(X|Y)$). Yet another interpretation of Somers' d is as a least-squares regression slope [4]. To see this more clearly, we rewrite $\hat{d}(Y|X)$ as $\hat{d}_{YX}$ and $\hat{d}(X|Y)$ as $\hat{d}_{XY}$, we note that

$$\begin{aligned} \tau_b^2 &= \hat{d}_{YX}\hat{d}_{XY} \\ .42^2 &= (.44)(.40). \end{aligned} \quad (48)$$

Recalling in least-squares regression that

$$r^2 = b_{YX}b_{XY}, \quad (49)$$

we see the analogy between τ_b^2 and r^2 as coefficients of determination, and $\hat{d}_{YX}$, $\hat{d}_{XY}$ and b_{YX} , b_{XY} as regression coefficients. In addition, we note, as is true for γ , Somers' $\hat{d}$ equals zero when the two variables are independent, but is not necessarily equal to unity when the two variables are completely dependent.

Measures Based on Derived Scores

Some methods for measuring the association between ordinal variables require assigning scores to the levels of the ordinal variables (see **Ordinal Regression**

Models). When a contingency table is involved, scale values are assigned to row and column categories, the data are treated as a grouped frequency distribution, and the Pearson product-moment correlation is computed.

Specifically, let x_i and y_j be the values assigned to the rows and columns. The Pearson product-moment correlation for grouped data is then

$$r = \frac{CP_{XY}}{\sqrt{SS_X SS_Y}}, \quad (50)$$

where CP_{XY} is the sum of cross products,

$$CP_{XY} = \sum_{i,j} x_i y_j n_{ij} - \frac{\left(\sum_i x_i n_{i+} \right) \left(\sum_j y_j n_{+j} \right)}{n_{++}}, \quad (51)$$

and SS_X and SS_Y are the sums of squares,

$$\begin{aligned} SS_X &= \sum_i x_i^2 n_{i+} - \frac{\left(\sum_i x_i n_{i+} \right)^2}{n_{++}}, \\ SS_Y &= \sum_j y_j^2 n_{+j} - \frac{\left(\sum_j y_j n_{+j} \right)^2}{n_{++}}. \end{aligned} \quad (52)$$

Therefore, if for our 3×3 contingency table for income and education, we assign the values, '1', '2', and '3' to the three levels of each variable, we have

$$\begin{aligned} CP_{XY} &= 6863 - \frac{(3475)(3282)}{1775} = 437.68, \\ SS_X &= 7585 - \frac{(3475)^2}{1775} = 781.83, \\ SS_Y &= 7212 - \frac{(3282)^2}{1775} = 1143.54, \\ r &= \frac{437.68}{\sqrt{(781.83)(1143.54)}} = .46. \end{aligned} \quad (53)$$

The value of $r = .46$ is also known as the *Spearman rank correlation coefficient* r_s , sometimes referred to as **Spearman's rho** [8]. The Spearman rank correlation coefficient can also be computed by using:

$$r_s = \frac{6 \sum d^2}{n(n^2 - 1)}, \quad (54)$$

where d is the difference in the rank scores between X and Y . Computed by using this formula,

$$r_s = \frac{6(20706)^2}{1775(1775^2 - 1)} = .46. \quad (55)$$

Instead of using the rank scores of '1', '2', and '3' directly, we can use *ridit scores* [2]. Ridit scores are cumulative probabilities; each ridit score represents the proportion of observations below category j plus half the proportion within j . We illustrate the calculation of ridit scores using the income and education contingency table. Beneath each marginal total is the proportion of observations in the row or column. The ridit score for the low income category is $.5(.409) = .205$; for medium income, $.409 + .5(.333) = .576$; and for high income, the ridit score is $.409 + .333 + .5(.258) = .871$. The ridit scores for less than high school education is $.5(.243) = .122$; for high school education, $.243 + .5(.557) = .522$; and for greater than high school education, the ridit score is $.243 + .557 + .5(.200) = .900$. Applying the Pearson product-moment correlation formula to the ridit scores, we obtain $r_s = .47$.

Scores for the ordered levels of categorical variable can also be derived by assuming that the variables have an underlying bivariate normal distribution. *Normal scores* computed from 2×2 contingency tables produce the **tetrachoric correlation**; when computed from larger tables, the correlation is known as the **polychoric correlation** [3]. The Tetrachoric/polychoric correlation is the **maximum likelihood estimate** of the Pearson product-moment

correlation between the bivariate normally distributed variables. The polychoric correlation between education and income is .57.

References

- [1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, Wiley, New York.
- [2] Bross, I.D.J. (1958). How to use ridit analysis, *Biometrics* **14**, 18–38.
- [3] Drasgow, F. (1986). Polychoric and polyserial correlations, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 68–74.
- [4] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, 2nd Edition, Chapman & Hall, Boca Raton.
- [5] Fleiss, J.L. *Statistical Methods for Rates and Proportions*, 2nd Edition, Wiley, New York.
- [6] Gibbons, J.D. (1993). *Nonparametric Measures of Association*, Sage Publications, Newbury Park.
- [7] Goodman, L.A. & Kruskal, W.H. (1972). Measures of association for cross-classification, IV, *Journal of the American Statistical Association* **67**, 415–421.
- [8] Hildebrand, D.K., Laing, J.D. & Rosenthal, H. (1977). *Analysis of Ordinal Data*, Sage Publications, Newbury Park.
- [9] Liebetrau, A.M. (1983). *Measures of Association*, Sage Publications, Newbury Park.
- [10] Reynolds, H.T. (1984). *Analysis of Nominal Data*, Sage Publications, Newbury Park.
- [11] Rudas, T. (1998). *Odds Ratios in the Analysis of Contingency Tables*, Sage Publications, Thousand Oaks.
- [12] Wickens, T.D. (1986). *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum, Hillsdale.

SCOTT L. HERSHBERGER AND DENNIS
G. FISHER

Median

DAVID CLARK-CARTER

Volume 3, pp. 1192–1193

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Median

The median is a measure of central tendency (or location). It is defined as the point in an ordered set of data such that 50% of scores fall below it and 50% above it. The median is thus the 50th **percentile** and the second **quartile** (Q_2).

To find the median, we first place a set of scores in order of their size and then locate the value that is in the middle of the ordered set. This is straightforward when there is an odd number of scores in the set, as the median will be the member of the set that has as many scores below it as above it. For example, for the set of seven numbers 1, 3, 4, 5, 7, 9, and 11, 5 is the median and there are three numbers below it and three above it.

Formally, when the number of scores in the set (N) is odd, the median is the $[(N + 1)/2]$ th value in the set. Accordingly, for the above-mentioned data, the median is the 4th in the ordered set, which is, as before, 5.

When there is an even number of values in the set, the median is taken to be the mean of the $(N/2)$ th and the $[N/2 + 1]$ th values. Suppose that the above-mentioned set had the value 15 added to it so that it now contained eight numbers, then the median would be the mean of the 4th and 5th scores in the ordered set, that is, the mean of 7 and 9, which is 8.

When data are presented as frequencies within ranges rather than as individual values, it is possible to estimate the value of the median using interpolation.

Suppose that we want to find the median age of a set of 100 young people from the data presented in Table 1.

It can be seen quite easily that the median is in the age range 10 to 14 years. However, as we have no

other information, we are forced to assume that the ages of the 35 people in that age range are evenly distributed across it. Given that assumption, we can see that the median age is at the beginning of that range, as the 48th percentile (from the cumulative frequency) is in the age range 5 to 9.

One equation for calculating an approximate value for the median in such a situation is

$$\text{Median} = L_m + \left[C_m \times \left\{ \frac{\left(\frac{1}{2} \times N \right) - F_{m-1}}{F_m} \right\} \right], \quad (1)$$

where L_m is the lowest value in the range that contains the median, C_m is the width of the range that contains the median, F_{m-1} is the cumulative frequency below the range that contains the median, F_m is the frequency within the range that contains the median, and N is the total sample size.

Therefore, in the tabulated data, the median would be found from

$$10 + \left[5 \times \left\{ \frac{\left(\frac{1}{2} \times 100 \right) - 48}{35} \right\} \right] = 10.29. \quad (2)$$

An advantage of the median is that it is not affected by extreme scores; for example, if instead of adding 15 to the original 7-item set above, we added 100, then the median would still be 8. However, as with the mean, with discrete distributions, a data set containing an even number of values can produce a median, which is not possible, for example, if it was found that the median number of people in a family is 3.5.

The median is an important robust measure of location for **exploratory data analysis** (EDA) and is a key element of a box plot [1].

Reference

- [1] Cleveland, W.S. (1985). *The Elements of Graphing Data*, Wadsworth, Monterey.

DAVID CLARK-CARTER

Table 1 Frequency table of ages of 100 young people

Age in years	Frequency	Cumulative frequency
0 to 4	26	26
5 to 9	22	48
10 to 14	35	83
15 to 19	17	100

Median Absolute Deviation

DAVID C. HOWELL

Volume 3, pp. 1193–1193

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Median Absolute Deviation

The median deviation (M.A.D.) is a robust measure of scale, usually defined as

$$\text{M.A.D.} = \text{median}|X_i - \text{median}|. \quad (1)$$

We can see that the M.A.D. is simply the average distance of observations from the median of the distribution.

For a normally distributed variable, the M.A.D. is equal to 0.6745 times the standard deviation (SD) and 0.8453 times the **average deviation** (AD). (For distributions of the same shape, these estimators are linearly related.) The M.A.D. becomes increasingly

smaller than the standard deviation for distributions with thicker tails.

The major advantage of the M.A.D. over the AD, both of which are more efficient than the SD for even slightly heavier than normal distributions, is that the M.A.D. is more robust. This led Mosteller and Tukey [1] to class the M.A.D. and the interquartile range as robust measures of scale, as ‘the best of an inferior lot.’

Reference

- [1] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Linear Regression*, Addison-Wesley, Reading.

DAVID C. HOWELL

Median Test

SHLOMO SAWILOWSKY AND CLIFFORD E. LUNNEBORG

Volume 3, pp. 1193–1194

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Median Test

This test for the equivalence of the medians of $k \geq 2$ sampled populations was introduced by Mood in 1950 [3]. The null hypothesis is that the k populations have the same median and the alternative is that at least two of the populations have differing medians.

The test requires random samples to be taken independently from the k populations and is based on the following assumption. If all the populations have a common median, then the probability that a randomly sampled value will lie above the grand sample median will be the same for each of the populations [1].

To carry out the test, one first finds the grand sample median, the median of the combined samples. Then, the sample data are reduced to a two-way table of frequencies. The k columns of the table correspond to the k samples and the two rows correspond to sample observations falling above and at or below the grand sample median.

As an example, [2] reports the results of an experiment on the impact of caffeine on finger tapping rates. Volunteers were trained at the task and then randomized to receive a drink containing 0 mg, 100 mg, or 200 mg caffeine. Two hours after

Table 1 Caffeine and finger tapping rates

Dosage (mg)	Tapping speeds
0	242, 245, 244, 248, 247, 248, 242, 244, 246, 242
100	248, 246, 245, 247, 248, 250, 247, 246, 243, 244
200	246, 248, 250, 252, 248, 250, 246, 248, 245, 250

Table 2 Distribution of tapping speeds about grand median

	0 mg	100 mg	200 mg
Above grand median	3	5	7
Below grand median	7	5	3

consuming the drink, the rates of tapping, given in Table 1, were recorded.

The grand sample median for these data is a value between 246 and 247. By custom, it is taken to be midway, at 246.5. Table 2 describes how the 30 tapping speeds, 10 in each caffeine dosage group, are distributed about this grand median.

Where, as here, $k > 2$, the test is completed as a chi-squared test of independence in a two-way table. The resulting P value is 0.2019. As the expected cell frequencies are small in this example, the result is only approximate.

Where $k = 2$, Fisher’s exact test (*see Exact Methods for Categorical Data*) can be used to test for independence. As the name implies, the resulting P value is exact rather than approximate.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [2] Draper, N.R. & Smith, N. (1981). *Applied Regression Analysis*, 2nd Edition, Wiley, New York.
- [3] Mood, A.M. (1950). *Introduction to the Theory of Statistics*, McGraw Hill, New York.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY AND
CLIFFORD E. LUNNEBORG

Mediation

DAVID A. KENNY

Volume 3, pp. 1194–1198

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mediation

Mediation presumes a causal chain. As seen in Figure 1, the variable X causes M (path a) and M in turn causes Y (path b). The variable M is said to *mediate* the X to Y relationship. *Complete mediation* is defined as the case in which variable X no longer affects Y after M has been controlled and so path c' is zero. *Partial mediation* is defined as the case in which the effect of X on Y is reduced when M is introduced, but c' is not equal to zero. The path from X to Y without M being controlled for is designated here as c . A mediator is sometimes called an *intervening variable* and is said to *elaborate* a relationship.

Given the model in the Figure 1, there are, according to Baron and Kenny [1], four steps in a mediational analysis. They are as follows.

- Step 1:* Show that X predicts the Y . Use Y as the criterion variable in a regression equation and X as a predictor. This step establishes that X is effective and that there is some effect that may be mediated.
- Step 2:* Show that X predicts the mediator, M . Use M as the criterion variable in a regression equation and X as a predictor (estimate and test path a in Figure 1). This step essentially involves treating the mediator as if it were an outcome variable.
- Step 3:* Show that the mediator, M , predicts Y . Use Y as the criterion variable in a regression equation and X and M as predictors (estimate and test path b in the Figure 1). It

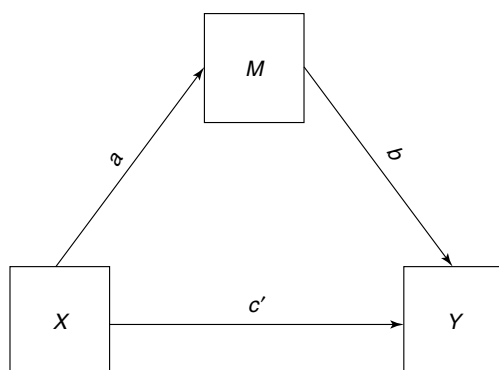


Figure 1 Basic mediational model: The variable M mediates the X to Y relationship

is insufficient just to correlate the mediator with Y because the mediator and Y may be both caused by X . The effect of X must be controlled in establishing the effect of the mediator on Y .

- Step 4:* To establish that M *completely* mediates the $X - Y$ relationship, the effect of X on Y controlling for M should be zero (estimate and test path c' in Figure 1). The effects in both steps 3 and 4 are estimated in the same regression equation where Y is the criterion and X and M are the predictors.

If all of these steps are met, then the data are consistent with the hypothesis that M completely mediates the $X - Y$ relationship. Partial mediation is demonstrated by meeting all but Step 4. It can sometimes happen that Step 1 fails because of a suppressor effect (c' and ab have different signs). Hence, the essential steps in testing for mediation are 2 and 3.

Meeting these steps does not, however, conclusively prove the mediation hypothesis because there are other models that are consistent with the data. So for instance, if one makes the mediator the outcome and vice versa, the results may look like 'mediation'. Design (e.g., measuring M before Y) and measurement (e.g., not using self-report to measure M and Y) features can strengthen the confidence that the model in the Figure 1 is correct.

If both steps 2 and 3 are met, then necessarily the effect of X on Y is reduced when M is introduced. Theoretically, the amount of reduction, also called the *indirect effect*, or $c - c'$ equals ab . So, the amount of reduction in the effect of X on Y is ab and not the change in an inferential statistic such as, F , p , or variance explained. Mediation or the indirect effect equals, in principle, the reduction in the effect of X to Y when the mediator is introduced. The fundamental equation in mediational analysis is

$$\text{Total effect} = \text{Direct Effect} + \text{Indirect Effect} \quad (1)$$

or mathematically

$$c = c' + ab \quad (2)$$

When **multiple linear regression** is used, ab exactly equals $c - c'$, regardless of whether standardized or unstandardized coefficients are used (see **Standardized Regression Coefficients**). However,

for **logistic regression**, multilevel modeling (*see Generalized Linear Mixed Models*), and **structural equation modeling**, ab only approximately equals $c - c'$. In such cases, it is probably advisable to estimate c by $ab + c'$ instead of estimating it directly.

There are two major ways to evaluate the null hypothesis that $ab = 0$. The simple approach is to test each of the two paths individually and if both are statistically significant, then mediation is indicated [4]. Alternatively, many researchers use a test developed by Sobel [11] that involves standard error of ab . It requires the standard error of a or s_a (which equals a/t_a , where t_a is the t Test of coefficient a) and the standard error of b or s_b (which equals b/t_b). The standard error of ab can be shown to equal approximately the square root of $b^2s_a^2 + a^2s_b^2$. The test of the indirect effect is given by dividing ab by the square root of the above variance and treating the ratio as a Z test (i.e., larger than 1.96 in absolute value is significant at the 0.05 level, two tailed).

The power of the Sobel test is very low and the test is very conservative, yielding too few Type I errors. For instance, in a simulation conducted by Hoyle and Kenny [2], in one condition, the Type I error rate was found sometimes to be below 0.05 even when the indirect effect was not zero. Hoyle and Kenny (1999) recommend having at least 200 cases for this test. Several groups of researchers (e.g., [7, 9]) are seeking an alternative to the Sobel test that is not so conservative.

The mediation model in the Figure 1 may not be properly specified. If the variable X is an intervention, then it can be assumed to be measured without error and that it may cause M and Y , and not vice versa. If the mediator has measurement error (i.e., has less than perfect reliability), then it is likely that effects are biased. The effect of the mediator on the outcome (path b) is underestimated and the effect of the intervention on the outcome (path c') tends to be overestimated (given that ab is positive). The overestimation of this path is exacerbated to the extent to which the mediator is caused by the intervention. To remove the biasing effect of measurement error, multiple indicators of the mediators can be found and a latent variable approach can be employed. One would need to use a structural equation modeling program (e.g., AMOS or LISREL – *see Structural Equation Modeling: Software*) to estimate such a model. If a **latent variable** approach is not used, the

researcher needs to demonstrate that the reliability of the mediator is very high.

Another specification error is that the mediator may be caused by the outcome variable (Y causes M). Because the intervention is typically a manipulated variable, the direction of causation from it is known. However, because both the mediator and the outcome variables are not manipulated variables, they may both cause each other. For instance, it might be the case that the ‘outcome’ may actually mediate the effect of the intervention on the ‘mediator’. Generally, reverse casual effects are ruled out theoretically. That is, a causal effect from Y to M does not make theoretical sense. Design considerations may also lessen the plausibility of reverse causation effects. The mediator should be measured prior to the outcome variable and efforts should be made to determine that the two do not share method effects (e.g., both self-reports from the same person). If one can assume that c' is zero, then reverse causal effects can be estimated. That is, if one assumes that there is complete mediation (X does not cause Y), one can allow for the mediator to cause the outcome and the outcome to cause the mediator.

Smith [10] discussed a method to allow for estimation of reverse causal effects. In essence, both the mediator and the outcome variable are treated as outcome variables and they each may mediate the effects of the other. To be able to employ the Smith approach, there must be for both the mediator and the outcome variable a variable that is known to cause each of them but does not cause the other.

Another specification error is that M does not cause Y , but rather some unmeasured variable causes both variables. For instance, the covariation between M and Y is due to social desirability. Other names for spuriousness are confounding, omitted variables, selection, and the third-variable problem.

The only real solution for spuriousness (besides randomization) is to measure potentially confounding variables. Thus, if there are concerns about social desirability, that variable should be measured and controlled for in the analysis. One can also estimate how strong spuriousness must be to explain the effect [8].

If M is a successful mediator, it is then necessarily correlated with the intervention (path a) and so there will be collinearity. This will affect the precision of the estimates of the last set of regression equations. Thus, the **power** of the test of the coefficients b and c' will be compromised. (The effective sample size for these

tests is approximately $N(1 - a^2)$ where N is the overall sample size and a is a standardized path.) Therefore, if M is a strong mediator (path a), to achieve equivalent power, the sample size might have to be larger than what it would be if M were a weak mediator.

Sometimes the mediator is chosen too close to the intervention and path a is relatively large and b small. Such a *proximal* mediator can sometimes be viewed as a manipulation check. Alternatively, sometimes the mediator is chosen too close to the outcome (a *distal* mediator), and so b is large and a is small. Ideally, the standardized values of a and b should be nearly equal. Even better in terms of maximizing power, standardized b should be somewhat larger than standardized a . The net effect is then that sometimes the power of the test of ab can be increased by lowering the size of a , thereby lowering the size of ab . Bigger is not necessarily better.

The model in the Figure 1 may be more complicated in that there may be additional variables. There might be covariates and there might be multiple interventions (X s), mediators (M s), or outcomes (Y s).

A *covariate* is a variable that a researcher wishes to control for in the analysis. Such variables might be background variables (e.g., age and gender). They would be included in all of the regression analyses. If such variables are allowed to interact with X or M , they would become moderator variables.

Consider the case in which there is X_1 and X_2 . For instance, these variables might be two different components of a treatment. It could then be determined if M mediates the $X_1 - Y$ and $X_2 - Y$ relationships. So in Step 1, Y is regressed on X_1 and X_2 . In Step 2, M is regressed on X_1 and X_2 . Finally, in Steps 2 and 3, Y is regressed on M , X_1 , and X_2 . Structural equation modeling might be used to conduct simultaneous tests.

Consider the case in which M_1 and M_2 mediate the $X - Y$ relationship. For Step 2, two regressions are performed, one for M_1 and M_2 . For Steps 3 and 4, one can test each mediator separately or both in combination. There is greater parsimony in a combined analysis but power may suffer if the two mediators are correlated. Of course, if they are strongly correlated, the researcher might consider combining them into a single variable or having the two serve as indicators of the same latent variable.

Consider the case in which there are two outcome variables, Y_1 and Y_2 . The Baron and Kenny steps would be done for each mediator. Step 2 would need

to be done only once since Y is not involved in that regression.

Sometimes the interactive effect of two independent variables on an outcome variable is mediated by the effect of a process variable. For instance, an intervention may be more effective for men versus women. All the steps would be repeated but now X would not be a main effect but an interaction. The variable X would be included in the analysis, but its effect would not be the central focus. Rather, the focus would be on the interaction of X .

Sometimes a variable acts as a mediator for some groups (e.g., males) and not others (e.g., females). There are two different ways in which the mediation may be moderated. The mediator interacts with some variable to cause the outcome (path b varies) or the intervention interacts with a variable to cause the mediator (path a varies).

In this case, all the variables cause one another. For simple models, one is better off using multiple regression, as Structural Equation Modeling (SEM) tests are only approximate. However, if one wants to test relatively complex hypotheses:

M_1 , M_2 , and M_3 do not mediate the $X - Y$ relationship,

M mediates the X to Y_1 and X to Y_2 relationship,

M mediates the X_1 to Y and X_2 to Y relationship, or a equals b ,

then SEM can be a useful method. Some structural equation modeling computer programs provide measures and tests of indirect effects.

In some cases, the variables X , M , or Y are measured with error, and multiple indicators are used to measure them. SEM can be used to perform the Baron and Kenny steps. However, the following difficulty arises: the measurement model would be somewhat different each time that the model is estimated. Thus, c and c' are not directly comparable because the meaning of Y varies. If, however, the measurement model were the same in both models (e.g., the loadings were fixed to one in both analyses), the analyses would be comparable.

Sometimes, X , M , and Y are measured repeatedly for each person. Judd, Kenny, and McClelland [3] discuss the case in which there are two measurements per participant. They discuss creating difference scores for X , M , and Y , and then examine the extent to which the differences in X and M predict

differences in Y . If the effect of the difference in X affects a difference in Y less when the difference in M is introduced, there is evidence of mediation.

Alternatively, mediational models can be tested using multilevel modeling [5]. Within this approach, the effect of X and M on Y can be estimated, assuming there is a sufficient number of observations for each person. The advantages of multilevel modeling over the previously described difference-score approach are several:

the estimate of the effect is more precise
a statistical evaluation of whether mediation differs by persons
missing data and unequal observations are less problematic

Within multilevel modeling, it does happen that c exactly equals $c' + ab$. The decomposition only approximately holds. Krull and MacKinnon [6] discuss the degree of difference in these two estimates.

One potential advantage of a multilevel mediational analysis is that it can be tested if the mediation effects vary by participant. So M might be a stronger mediator for some persons than for others. If the mediator were coping style and the causal effect were the effect of stress on mood, it might be that a particular coping style is more effective for some people than for others (moderated mediation).

In conclusion, mediational analysis appears to be a simple task. A series of multiple regression analyses are performed, and, given the pattern of statistically significant regression coefficients, mediation can be inferred. In actuality, mediational analysis is not formulaic and careful attention needs to be given to questions concerning the correct specification of the causal model, a thorough understanding of the area of study, and knowledge of the measures and design of the study.

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] Hoyle, R.H. & Kenny, D.A. (1999). Statistical power and tests of mediation, in *Statistical Strategies for Small Sample Research*, R.H. Hoyle ed., Sage Publications, Newbury Park.
- [3] Judd, C.M., Kenny, D.A. & McClelland, G.H. (2001). Estimating and testing mediation and moderation in within-participant designs, *Psychological Methods* **6**, 115–134.
- [4] Kenny, D.A., Kashy, D.A. & Bolger, N. (1998). Data analysis in social psychology, in *The handbook of social psychology*, Vol. 1, 4th Edition, D. Gilbert, S. Fiske & G. Lindzey, eds, McGraw-Hill, Boston, pp. 233–265.
- [5] Kenny, D.A., Korchmaros, J.D. & Bolger, N. (2003). Lower level mediation in multilevel models, *Psychological Methods* **8**, 115–128.
- [6] Krull, J.L. & MacKinnon, D.P. (1999). Multilevel mediation modeling in group-based intervention studies, *Evaluation Review* **23**, 418–444.
- [7] MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G. & Sheets, V. (2002). A comparison of methods to test the significance of the mediated effect, *Psychological Methods* **7**, 82–104.
- [8] Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): a method for the estimating the effects of omitted variables, *Psychological Bulletin* **108**, 314–329.
- [9] Shrout, P.E. & Bolger, N. (2002). Mediation in experimental and nonexperimental studies. New procedures and recommendations, *Psychological Methods* **7**, 422–445.
- [10] Smith, E.R. (1982). Beliefs attributions, and evaluations: nonhierarchical models of mediation in social cognition, *Journal of Personality and Social Psychology* **43**, 348–259.
- [11] Sobel, M.E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models, in *Sociological Methodology 1982*, S. Leinhardt, ed., American Sociological Association, Washington, pp. 290–312.

(See also **Structural Equation Modeling: Multilevel; Structural Equation Modeling: Nonstandard Cases**)

DAVID A. KENNY

Mendelian Genetics Rediscovered

DAVID DICKINS

Volume 3, pp. 1198–1204

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mendelian Genetics Rediscovered

Introductory Remarks

The schoolchild's notion of the 'rediscovery of Mendel' is that

1. Mendel was the first person to make systematic studies of heredity.
2. He used large enough numbers and mathematical skill to get firm results from which his famous laws could be derived.
3. Because Mendel worked in a monastery in Moravia and published in an obscure journal, Darwin (whose own understanding of genetics was incorrect, thereby leaving a logical flaw in his theory of evolution) never got to discover this work.
4. By coincidence, three scientists, De Vries in Holland, Correns in Germany, and Tschermak in Austria, all independently arrived at the same laws of heredity in 1900, and only then realized Mendel had achieved this 34 years earlier.
5. Rediscovery of these laws paved the way for a complete acceptance of Darwin's theory, with natural selection as the primary mechanism of evolutionary change.

All of these need correction.

Mendel was the First Person to Make Systematic Studies of Heredity

The first of such experiments should probably be associated with the name of Joseph Gottlieb Koelreuter, a German botanist who made seminal contributions to plant hybridization [22] a century before Mendel. Cyrill Napp, a previous abbot of the monastery, of which Mendel himself eventually became abbot, through contact with sheep-breeders, realized the economic importance to the locality of understanding heredity, and sent Mendel to the University of Vienna to be trained in experimental science¹ [21].

In Vienna, Mendel received a very good scientific education in biology and especially in Physics: he had courses from Doppler and von Ettinghausen, and

the latter, in particular, gave him the methodological sophistication to deliberately employ a hypothetico-deductive approach. His Botany professor was Franz Unger, who was a forerunner of Darwin, and who published an evolutionary theory in an *Attempt of a History of the Plant World* in 1852. Mendel reported it was Unger's ponderings on how new species arose from old that prompted him to start his experiments. As the title of his famous paper [13, 14] suggests, rather than seeking some general laws of heredity, Mendel was actually trying to study hybridization, in the tradition, started by Linnaeus, that this might lead to the formation of new species. This was not a mechanism for substantive, macroevolutionary change, and some [5] have doubted whether Mendel actually believed in evolution as an explanation for the entirety of living organisms. He had a copy, annotated by him, of the sixth edition of the *Origin of Species* [2], (which is preserved in the museum in Brno), but certainly Darwin's idea of natural selection being the main mechanism of evolutionary change did not drive Mendel to search for appropriate units of heredity on which this process could work.

Using the garden pea (*Pisum sativum*), Mendel sought contrasting pairs of definitive characteristics, producing a list of seven, namely:

1. ripe seeds smooth or wrinkled
2. cotyledon yellow or green
3. seed coat white or grey
4. ripe pod smooth and not constricted anywhere, or wrinkled and constricted between the seeds;
5. unripe pod green or vivid yellow
6. lots of flowers along the main stem, or only at the end
7. stem tall (6–7 ft) or short (9–18 in.).

When members of these contrasting pairs were crossed, he found a 3:1 ratio in their offspring for that pair of characters. When they were then self-pollinated, the less frequent variant, and one third of the more frequent variant, bred true (gave only offspring the same as themselves), whereas the remaining two thirds of the more frequent variant produced a similar 3:1 ratio. This led him to postulate (or confirmed his postulation), that some hypothetical entity responsible for influencing development in either one way or the other must exist as two copies in a particular plant, but as only one copy in a pollen grain or ovum. If the two copies were of

2 Mendelian Genetics Rediscovered

opposite kind, one kind would always be dominant, and the other recessive, so that the former character would always be expressed in a hybrid², while the other would lie dormant, yet be passed on to offspring. These were only conceptual ‘atoms’, probably not considered by Mendel to be material entities that might one day be chemically isolated. All of this was prior to the discovery of chromosomes, the individual character and continuity of which was suspected, especially by Boveri, in the 1890s, but only finally shown by Montgomery and Sutton in 1901 and 1902.

The three so-called laws of Mendel were neither proposed by him, nor were they really laws. Independent segregation relates to the really important idea that a character which can exist (as we would now say *phenotypically* in two forms can be due to two factors, one inherited from one parent, and one from the other. That such factors (corresponding only with the wisdom of hindsight to today’s genes) may exist, and that there may be two or more alternative forms of them (today’s set of *alleles*) were the really important ideas of Mendel that his experiments revolutionarily enshrine. Independent assortment of one pair of factors in relation to some other such pair is not a law, since it depends upon the location of the different pairs on different chromosomes. Mendel was either extremely lucky to select seven pairs of characteristics presumably related to the seven different chromosomes of this species, a chance of 1/163, or he neglected noise in his data (in fact due to what we now know as *linkage*) or perhaps attributed it to a procedural error, and rejecting such data, and arrived at the ‘correct’ result by calculation based on prior assumptions. And the third ‘law’, of dominance and recessiveness, of course (as Mendel himself knew) only applies to certain pairs of alleles, such as those in which the recessive allele results in the inability to synthesize a particular enzyme, which may be manufactured in sufficient quantity in the presence of just one copy of the dominant allele [8]. In many cases the heterozygote may be intermediate between, or qualitatively different from either homozygote.

He used Large Enough Numbers and Mathematical Skill to Get Firm Results from which his famous Laws could be Derived

Certainly, the numbers of plants and crosses Mendel used were appropriately large for the probabilistic

nature of the task in hand. Mendel knew nothing of statistical proofs, but realized that large samples would be needed, and on the whole tested large numbers in order to produce reliable results. Mendel’s carefully cast three laws led with impeccable logic to the prediction of a 9:3:3:1 ratio for the independent segregation of two pairs of ‘factors’ (*Elemente*)³. Now, the actual ratios which Mendel seems to have claimed to have found, fit these predictions so well that they have been deemed (by **R. A. Fisher**, no less [7]) to be too good, that is very unlikely to have been obtained by chance, given the numbers of crosses Mendel carried out. Charitably, Jacob Bronowski [1] created the memorable image of Mendel’s brother friar ‘research assistants’ innocently counting out the relative numbers of types of peas so as to please Gregor, rather than Mendel himself inventing them. Another explanation might be that Mendel went on collecting data until the predicted ratios emerged, and then stopped, poor practice leading in effect to an unwitting rigging of results. Fisher made his criticism, which is by no means universally accepted [12, 19], in the context of a paper [7] that also patiently attempts to reconstruct what actual experiments Mendel may have done, and overall eulogizes unstintingly the scientific prestige of Mendel. The ethics do not matter⁴: the principles by means of which he set out to predict the expected outcomes are what Mendel is famous for, since we now know, for pairs of nonlethal alleles with loci on separate chromosomes, that these rules correctly describe the default situation. It is a long way from here, however, to the ‘modern synthesis’ [6, 9], and neither Mendel, nor initially ‘Mendelism’ (in the hands of the three ‘Rediscoverers’), were headed in that direction.

Because Mendel worked in a monastery in Moravia and published in an obscure journal, Darwin (whose own understanding of genetics was incorrect, thereby leaving a logical flaw in his theory of evolution) never got to discover this work

Gregor Mendel gave an oral paper, and later published his work in the proceedings of the local society of naturalists in Brünn, Austria (now Brno, Czech Republic), in 1866. The *Verhandlungen* of the Brünn Society were sent to the Royal Society and the Linnaean Society (*inter alia*), but there seems to be no

evidence for the notion that Darwin had a personal copy (uncut, and therefore unread).

Even if Darwin had come across Mendel's paper, the fact that it was in German and proposed ideas different from his own theory of *pangenesis* would have been sufficient, some have suggested, to discourage Darwin from further browsing. A more interesting (alternative history) question (put by Mayr [12], and answered by him in the negative) is whether Darwin would have appreciated the significance of Mendel's work, had he carefully read it.

Darwin was inclined towards a 'hard' view of inheritance (that it proceeded to determine the features of offspring with no relation to prevailing environmental conditions). He nevertheless made some allowance, increasingly through his life, for 'soft' inheritance, for example, in the effects of use and disuse on characters, such as the reduction or loss of eyes in cave-dwelling species. (This we would now attribute to the cessation of stabilizing selection, plus some possible benefit from jettisoning redundant machinery.) Such 'Lamarckian' effects were allowed for in Darwin's own genetic theory, his 'provisional hypothesis of pangenesis' [3], in which hereditary particles, which Darwin called *gemmules*, dispersed around the individual's body, where they would be subject to environmental influences, before returning to the reproductive organs to influence offspring. The clear distinction, made importantly by the great German biologist Weismann [23], between germ plasm and somatic cells, with the hereditary particles being passed down an unbroken line in the germ plasm, uninfluenced by what happened in the soma, thus banishing soft inheritance⁵, came too late for Darwin and Mendel. After it, the discoveries of Mendel became much easier to 'remake'.

Darwin always ranked natural selection as a more potent force in evolutionary change, however, and this was a gradual process working on the general variation in the population. Darwin was aware of 'sports' and 'freaks', and thus in a sense had the notion of mutations (which De Vries – see subsequent passage – made the basis of his theory of evolution), but these were not important for Darwin's theory, and he might have seen the explicit discontinuities of Mendel's examples as exceptions to general population variation. The work of Fisher and others in population genetics, which explains the latter in terms of the former, thereby reconciling Mendelism

with natural selection, had yet to be done: this provided the key to the so-called 'Modern Synthesis' [9] of the 1940s.

We have seen that Darwin, with the gemmules of his pangenesis theory, had himself postulated a particulate theory of heredity. However, he, together with Galton, Weismann, and De Vries, in their particulate theories of inheritance, postulated the existence of multiple identical elements for a given character in each cell nucleus, including the germ cells. This would not have led to easily predictable ratios. For Mendel, only one *Elemente* would enter a gamete. In the heterozygote, the unlike elements (*differierenden Elemente*) would remain separate, though Mendel for some reason assumed that blending would occur (in the homozygote) if the two elements were the same (*gleichartigen Elemente*). So perhaps Darwin would not have seen Mendel's work as the answer to Jenkin's devastating criticism [10] of his theory, based on the notion of blending inheritance, that after a generation or two, the effects of natural selection would be diluted to nothing. This, of course, is the statistical principle of **regression to the mean** in a Gaussian distribution, which Galton, Darwin's cousin, later emphasized. Galton could not see how the average of a population could shift through selection, and looked to genetic factors peculiar to those at the tail ends of distributions to provide a source of change, as in positive eugenic plans to encourage those possessed of 'hereditary genius' to carefully marry those with similar benefits.

Of the 40 reprints of [14], Mendel is known to have sent copies to two famous botanists, A. Kerner von Marilaun at Innsbruck, and Nägeli, a Swiss, but by that time professor of botany in Munich, and famous for his own purely speculative theory of inheritance in conformity with current physics (a kind of String theory of its day!). Now Nägeli was a dampening influence upon Mendel (we only have Mendel's letters to Nägeli, not vice versa). Nägeli did not cite Mendel in his influential book [17], nor did he encourage Mendel to publish his voluminous breeding experiments with *Pisum* and his assiduous confirmations in other species. Had Mendel ignored this advice the meme of his work would probably not have lain dormant for a generation, nor would it have needed to be 'rediscovered'. Nägeli simply urged Mendel to test his theory on hawkweeds (*Hieracium*), which he did, with negative results: we now know parthenogenesis is common in this genus. With his

4 Mendelian Genetics Rediscovered

training in physics (a discipline famous for vaulting general laws, perhaps bereft of the only Golden Rule of Biology – like that of the Revolutionist in George Bernard Shaw’s *Handbook* – which is that there is no Golden Rule), Mendel asserted that ‘A final decision can be reached only when the results of detailed experiments from the most diverse plant families are available.’ The *Hieracium* work he did publish in 1870: it was to be his only other paper, before he ascended his institution’s hierarchy and was lost to administration the following year.

By Coincidence Three Scientists, De Vries in Holland, Correns in Germany, and Tschermak in Austria, all Independently Arrived in 1900 at the Same Laws of Heredity and only then Realized Mendel had done this 34 Years Earlier

At the turn of the twentieth century, Correns and Tschermak were both working on the garden pea also, and found references to Mendel’s paper during their literature searches. Most important was its citation in Focke’s review of plant hybridization, *Die Pflanzen-Mischlinge* (1881), (though Focke did not understand Mendel’s work), and is the only one of the 15 actual citations that is relevant to its content that does little to stimulate the reader to consult the original paper⁶. De Vries also found the reference. It would be like harking back to a 1970 paper now, but in a strikingly smaller corpus of publications.

Hugo (Marie) de Vries was a Dutchman, in his early fifties in 1900, and Professor of Botany at the University of Amsterdam. In 1886, he became interested in the many differences between wild and cultivated varieties of the Evening Primrose (*Oenothera lamarckiana*), and the sudden appearance, apparently at random, of new forms or varieties when he cultivated this plant, for which he coined the term ‘mutations’. Their sudden appearance seemed to him to contradict the slow, small variations of Darwin’s natural selection, and he thought that it was through mutations that new species were formed, so that their study provided an experimental way of understanding the mechanism of evolution.

De Vries had already proposed a particulate theory of inheritance in his *Intracellular Pangenesis* (1889) [4], and in 1892 began a program of breeding experiments on unit characters in several plant

species. With large samples, he obtained clear segregations in more than 30 different species and varieties, and felt justified in publishing a general law. In a footnote to one of three papers given at meetings and quickly published in 1900, he stated that he had only learned of the existence of Mendel’s paper⁷ after he had completed most of these experiments and deduced from his own results the statements made in this text. Though De Vries carefully attributed originality to Mendel in all his subsequent publications, scholars [5] have argued about the veracity of this assertion. Some of his data, based on hundreds of crosses, we would now see as quite good approximations to 3 : 1 ratios in F2 crosses, but in his writings around this time, De Vries talked of 2 : 1 or 4 : 1 ratios, or quoted percentage splits such as 77.5% : 22.5% or 75.5% : 24.5%. Mayr [12] opines that it cannot be determined whether or not De Vries had given up his original theory of multiple particles determining characteristics in favor of Mendel’s single elements. Perhaps through palpable disappointment at having been anticipated by Mendel, he did not follow through to the full implications of Mendel’s results as indicative of a general genetic principle, seeing it as only one of several mechanisms, even asserting to Bateson that ‘Mendelism is an exception to the general rule of crossing.’

De Vries became more concerned about his mutation theory of evolution (*Die Mutationstheorie*, 1901); he was the first to develop the concept of mutability of hereditary units. However, rather than the point gene mutations later studied by Morgan, De Vries’ mutations were mostly chromosome rearrangements resulting from the highly atypical features that happen to occur in *Oenothera*, the single genus in which De Vries described them.

Carl Erich Correns, in his mid-thirties in 1900, was a German botany instructor at the University of Tübingen, was also working with garden peas, and similarly claimed to have achieved a sudden insight into Mendelian segregation in October 1899. Being busy with other work, he only read Mendel a few weeks later, and he only rapidly wrote up his own results after he had seen a reprint of one of the papers of De Vries in 1900. He readily acknowledged Mendel’s priority, and thought his own rediscovery had been much easier, given that he, unlike Mendel, was following the work of Weismann. Correns went on to produce a lot more supportive evidence throughout his career in a number of German

universities, and he postulated a physical coupling of genetic factors to account for the consistent inheritance of certain characters together, anticipating the development of the concept of linkage by the American geneticist Thomas Hunt Morgan.

Though his 1900 paper⁸ showed no understanding of the basic principles of Mendelian inheritance [15, 16], the third man usually mentioned as a rediscoverer of Mendel is the Austrian botanist Erich Tschermak von Seysenegg, not quite 30 in 1900. Two years before, he had begun breeding experiments, also on the garden pea, in the Botanical Garden of Ghent, and the following year he continued these in a private garden, whilst doing voluntary work at the Imperial Family's Foundation at Esslingen near Vienna. Again, it was while writing up his results that he found a reference to Mendel's work that he was able to access in the University of Vienna. It duplicated, and in some ways went beyond his own work. In his later work, he took up a position in the Academy of Agriculture in Vienna in 1901 and was made a professor five years later. He applied Mendel's principles to the development of new plants such as a new strain of barley, an improved oat hybrid, and hybrids between wheat and rye.

It is really William Bateson (1861–1926) who should be included as the most influential of those who were the first to react appropriately to the small opus by Mendel lurking in the literature, even though his awareness of it was due initially to De Vries. Bateson was the first proponent of Mendelism in English-speaking countries, and coined the term 'genetics' (1906). Immediately after the appearance of papers by De Vries, Correns and Tschermak, Bateson reported on these to the Royal Horticultural Society in May 1900, and then read, and was inspired by, Mendel's paper *Experiments in Plant Hybridisation* shortly thereafter. Bateson was also responsible for the first English translation of this paper, which was published in the *Journal of the Royal Horticultural Society* in 1901.

Bateson's is a fascinating case of a brilliant and enthusiastic scientist, who turned against his earlier Darwinian insights, and adopted, with considerable grasp, the new principles of Mendelism. Paradoxically, this served to obstruct the eventual realization that these were in fact the way to vindicate and finally establish the overarching

explanatory power of Darwin's evolution by natural selection.

Initially, Bateson shared the interests of his friend Weldon, who had been influenced by Pearson's statistical approach to study continuous variation in populations, but when he himself came to do this in the field, choosing fishes in remote Russian lakes, he strengthened his conviction that natural selection could not produce evolutionary change from the continuum of small variations he discovered. He urgently sought some particulate theory of heredity to account for the larger changes he thought necessary to drive evolution, and thought he had found this in Mendelism. He was stimulated by De Vries's notion that species formation occurred when mutations spread in a population, and when he later read a 1900 paper by De Vries describing the 3:1 ratio results (in the train on his way to a meeting of the Royal Horticultural Society), he changed the text of his paper to claim that Galton's regression law was in need of amendment. Bateson became an insightful and influential exponent of the new Mendelism, carrying out confirmatory experiments in animals, and coining terms (in addition to 'genetics') which soon became key conceptual tools, including 'allelomorph' (later shortened to 'allele'), 'homozygote' and 'heterozygote'.

The term 'gene', however, was coined by the Danish botanist William Johannsen. Unfortunately, he sought to define each gene in terms of the specific phenotypic character for which it was deemed responsible. The distinction between 'genotype' and 'phenotype' was drawn early by **G. Udny Yule** in Cambridge. Yule, unlike Johannsen, insightfully saw that if many genes combined to influence a particular character, and if this character were a relatively simple unit, which a phenotypic feature arbitrarily seized upon an experimenter might well not be, then Mendelism might deliver the many tiny changes beloved of Darwin. This was an idea whose time was yet to come.

Rediscovery of these Laws Paved the way for a Complete Acceptance of Darwin's Theory with Natural Selection as the Primary Mechanism of Evolutionary Change

In the end it did, but not until after a long period of misunderstanding and dispute, which was finally brought to an end by means of the Modern Synthesis.

As we have seen, there were two main sides to the dispute. On the one hand were the so-called biometricians, led by **Karl Pearson**, who defended Darwinian natural selection as the major cause of evolution through the cumulative effects of small, continuous, individual variations (which they assumed passed from one generation to the next without being limited by Mendel's laws of inheritance). On the other were the champions of 'Mendelism', who were mostly also mutationists: they felt that if we understood how heredity worked to produce big changes, then *ipso facto* we would have an explanation for the surges of the evolutionary process.

The resolution of the dispute was brought about in the 1920s and 1930s by the mathematical work of the population geneticists, particularly J. B. S. Haldane and R. A. Fisher in Britain, and Sewall Wright in the United States. They showed that continuous variation was entirely explicable in terms of Mendel's laws, and that natural selection could act on these small variations to produce major evolutionary changes. Mutationism became discredited. Recent squabbles about the relative size and continuity of evolutionary changes, between those tilted towards relative, rapid, quite large changes 'punctuating' long periods of unchanging equilibrium, and those in favor of smaller and more continuous change, are but a pale reflection of the earlier dispute.

In its original German, Mendel's paper repetitively used the term *Entwicklung*, the nearest English equivalent to which is probably 'development' [20]. It was also combined, German-fashion, with other words, as in *Entwicklungsreihe*, (developmental series); *die Entwicklungsgeschichte* (the history of development); *das Entwicklungs-Gesetz* (the law of development). Mendel at least (if not all of the 'rediscoverers' and their colleagues who thought they were addressing the same problems as Mendel), would have been delighted by the present day rise of developmental genetics, that is, the study of how the genetic material actually shapes the growth and differentiation of the organism, made possible, among other advances, by the discovery of the genetic code and the mapping of more and more genomes. He might have been surprised too, by the explanatory power of theories derived from population genetics, based on his own original work, combined with natural selection, to account for the evolution of behavior, and of human nature.

Notes

1. According to the presentation by Roger Wood (Manchester University, UK), in a joint paper with Vitezslav Orel (Mendelianum, Brno, Czech Republic) at a conference in 2000, at the Academy of Sciences in Paris, to mark the centenary of the conference of the Academy to which De Vries reported.
2. Mendel was not clear about the status of species versus varieties (not an easy issue today) and used the term hybrid for crosses at either level, whereas we would confine it to species crosses today.
3. With the first two pairs of characters listed above Mendel actually seems to have obtained 315 round yellow: 108 round green: 101 wrinkled yellow: 32 wrinkled green.
4. As the noted historian of psychology Leslie Hearnshaw once suggested to me, the fact that Mendel went on to publish his subsequent failures to obtain similar results in the Hawkweed (*Hieracium*) does much to allay any suspicions as to his honesty.
5. Weismann was right in principle, but instead of the cytological distinction, we have today the 'central dogma' of biochemical genetics, that information can pass from DNA to proteins, and from DNA to DNA, but not from protein to DNA [11].
6. 'Mendel's numerous crossings gave results which were quite similar to those of Knight, but Mendel believed that he found constant numerical relationships between the types of the crosses.' (Focke 1881, quoted by Fisher [7].)
7. In his book entitled *Origins of Mendelism* [18], the historian Robert Olby relates that just prior to publication 'from his friend Professor Beijerinck in Delft [De Vries] received a reprint of Mendel's paper with the comment: "I know that you are studying hybrids, so perhaps the enclosed reprint of the year 1865 by a certain Mendel is still of some interest to you"'.
8. <http://www.esp.org/foundations/genetics/classical/holdings/t/et-00.pdf>

References

- [1] Bronowski, J. (1973). *The Ascent of Man*, British Broadcasting Corporation Consumer Publishing, London.
- [2] Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, 1st Edition, John Murray, London.

-
- [3] Darwin, C. (1868). *The Variation of Plants and Animals Under Domestication*, 1–2 John Murray, London.
 - [4] De Vries, H. (1889). *Intracelluläre Pangenesis*, Gustav Fischer, Jena.
 - [5] Depew, D.J. & Weber, B.H. (1996). *Darwinism Evolving: Systems Dynamics and the Genealogy of Natural Selection*, The MIT Press, Cambridge & London, England.
 - [6] Dobzhansky, T. (1937). *Genetics and the Origin of Species*, Columbia University Press, New York.
 - [7] Fisher, R.A. (1936). Has Mendel's work been rediscovered? *Annals of Science* **1**, 115–137.
 - [8] Garrod, A.E. (1909). *Inborn Errors of Metabolism*, Frowde & Hodder, London.
 - [9] Huxley, J. (1942). *Evolution, the Modern Synthesis*, Allen & Unwin, London.
 - [10] Jenkin, F. (1867). The origin of species, *North British Review* **45**, 277–318.
 - [11] Maynard Smith, J. (1989). *Evolutionary Genetics*, Oxford University Press, Oxford.
 - [12] Mayr, E. (1985). *The Growth of Biological thought: Diversity, Evolution, and Inheritance*, Harvard University Press, Cambridge.
 - [13] Mendel, J.G. (1865). *Experiments in Plant Hybridisation*. http://www.laskerfoundation.org/news/gnn/timeline/mendel_1913.pdf
 - [14] Mendel, J.G. (1865). Versuche über Pflanzen-hybriden, *Verhandlungen des Naturforschenden Vereines in Brunn* **4**, 3–47.
 - [15] Monaghan, F. & Corcos, A. (1986). Tschermak: a nondiscoverer of mendelism. I. An historical note, *Journal of Heredity* **77**, 468–469.
 - [16] Monaghan, F. & Corcos, A. (1987). Tschermak: a non-discoverer of mendelism. II. A critique, *Journal of Heredity* **78**, 208–210.
 - [17] Nägeli, C. (1884). *Mechanisch-Physiologische Theorie der Abstammungslehre*, Oldenbourg, Leipzig.
 - [18] Olby, R.C. (1966). *Origins of Mendelism*, Schocken Books, New York.
 - [19] Pilgrim, I. (1986). A solution to the too-good-to-be-true Paradox and Gregor Mendel, *The Journal of Heredity* **77**, 218–220.
 - [20] Sandler, I. (2000). Development: Mendel's legacy to genetics, *Genetics* **154**(1), 7–11.
 - [21] Veuille, M. (2000). 1900–2000: How the mendelian revolution came about: the Rediscovery of Mendel's Laws (1900), *International Conference, Paris, 23–25 March 2000; Trends in Genetics* **16**(9), 380.
 - [22] Waller, J. (2002). *Fabulous Science: Fact and Fiction in the History of Scientific Discovery*, Oxford University Press, Oxford.
 - [23] Weismann, A. (1892). *Das Keimplasma: Eine Theorie der Vererbung*, Gustav Fischer, Jena.

DAVID DICKINS

Mendelian Inheritance and Segregation Analysis

P. SHAM

Volume 3, pp. 1205–1206

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mendelian Inheritance and Segregation Analysis

Gregor Mendel (1822–1884), Abbot at the St Thomas Monastery of the Augustinian Order in Brunn, conducted the seminal experiments that demonstrated the existence of genes and characterized how they are transmitted from parents to offspring, thus laying the foundation of the science of genetics. Mendel chose the garden pea as his experiment organism, and selected seven characteristics that are dichotomous and therefore easy to measure. Mendel's experiments, and their results, can be summarized as follows:

1. After repeated inbreeding, plants became uniform in each characteristic (e.g., all tall). These are known as pure lines.
2. When two pure lines with opposite characteristics are crossed (e.g., tall and short), one of the characteristics is present in all the offspring (e.g., they are all tall). The offspring are said to be the F1 generation, and the characteristic present in F1 (e.g., tall) is said to be dominant, while the alternative, absent characteristic (e.g., short) is said to be recessive.
3. When two plants of the F1 generation are crossed, the offspring (F2) display the dominant (e.g., tall) and recessive (e.g., short) characteristics in the ratio 3:1. This cross is called an *intercross*.
4. When an F1 plant is crossed with the parental recessive pure line, the offspring display the dominant and recessive characteristics in the ratio 1:1. This cross is called a *backcross*.

Mendel explained these observations by formulating the law of segregation. This law states that each individual contains two inherited factors (or genes) for each pair of characteristics, and that during reproduction one of these two factors is transmitted to the offspring, each with 50% probability. There are two alternative forms of the genes, called *alleles*, corresponding to each dichotomous character. When the two genes in an individual are of the same allele, then the individual's genotype is said to be homozygous, otherwise it is said to be heterozygous. An individual with heterozygous genotyping (e.g., Aa) has the same characteristic as an individual with homozygous

genotype (e.g., AA) of the dominant allele (e.g., A). The explanations of the above observations are then as follows:

1. Repeated inbreeding produces homozygous lines (e.g., AA, aa) that will always display the same characteristic in successive generations.
2. When two pure lines (AA and aa) are crossed, the offspring (F1) will all be heterozygous (Aa), and therefore have the same characteristic as the homozygous of the dominant allele (A).
3. When two F1 (Aa) individuals are crossed, the gametes A and a from one parent combine at random with the gametes A and a from the other to form the offspring genotypes AA, Aa, aA, and aa, in equal numbers. The ratio of offspring with dominant and recessive characteristics is therefore 3:1.
4. When an F1 (Aa) individual is crossed with the recessive (aa) pure line, the offspring will be 50:50 mixture of Aa and aa genotypes, so that the ratio of offspring with dominant and recessive characteristics is therefore 1:1.

The characteristic 3:1 and 1:1 ratios among the offspring of intercross and backcross, respectively, are known as Mendelian segregation ratios. Mendel's work was the first demonstration of the existence of discrete heritable factors, although the significance of this was overlooked for many years, until 1900 when three botanists, separately and independently, rediscovered the same principles.

Classical Segregation Analysis

The characteristic 1:1 and 3:1 segregation ratios provide a method of checking whether a disease in humans is caused by mutations at a single gene. For rare dominant disease, the disease mutation (A) is likely to be rare, so that individuals homozygous for the disease mutations (AA) are likely to be exceedingly rare. A mating between affected and unaffected individuals is therefore very likely to be a backcross (Aa × aa), with a predicted segregation ratio of 1:1 among the offspring. An investigation of such matings to test the segregation ratio of offspring against the hypothetical value of 1:1 is a form of classical segregation analysis.

For rare recessive conditions, the most informative mating is the intercross (Aa × Aa) with a predicted

2 Mendelian Inheritance and Segregation Analysis

segregation ratio of 3:1. However, because the condition is recessive, the parents are both normal and are indistinguishable from other mating types (e.g., $aa \times aa$, $Aa \times aa$). It is therefore necessary to recognize $Aa \times Aa$ matings from the fact that an offspring is affected. This, however, introduces an ascertainment bias to the “apparent segregation ratio”. To take an extreme case, if all families in the community have only one offspring, then ascertaining families with at least one affected offspring will result in all offspring being affected. A number of methods have been developed to take ascertainment procedure into account when conducting segregation analysis of putative recessive disorders. These include the proband method, the singles method, and maximum likelihood methods.

Complex Segregation Analysis

Complex segregation analysis is concerned with the detection of a gene that has a major impact on the phenotype (called a *major locus*), even though it is not the only influence and that other genetic and environmental factors are involved in determining the phenotype. The involvement of other factors reduces the strength of relationship between the major locus and the phenotype. Because of this, it is not possible to deduce the underlying mating type from the phenotypes of family members. Instead, it is necessary to consider all the possible mating types for each family. Even if a specific mating type could be isolated, the segregation ratio would not be expected to follow classical Mendelian ratios. For

these reasons, this form of segregation analysis is said to be complex.

Complex segregation analysis is usually conducted using **maximum likelihood methods** under one of two models. The first is a generalized single locus model with generalized transmission parameters. This differs from a Mendelian model in two ways. First, the probabilities of disease given **genotype** (called *penetrances*) are not necessarily 0 or 1, but can take intermediate values. Secondly, the probabilities of transmitting an allele (e.g., A) given parental genotype (e.g., AA, Aa, and aa) are not necessarily 1, 1/2, or 0, but can take other values. A test for a major locus is provided by a test of whether these transmission probabilities conform to the Mendelian values (1, 1/2, and 0). The second model is the mixed model, which contains a single major locus against a polygenic background. A test for a major locus is provided by a test of this mixed model against a pure polygenic model without a major locus component. Both forms of analyses can be applied to both qualitative (e.g., disease) and quantitative traits, and are usually conducted in a maximum likelihood framework, with adjustment for the ascertainment procedure. The generalized transmission test and the mixed model test have been combined into a unified model and implemented in the POINTER program. Other methods for complex segregation analysis, for example, using regressive models, have also been developed.

P. SHAM

Meta-Analysis

WIM VAN DEN NOORTGATE AND PATRICK ONGHENA

Volume 3, pp. 1206–1217

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Meta-Analysis

Introduction

Small Progress in Behavioral Science?

It is well known that the results of empirical studies are subject to random fluctuation if they are based on samples of subjects instead of on the complete population on which the study is focused. In a study evaluating the effect of a specific treatment, for instance, the population **effect size** is typically estimated using the effect size that is observed in the sample. Traditionally, researchers deal with the uncertainty associated with the estimates by performing significance tests or by constructing **confidence intervals** around the effect size estimates. In both procedures, one refers implicitly to the sampling distribution of the effect sizes, which is the distribution of the observed effect sizes if the study had been replicated an infinite number of times.

Unfortunately, research in behavioral science is characterized by relatively small-sample studies, small population effects and large initial differences between subjects [49, 58]. Consequently, confidence intervals around effect sizes are often unsatisfactorily large and thus relatively uninformative, and the **power** of the significance tests is small. Researchers, at least those who are aware of the considerable uncertainty about the study results, therefore often conclude their research report by a call for more research.

During the last decennia, several topics were indeed investigated several times, some of them even dozens of times. Mosteller and Colditz [44] talk about an information explosion. Yet, reviewers of the results of studies evaluating a similar treatment often are disappointed. While in some studies positive effects are found, in other studies, no effects or even negative effects are obtained. This is not only true for sets of studies that differ from each other in the characteristics of the subjects that were investigated or in the way the independent variable is manipulated, but also for sets of studies that are more or less replications of each other. Reviewers have a hard job to see the wood for the trees and often fall back on personal strategies to summarize the results of a set of studies. Different reviewers therefore often come to different conclusions, even if they discuss the

same set of studies [49, 58]. The conflicting results from empirical studies have brought some to the pessimistic idea that researchers in these domains do not progress, and have driven some politicians and practitioners toward relying on their own feelings instead of on scientific results. The rise of the meta-analysis offered new perspectives.

The Meta-analytic Revolution

Since the beginning of the twentieth century, there have been some modest attempts to summarize study results in a quantitative and objective way (e.g., [48] and [29]; see [59]). It was, however, not before the appearance of an article from Glass in 1976 [23] that the idea of a quantitative integration of study results was explicitly described. Glass coined the term ‘meta-analysis’ and defined it as

‘... the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings’. (p. 3)

The introduction of the term meta-analysis and the description of simple meta-analytic procedures were the start of a spectacular rise of the popularity of quantitative research synthesis. Fiske [17] even used the term meta-analytic revolution. Besides the popularity of meta-analysis in education [41], applications showed up in a wide variety of research domains [46]. The quantitative approach gradually supplemented or even pushed out the traditional narrative reviews of research literature [35]. Another indication of the growing popularity of meta-analysis was the mounting number and size of meta-analytic handbooks [1].

The growing attention for meta-analysis gradually affected the views on research progress in behavioral science. Researchers became aware that even if studies are set up in a very similar way, conflicting results due to random fluctuation alone are not surprising, especially when study results are described on the significant–nonsignificant dichotomy. By quantitatively combining the results of several similar trials, meta-analyses have the potential of averaging out the influence of random fluctuation, resulting in more steady estimates of the overall effect and a higher power in testing this overall effect size. Meta-analytic methods were developed in order to distinguish the random variation in study results from ‘true’ between

study heterogeneity, and in order to explain the latter by estimating and testing moderating effects of study characteristics.

Performing a Meta-analysis

A meta-analysis consists of several steps (see e.g., [9] for an extensive discussion):

1. *Formulating the research questions or hypotheses and defining selection criteria.* Just like in primary research, a clear formulation of the research question is crucial for meta-analytic research. Together with practical considerations, the research question results in a set of criteria that are used to select studies. The most common selection criteria relate to the population from which study participants are sampled, the dependent and the independent variables and their indicators, and the quality of the study. Note that formulating a very specific research question and using very strict selection criteria eventually results in a set of similar studies for which the results of the meta-analysis are relatively easy to interpret, but at the other side of the coin, the set of studies will also be relatively small, at the expense of the reliability of the results.

2. *Looking for studies investigating these questions.* A thorough search includes an exploration of journals, books, doctoral dissertations, published or unpublished research reports, conference papers, and so on. Sources of information are, for instance, databases that are printed or available online or on CD-ROM, contacts with experts in the research domain, and reference lists of relevant material. To avoid bias in the meta-analytic results (see below), it is a good idea to use different sources of information and to include published as well as unpublished study results. Studies are selected on the basis of selection criteria from the first step.

3. *Extracting relevant data.* Study outcomes and characteristics are selected, coded for each study, and assembled in a database. The study characteristics can be used later in order to account for possible heterogeneity in study outcomes.

4. *Converting study results to a comparable measure.* Study results are generally reported by means of descriptive statistics (means, standard deviations, etc.), test statistics (t , F , χ^2 , ...) or P values, but the way of reporting is usually very different from

study to study. Moreover, variables are typically not measured on the same scale in all studies. The comparison of the study outcomes, therefore, requires a conversion to a common standardized measure. One possibility is to use P values. The meaning of P values does not depend on the way the variables are measured, or on the statistical test that is used in the study. A disadvantage of P values, however, is that it depends not only on the effect that is observed but also on the sample size. A very small difference between two groups, for instance, can be statistically significant if the groups are large, while a large difference can be statistically nonsignificant if the groups are relatively small. Although techniques for combining P values have been described, in current meta-analyses, usually measures are combined that express the magnitude of the observed relation, independently of the sample size and the measurement scale. Examples of such standardized effect size measures are the **Pearson's correlation coefficient** and the **odds ratio**. A popular effect size measure in behavioral science is the standardized mean difference, used to express the difference between the means of an experimental and a control condition on a continuous variable, or more generally the difference between the means of two groups:

$$\delta = \frac{\mu_E - \mu_C}{\sigma}, \quad (1)$$

with μ_E and μ_C equal to the population mean under the experimental and the control condition, respectively, and σ equal to the common population standard deviation. The population effect size δ is estimated by its sample counterpart:

$$d = \frac{\bar{x}_E - \bar{x}_C}{s_p}, \quad (2)$$

with $\bar{x}_E$ and $\bar{x}_C$ equal to the sample means, and s_p equal to the pooled sample standard deviation. Assuming normal population distributions with a common variance under both conditions, the sampling distribution of d is approximately normal with mean δ and variance equal to [30]:

$$\hat{\sigma}_d^2 = \frac{n_E + n_C}{n_E n_C} + \frac{d^2}{2(n_E + n_C)} \quad (3)$$

5. *Combining and/or comparing these results and possibly looking for moderator variables.* This step

forms the essence of the meta-analysis. The effect sizes from the studies are analyzed statistically. The unknown parameters of one or more meta-analytic models are estimated and tested. Common statistical models and techniques to combine or compare effect sizes will be illustrated extensively below by means of an example.

6. Interpreting and reporting the results. In this phase, the researcher returns to the research question(s) and tries to answer these questions based on the results of the analysis. The research report ideally describes explicitly the research questions, the inclusion criteria, the sources used in the search for studies, a list of the studies that were selected, the observed effect sizes, the study sample sizes, and the most important study characteristics. It should also contain a description of how study results were converted to a common measure, which models, techniques, estimation procedures, and software were used to combine these measures, and which assumptions were made for the analyses. It is often a good idea to illustrate the meta-analytic results graphically, for instance, using a funnel plot, a **stem-and-leaf plot**, a **histogram** or a plot of the interval estimates of the observed study effect sizes and the overall effect size estimate (see below).

An Example

Raudenbush [50, 51] combined the results of 18 experiments investigating the effect of teacher expectations on the intellectual development of their pupils. In each of the studies, the researcher tried to create high expectancies for an experimental group of pupils, while this was not the case for a control group. The observed differences between groups were converted to the standardized mean difference (Table 1). Table 1 also describes for each study the length of contact between teacher and pupil prior to the expectancy induction, as well as the standard error of the observed effect sizes (3).

Since the sampling distribution of the standardized mean difference is approximately normal with a standard deviation estimated by the standard error of estimation, one could easily construct confidence intervals around the observed effect sizes. The confidence intervals are presented in Figure 1.

Note that somewhat more than half of the observed effect sizes are larger than zero. In three studies, zero is not included in the confidence interval, and the (positive) observed effect sizes therefore are statistically significantly different from zero. Following Cohen [6], calling a standardized mean difference of

Table 1 Summary results of experiments assessing the effect of teacher expectancy on pupil IQ, reproduced from Raudenbush and Bryk [51], with permission of AERA

	Study	Weeks of prior contact	Effect size	Standard error of effect size estimate
1	[55]	2	0.03	0.125
2	[7]	21	0.12	0.147
3	[36]	19	-0.14	0.167
4	[47]	0	1.18	0.373
5	[47]	0	0.26	0.369
6	[12]	3	-0.06	0.103
7	[15]	17	-0.02	0.103
8	[4]	24	-0.32	0.22
9	[38]	0	0.27	0.164
10	[42]	1	0.8	0.251
11	[3]	0	0.54	0.302
12	[19]	0	0.18	0.223
13	[37]	1	-0.02	0.289
14	[32]	2	0.23	0.29
15	[16]	17	-0.18	0.159
16	[28]	5	-0.06	0.167
17	[56]	1	0.3	0.139
18	[18]	2	0.07	0.094
19	[21]	7	-0.07	0.174

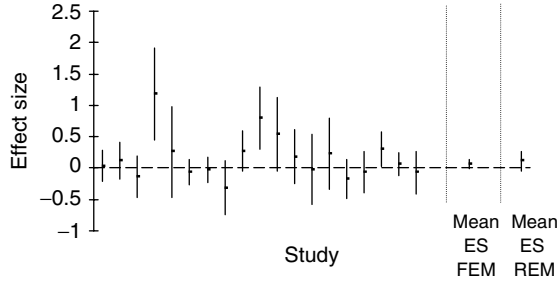


Figure 1 Confidence intervals for the observed effect sizes and for the estimated mean effect size under a fixed effects model (FEM) or a random effects model (REM)

0.20, 0.50, and 0.80 a small, moderate, and large effect respectively, almost all observed effect sizes suggest that there is generally only a small effect. With the naked eye, it is difficult to see whether there is a general effect of teacher expectancy, and how large this effect is. Moreover, it is difficult to see whether the observed differences between study outcomes are due to sampling variation, or whether the observed differences between study outcomes are due to the intrinsic differences between studies. In the following section, we show how these questions can be dealt with in meta-analysis. Different models are illustrated, which differ in complexity and in the underlying assumptions.

Fixed Effects Model

As discussed above, observed effect sizes can be conceived of as randomly fluctuating. If in a specific study another sample would have been taken (from the same population), the observed effect size would probably not have been exactly the same. Suppose that differences in the meta-analytic data set between the observed effect sizes can be entirely accounted for by sampling variation. In this case, we can model the study results as:

$$d_j = \delta + e_j, \quad (4)$$

with d_j the observed effect size in study j , δ the common population effect size, and e_j a residual due to sampling variation. The primary purpose of the meta-analysis is then to estimate or test the overall effect, δ . The overall effect is usually estimated by averaging the observed effects. Since, in general, the population effect size is estimated more reliably in

large studies, effect sizes are sometimes weighted by the study sample sizes in calculating the average. A similar approach, resulting in a decreased *MSE*, is to weight by the precision of the estimates, this is the inverse of the squared standard error. The precision is estimated as the inverse of the sampling variance estimate (as given in Equation 3).

$$\hat{\delta} = \frac{\sum_{j=1}^k w_j d_j}{\sum_{j=1}^k w_j},$$

$$\text{with } w_j = \frac{1}{\hat{\sigma}_{d_j}^2} \text{ and } k \text{ the total number of studies} \quad (5)$$

The precision of the estimate of the overall effect is the sum of the individual precisions. The standard error of the estimate of the overall effect is estimated by:

$$SE(\hat{\delta}) = (\text{precision})^{-1/2} = \left(\sum_{j=1}^k \frac{1}{\hat{\sigma}_{d_j}^2} \right)^{-1/2} \quad (6)$$

For the example, the estimate of the overall effect is 0.060, with a corresponding standard error of 0.036. Since the sampling distribution of the overall effect size estimate is again approximately normal, an approximate 95% confidence interval equals $[0.060 - 1.96 * 0.036; 0.060 + 1.96 * 0.036] = [-0.011; 0.131]$. Because zero is included in the confidence interval, we conclude that the overall effect size estimate is statistically not significant at the 0.05-level. This is equivalent to comparing the estimate divided by the standard error with a standard normal distribution, $z = 1.67$, $p = 0.09$. The confidence interval around the overall effect size estimate is also presented in Figure 1. It can be seen that the interval is much smaller than the confidence intervals of the individual studies, illustrating the increased precision or power when estimating or testing a common effect size by means of a meta-analysis.

The assumption that the population effect size is the same in all studies is frequently unlikely. Studies often differ in, for example, the operationalization of the dependent or the independent variable and the

population from which the study participants are sampled, and it is often unlikely that these study or population characteristics are unrelated to the effect that is investigated. Before using the fixed effects techniques, it is therefore recommended to test the homogeneity of the study results. A popular homogeneity test is the test of Cochran [5]. The test statistic

$$Q = \sum_{j=1}^k \frac{(d_j - \hat{\delta})^2}{\hat{\sigma}_{d_j}^2} \quad (7)$$

follows approximately a chi-square distribution with $(k - 1)$ degrees of freedom¹. For the example, the homogeneity test reveals that it is highly unlikely that such relatively large differences in observed effect sizes are entirely due to sampling variation, $\chi^2(18) = 35.83$, $p = 0.007$. The assumption of a common population effect size is relaxed in the following models.

For the sake of completeness, we want to remark that the fixed effects techniques (*see Fixed and Random Effects; Fixed Effect Models*) described above are actually not only appropriate when the population effect size is the same for all studies, but also when the population effect size differs from study to study but the researcher is interested only in the population effect sizes that are studied. In the last case, (5) is not used to estimate a common population effect size, but rather to estimate the mean of the population effect sizes studied in the studies included in the meta-analytic data set.

Random Effects Model

Cronbach [10] argued that the treatments that are investigated in a meta-analytic set of empirical studies often can be regarded as a random sample from a population of possible treatments. As a result, studies often do not investigate a common population effect size, but rather a distribution of population effect sizes, from which the population effect sizes from the studies in the data set represent only a random sample. While in the fixed effects model described above (4) the population effect size is assumed to be constant over studies, in the random effects model, the population effect size is assumed to be a stochastic variable (*see Fixed and Random Effects*). The population effect size that is estimated in a specific study is modeled as the mean of the population distribution of effect sizes plus a random residual [11,

31, 35]:

$$\delta_j = \gamma + u_j \quad (8)$$

An observed effect size deviates from this study-specific population effect size due to sampling variation:

$$d_j = \delta_j + e_j \quad (9)$$

Combining (8 and 9) results in:

$$d_j = \gamma + u_j + e_j \quad (10)$$

If d_j is an unbiased estimate of δ_j , the variance of the observed effect size is the sum of the variance of the population effect size and the variance of the observed effect sizes around these population effect sizes:

$$\sigma_{d_j}^2 = \sigma_{\delta}^2 + \sigma_{(d_j|\delta_j)}^2 = \sigma_u^2 + \sigma_e^2 \quad (11)$$

Although better estimators are available [70], an obvious estimate of the population variance of the effect sizes therefore is [31]:

$$\hat{\sigma}_{\delta}^2 = s_d^2 - \frac{\sum_{j=1}^k \hat{\sigma}_{d_j|\delta_j}^2}{k} \quad (12)$$

To estimate γ , the mean of the population distribution of effect sizes, one could again use the precision weighted average of the observed effect sizes (5). While in the fixed effects model the precision associated with an observed effect size equals the inverse of the sampling variance alone, the precision based on a random effects model equals the inverse of the sampling variance plus the population variance. Since this population variance is the same for all studies, assuming a random effects model instead of a fixed effects model has an equalizing effect on the weights.

For the example, the estimate of the variance in population effect sizes is 0.080, the estimate of the mean effect size 0.114. Assuming that the distribution of the true effect sizes is normal, 95% of the estimated population distribution of effect sizes therefore is located between -0.013 and 0.271. For the example, the equalizing effect of using a random effects model resulted in an estimate of the mean effect size that is larger than that for the fixed effects model (0.060), since in the example larger standard errors are in general associated with larger observed effect sizes.

The standard error of the estimate can again be estimated using (6), but in this case, (11) is used to

estimate the variance of the observed effect sizes. The standard error therefore will be larger than for the fixed effects model, which is not surprising: the mean effect size can be estimated more precisely if one assumes that in all studies exactly the same effect size is estimated. For the example, the standard error of the estimate of the mean effect size equals 0.079, resulting in a 95% confidence interval equal to $[-0.041; 0.269]$. Once more, the interval includes zero, which means that even if there is no real effect, it is not unlikely that a mean effect size estimate of 0.114 or a more extreme one is found, $z = 1.44$, $p = 0.15$.

Since the effect is assumed to depend on the study, the researcher may be interested in the effect in one or a few specific studies. Note that the observed effect sizes are in fact estimates of the study-specific population effect sizes. Alternatively, one could use the mean effect size estimate to estimate the effect in each single study. While the first kind of estimate seems reasonable if studies are large (and the observed effect sizes are more precise estimates of the population effect sizes), the second estimate is sensible if studies are much alike (this is if the between study variance is smaller). The *empirical Bayes* estimate of the effect in a certain study is an optimal combination of both kinds of estimates. In this combination, more weight will be given to the mean effect if studies are more similar and if the study is small. Hence, in these situations, the estimates are more ‘shrunk’ to the mean effect. Because of this property of shrinkage, empirical Bayes estimates are often called *shrinkage estimates*. Empirical Bayes estimates ‘borrow strength’ from the data from other studies: the *MSE* associated with the empirical Bayes estimates of the study effect sizes is, in general, smaller than the *MSE* of the observed effect sizes. For more details about empirical Bayes estimates, see, for example, [51] and [52].

In the random effects model, population effect sizes are assumed to be exchangeable. This means that it is assumed that there is no prior reason to believe that the true effect in a specific study is larger (or smaller) than in another study. This assumption is quite often too restrictive, since frequently study characteristics are known that can be assumed to have a systematic moderating effect. In Table 1, for instance, the number of weeks of prior contact between pupils and teachers is given for each study. The number of weeks of prior contact could be

supposed to affect the effect of manipulating the expectancy of teachers toward their pupils, since manipulating the expectancies is easier if teachers do not know the pupil yet. In the following models, the moderating effect of study characteristics is modeled explicitly.

Fixed Effects Regression Model

Several methods have been proposed to account for moderator variables. Hedges and Olkin [31], for example, proposed to use an adapted **analysis of variance** to explore the moderating effect of a categorical study characteristic. For a continuous study characteristic, they proposed to use a fixed effects regression model. In this model, study outcomes differ due to sampling variation and due to the effect of study characteristics:

$$d_j = \delta_j + e_j = \gamma_0 + \sum_{s=1}^S \gamma_s W_{sj} + e_j, \quad (13)$$

with W_{sj} equal to the value of study j on the study characteristic s , and S the total number of study characteristics included in the model. Note that the effect of a categorical study characteristic can also be modeled by means of such a fixed effects regression model, by means of dummy variables indicating the category the study belongs to. The fixed effects regression model simplifies to the fixed effects model described above in case the population effect sizes do not depend on the study characteristics.

Unknown parameters can be estimated using the weighted least squares procedure, weighting the observed effect sizes by their (estimated) precision as we did before for the fixed effects model (5). Details are given by Hedges and Olkin [31]. If, for the example, one study characteristic is included with levels 0, 1, 2, and 3 for respectively 0, 1, 2, and 3 or more weeks of prior contact, the estimate of the regression intercept equals 0.407, with a standard error of 0.087, while the estimated moderating effect of the number of weeks equals -0.157 with a standard error of 0.036. This means that if there was no prior contact between pupils and teachers, elevating the expectancy of teachers can be expected to have a positive effect, $z = 4.678$, $p < 0.001$, but this treatment effect decreases significantly with the length of prior contact, $z = -4.361$, $p < 0.001$.

Mixed Effects Model

In the fixed effects regression model, possible differences in population effect sizes are entirely attributed to (known) study characteristics. The dependence of δ_j on the study characteristics is considered to be nonstochastic. The *random effects regression model* accounts for the possibility that population effect sizes vary partly randomly, partly according to known study characteristics:

$$d_j = \delta_j + e_j = \gamma_0 + \sum_{s=1}^s \gamma_s W_{sj} + u_j + e_j, \quad (14)$$

Since in the random effects regression model the population effect sizes depend on fixed effects (the γ 's) and random effects (the u 's), the model is also called a *mixed effects model*. Raudenbush and Bryk [51] showed that the mixed effects meta-analytic model is a special case of a hierarchical linear model or **linear multilevel model**, and proposed to use **maximum likelihood estimation** procedures that are commonly used to estimate the parameters of multilevel models, assuming normal residuals. For the example, the maximum likelihood estimate of the residual between study variance equals zero. This means that for the example the model simplifies to the fixed effects regression model, and the parameter estimates and corresponding standard errors for the fixed effects are the same as the ones given above. Differences between the underlying population effect sizes are explained entirely by the length of prior contact between pupils and teachers.

Threats for Meta-analysis

Despite its growing popularity, meta-analysis has always been a point of controversy and has been the subject of lively debates (see, e.g., [13], [14], [33], [34], [62], [67] and [72]). Critics point to interpretation problems due to, among other things, combining studies of dissimilar quality, including dependent study results, incomparability of different kinds of effect size measures, or a lack of essential data to calculate effect size measures. Another problem is the 'mixing of apples and oranges' due to combining studies investigating dissimilar research questions or using dissimilar study designs, dissimilar independent, or dependent variables or participants from dissimilar populations.

The criticism on meta-analysis that probably received most attention is the *file drawer problem* [57], which refers to the idea that the drawers of researchers may be filled with statistically nonsignificant unpublished study results, since researchers are more inclined to submit manuscripts describing significant results, and manuscripts with significant results are more likely to be accepted for publication. In addition, the results of small studies are less likely to be published, unless the observed effect sizes are relatively large. One way to detect the file drawer problem and the resulting *publication bias* is to construct a *funnel plot* [40]. A funnel plot is a scatter plot with the study sample size as the vertical axis and the observed effect sizes as the horizontal axis. If there is no publication bias, observed effect sizes of studies with smaller sample sizes will generally be more variable, while the expected mean effect size will be independent of the sample size. The shape of the scatter plot therefore will look like a symmetric funnel. In case of publication bias, most extreme negative effect sizes, or statistically nonsignificant effect sizes are not published and thus less likely to be included in the meta-analytic data set, resulting in a gap in the funnel plot for small studies and small or negative effect sizes. A funnel plot of the data from the example² is presented in Figure 2.

It can be seen that there is indeed some evidence for publication bias: for small studies there are some extreme positive observed effect sizes but no extreme negative observed effect sizes, resulting in a higher mean effect size for smaller studies. The asymmetry of the funnel plot, however, is largely due to only two effect sizes, and may well be caused by coincidence. This is confirmed by performing a distribution free statistical test for testing the correlation between effect size and sample size, based

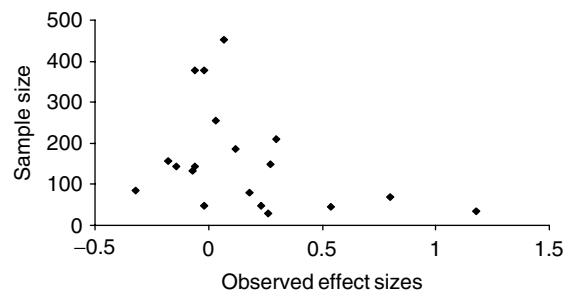


Figure 2 Funnel plot for the example data

on **Spearman's rho** [2]. While there is a tendency of a negative correlation, this relation is statistically not significant at the 0.05-level, $z = -1.78$, $p = .08$. Since a negative correlation between effect size and the sample size is expected in the presence of publication bias, an alternative approach to assess publication bias is to include the sample size as a moderator variable in a meta-analytic regression model. More information about methods for identifying and correcting for publication bias can be found in [2].

One might expect that some of the problems will become less important due to the fact that since the rise of meta-analysis, researchers and editors became aware of the importance of publishing nonsignificant results and of reporting exact P values, effect sizes, or test statistics, and meta-analysts became aware of the importance of looking for nonpublished study results. In addition, some of the criticisms are especially applicable to the relatively simple early meta-analytic techniques. The problem of 'mixing apples and oranges', for instance, is less pronounced if the heterogeneity in effect sizes is appropriately modeled using a random effects and/or by using moderator variables. 'Mixing apples and oranges' can even yield interesting information if a regression model is used, and the 'kind of fruit' is included in the model by means of one or more moderator variables. The use of the framework of multilevel models for meta-analysis further can offer an elegant solution for the problem of multiple effect size measures in some or all studies: a three-level model can be used, modeling within-study variation in addition to sampling variation and between study variation (see [20] for an example). Nevertheless, a meta-analysis remains a tenuous statistical analysis that should be performed rigorously.

Literature and Software

The article by Glass [23] and the formulation of some of the simple meta-analytic techniques by Glass and colleagues, (see e.g., [24], [25] and [64]), might be considered as the breakthrough of meta-analysis. Besides the calculation of the mean and standard deviation of the effect sizes, moderator variables are looked for by calculating correlation coefficients between study characteristics and effect sizes, by means of a multiple regression analysis

or by performing separate meta-analyses for different groups of studies. In the 1980s, these meta-analytic techniques were further developed, and the focus moved from estimating the mean effect size to detecting and explaining study heterogeneity. Hunter, Schmidt, and Jackson [35, 39] especially paid attention to possible sources of error (e.g., sampling error, measurement error, range restriction, computational, transcriptional, and typographical errors) by which effect sizes are affected, and to the correction of effect sizes for these artifacts. Rosenthal [58] described some simple techniques to compare and combine P values or measures of effect size. A more statistically oriented introduction in meta-analysis, including an overview and discussion of methods for combining P values, is given by Hedges and Olkin [31].

Rubin [61] proposed the random effects model for meta-analysis, which was further developed by DerSimonian and Laird [11] and Hedges and Olkin [31]. Raudenbush and Bryk [51] showed that the general framework of hierarchical linear modeling encompasses a lot of previously described meta-analytic methodology, yielding similar results [70], but at the same time extending its possibilities by allowing modeling random and fixed effects simultaneously in a mixed effects model. Goldstein, Yang, Omar, Turner, and Thompson [27] illustrated the flexibility of the use of hierarchical linear models for meta-analysis. The flexibility of the hierarchical linear models makes them also applicable for, for instance, combining the results from single-case empirical studies [68, 69]. Parameters of these hierarchical linear models are usually estimated using maximum likelihood procedures, although other estimation procedures could be used, for instance Bayesian estimation [22, 65].

An excellent book for basic and advanced meta-analysis, dealing with each of the steps of a meta-analysis is Cooper and Hedges [8]. This and other instructional, methodological, or application-oriented books on meta-analysis are reviewed by Becker [1]. More information about one of the steps in a meta-analysis, converting summary statistics, test statistics, P values and measures of effect size to a common measure is found in [8], [43], [53], [54], [60] and [66].

Several packages for performing meta-analysis are available, such as META [63] that implements techniques for fixed and random effects models and can

be downloaded free of charge together with a manual from http://www.fu-berlin.de/gesund/gesu_engl/meta_e.htm, Advanced Basic meta-analysis [45], implementing the ideas of Rosenthal, and MetaWin for performing meta-analyses using fixed, random, and mixed effects models and implementing parametric or resampling based tests (<http://www.metawinsoft.com>). Meta-analyses however can also be performed using general statistical packages, such as SAS [73]. A more complete overview of specialized and general software for performing meta-analysis, together with references to software reviews is given on the homepages from William Shadish (<http://faculty.ucmerced.edu/wshadish/Meta-Analysis%20Links.htm>) and Alex Sutton (<http://www.prw.le.ac.uk/epidemio/personal/ajs22/meta>). Since meta-analytic models can be considered as special forms of the hierarchical linear model, software, and estimation procedures for hierarchical linear models can be used. Examples are MLwiN (<http://multilevel.ioe.ac.uk>), based on the work of the *Centre for Multi-level Modelling* from the *Institute of Education* in London [26], HLM (<http://www.ssicentral.com/hlm/hlm.htm>), based on the work of Raudenbush and Bryk [52] and SAS proc MIXED [74].

Notes

1. The sampling distribution of Q often is only roughly approximated by a chi-square distribution, resulting in a conservative or liberal homogeneity test. Van den Noortgate & Onghena [71], therefore, propose a bootstrap version of the homogeneity test.
2. Equation 3 was used to calculate group sizes based on the observed effect sizes and standard errors given in Table 1. In each study, the two group sizes were assumed to be equal.

References

- [1] Becker, B.J. (1998). Mega-review: books on meta-analysis, *Journal of Educational and Behavioral Statistics* **1**, 77–92.
- [2] Begg, C. (1994). Publication bias, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds, Sage, New York, pp. 399–409.
- [3] Carter, D.L. (1971). The effect of teacher expectations on the self-esteem and academic performance of seventh grade students (Doctoral dissertation, University of Tennessee, 1970). *Dissertation abstracts International*, **31**, 4539–A. (University Microfilms No. 7107612).
- [4] Claiborn, W. (1969). Expectancy effects in the classroom: A failure to replicate. *Journal of Educational Psychology*, **60**, 377–383.
- [5] Cochran, W.G. (1954). The combination of estimates from different experiments, *Biometrics* **10**, 101–129.
- [6] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.
- [7] Conn, L.K., Edwards, C.N., Rosenthal, R. & Crowne, D. (1968). Perception of emotion and response to teachers' expectancy by elementary school children. *Psychological Reports*, **22**, 27–34.
- [8] Cooper, H. & Hedges, L.V. (1994a). *The Handbook of Research Synthesis*, Sage, New York.
- [9] Cooper, H. & Hedges, L.V. (1994b). Research synthesis as a scientific enterprise, in *The Handbook of Research Synthesis*, H. Cooper & L.V. Hedges, eds, Sage, New York, pp. 3–14.
- [10] Cronbach, L.J. (1980). *Toward Reform of Program Evaluation*, Jossey-Bass, San Francisco.
- [11] DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177–188.
- [12] Evans, J. & Rosenthal, R. (1969). Interpersonal self-fulfilling prophecies: Further extrapolations from the laboratory to the classroom. *Proceedings of the 77th Annual Convention of the American Psychological Association*, **4**, 371–372.
- [13] Eysenck, H.J. (1978). An exercise in mega-silliness, *American Psychologist* **39**, 517.
- [14] Eysenck, H.J. (1995). Meta-analysis squared – Does it make sense? *The American Psychologist* **50**, 110–111.
- [15] Fielder, W.R., Cohen, R.D. & Feeney, S. (1971). An attempt to replicate the teacher expectancy effect. *Psychological Reports*, **29**, 1223–1228.
- [16] Fine, L. (1972). The effects of positive teacher expectancy on the reading achievement of pupils in grade two (Doctoral dissertation, Temple University, 1972). *Dissertation Abstracts International*, **33**, 1510–A. (University Microfilms No. 7227180).
- [17] Fiske, D.W. (1983). The meta-analytic revolution in outcome research, *Journal of Consulting and Clinical Psychology* **51**, 65–70.
- [18] Fleming, E. & Anttonen, R. (1971). Teacher expectancy or my fair lady. *American Educational Research Journal*, **8**, 241–252.
- [19] Flowers, C.E. (1966). Effects of an arbitrary accelerated group placement on the tested academic achievement of educationally disadvantaged students (Doctoral dissertation, Columbia University, 1966). *Dissertation Abstracts International*, **27**, 991–A. (University Microfilms No. 6610288).
- [20] Geeraert, L., Van den Noortgate, W., Hellinckx, W., Grietens, H., Van Assche, V. & Onghena, P. (2004). The effects of early prevention programs for families with young children at risk for physical child abuse and neglect. A meta-analysis, *Child Maltreatment* **9**(3), 277–291.

- [21] Ginsburg, R.E. (1971). An examination of the relationship between teacher expectations and student performance on a test of intellectual functioning (Doctoral dissertation, University of Utah, 1970). *Dissertation Abstracts International*, **31**, 3337–A. (University Microfilms No. 710922).
- [22] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, London.
- [23] Glass, G.V. (1976). Primary, secondary, and meta-analysis of research, *Educational Researcher* **5**, 3–8.
- [24] Glass, G.V. (1977). Integrating findings: the meta-analysis of research, *Review of Research in Education* **5**, 351–379.
- [25] Glass, G.V., McGraw, B. & Smith, M.L. (1981). *Meta-Analysis in Social Research*, Sage, Beverly Hills.
- [26] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold, London.
- [27] Goldstein, H., Yang, M., Omar, R., Turner, R. & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects, *Journal of the Royal Statistical Society. Series C, Applied Statistics* **49**, 339–412.
- [28] Grierger, R.M.H. (1971). The effects of teacher expectancies on the intelligence of students and the behaviors of teachers (Doctoral dissertation, Ohio State University, 1970). *Dissertation Abstracts International*, **31**, 3338–A. (University Microfilms No. 710922).
- [29] Gosset, W.S. (1908). The probable error of a mean, *Biometrika* **6**, 1–25.
- [30] Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators, *Journal of Educational Statistics* **6**, 107–128.
- [31] Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta-analysis*, Academic Press, Orlando.
- [32] Henrikson, H.A. (1971). An investigation of the influence of teacher expectation upon the intellectual and academic performance of disadvantaged children (Doctoral dissertation, University of Illinois Urbana, Champaign, 1970). *Dissertation Abstracts International*, **31**, 6278–A. (University Microfilms No. 7114791).
- [33] Holroyd, K.A. & Penzien, D.B. (1989). Letters to the editor. Meta-analysis minus the analysis: a prescription for confusion, *Pain* **39**, 359–361.
- [34] Hoyle, R.H. (1993). On the relation between data and theory, *The American Psychologist* **48**, 1094–1095.
- [35] Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*, Sage, Newbury Park.
- [36] Jose, J. & Cody, J. (1971). Teacher-pupil interaction as it relates to attempted changes in teacher expectancy of academic ability achievement. *American Educational Research Journal*, **8**, 39–49.
- [37] Keshock, J.D. (1971). An investigation of the effects of the expectancy phenomenon upon the intelligence, achievement, and motivation of inner-city elementary school children (Doctoral dissertation, Case Western Reserve, 1970). *Dissertation Abstracts International*, **32**, 243–A. (University Microfilms No. 7119010).
- [38] Kester, S.W. & Letchworth, G.A. (1972). Communication of teacher expectations and their effects on achievement and attitudes of secondary school students. *Journal of Educational Research*, **66**, 51–55.
- [39] Hunter, J.E., Schmidt, F.L. & Jackson, G.B. (1982). *Meta-analysis: Cumulating Findings Across Research*, Sage, Beverly Hills.
- [40] Light, R.J. & Pillemer, D.B. (1984). *Summing up: The Science of Reviewing Research*, Harvard University Press, Cambridge.
- [41] Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis, *American Psychologist* **48**, 1181–1209.
- [42] Maxwell, M.L. (1971). A study of the effects of teacher expectations on the IQ and academic performance of children (Doctoral dissertation, Case Western Reserve, 1970). *Dissertation Abstracts International*, **31**, 3345–A. (University Microfilms No. 7101725).
- [43] Morris, S.B. & DeShon, R.P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs, *Psychological Methods* **7**, 105–125.
- [44] Mosteller, F. & Colditz, G.A. (1996). Understanding research synthesis (Meta-analysis), *Annual Review Public Health* **17**, 1–23.
- [45] Mullen, B. (1989). *Advanced BASIC Meta-analysis*, Erlbaum, Hillsdale.
- [46] National Research Council. (1992). *Combining Information. Statistical Issues and Opportunities for Research*, National Academy Press, Washington.
- [47] Pellegrini, R. & Hicks, R. (1972). Prophecy effects and tutorial instruction for the disadvantaged child. *American Educational Research Journal*, **9**, 413–419.
- [48] Pearson, K. (1904). Report on certain enteric fever inoculation statistics, *British Medical Journal* **3**, 1243–1246.
- [49] Pillemer, D.B. (1984). Conceptual issues in research synthesis, *Journal of Special Education* **18**, 27–40.
- [50] Raudenbush, S.W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: a synthesis of findings from 18 experiments, *Journal of Educational Psychology* **76**, 85–97.
- [51] Raudenbush, S.W. & Bryk, A.S. (1985). Empirical Bayes meta-analysis, *Journal of Educational Statistics* **10**, 75–98.
- [52] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, London.
- [53] Ray, J.W. & Shadish, R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology* **64**, 1316–1325.
- [54] Richardson, J.T.E. (1996). Measures of effect size, *Behavior Research Methods, Instruments & Computers* **28**, 12–22.

-
- [55] Rosenthal, R., Baratz, S. & Hall, C.M. (1974). Teacher behavior, teacher expectations, and gains in pupils' rated creativity. *Journal of Genetic Psychology*, **124**, 115–121.
- [56] Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- [57] Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results, *Psychological Bulletin* **86**, 638–641.
- [58] Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*, Sage, Newbury Park.
- [59] Rosenthal, R. (1998). Meta-analysis: concepts, corollaries and controversies, in *Advances in Psychological Science, Vol. 1: Social, personal, and Cultural Aspects*, J.G. Adair, D. Belanger & K.L. Dion, eds, Psychology Press, Hove.
- [60] Rosenthal, R., Rosnow, R.L. & Rubin, D.B. (2000). *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*, Cambridge University Press, Cambridge.
- [61] Rubin, D.B. (1981). Estimation in parallel randomized experiments, *Journal of Educational Statistics* **6**, 377–400.
- [62] Schmidt, F.L. (1993). Data, theory, and meta-analysis: response to Hoyle, *The American Psychologist* **48**, 1096.
- [63] Schwarzer, R. (1989). *Meta-analysis Programs. Program Manual*, Institut für psychologie, Freie Universität Berlin, Berlin.
- [64] Smith, M.L. & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies, *American Psychologist* **32**, 752–760.
- [65] Smith, T.S., Spiegelhalter, D.J. & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine* **14**, 2685–2699.
- [66] Tatsuoaka, M. (1993). Effect size, in *Data Analysis in the Behavioral Sciences*, G. Keren & C. Lewis, eds, Erlbaum, Hillsdale, pp. 461–479.
- [67] Thompson, S.G. & Pocock, S.J. (1991). Can meta-analyses be trusted? *Lancet* **338**, 1127–1130.
- [68] Van den Noortgate, W. & Onghena, P. (2003a). Combining single-case experimental studies using hierarchical linear models, *School Psychology Quarterly* **18**, 325–346.
- [69] Van den Noortgate, W. & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research, *Behavior Research Methods, Instruments, & Computers* **35**, 1–10.
- [70] Van den Noortgate, W. & Onghena, P. (2003c). Multi-level meta-analysis: a comparison with traditional meta-analytical procedures, *Educational and Psychological Measurement* **63**, 765–790.
- [71] Van den Noortgate, W. & Onghena, P. (2003d). A parametric bootstrap version of Hedges' homogeneity test, *Journal of Modern Applied Statistical Methods* **2**, 73–79.
- [72] Van Mechelen, I. (1986). In search of an interpretation of meta-analytic findings, *Psychologica Belgica* **26**, 185–197.
- [73] Wang, M.C. & Bushman, B.J. (1999). *Integrating Results Through Meta-analytic Review Using SAS Software*, SAS Institute, Inc., Cary.
- [74] Yang, M. (2003). *A Review of Random Effects Modelling in SAS (Release 8.2)*, Institute of Education, London.

WIM VAN DEN NOORTGATE AND
PATRICK ONGHENA

Microarrays

LEONARD C. SCHALKWYK
Volume 3, pp. 1217–1221

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Microarrays

The culture of molecular biology values categorical results, and a generation of scientists used to bands on gels and DNA sequence is confronting statistics for the first time as they design microarray experiments and analyze data. This can sometimes be seen in the style of data presentation used, for example, the red-green false color images often used to summarize differences between the signal from a pair of RNA samples hybridized to the same array in an experiment using two fluorescent labels. This kind of experimental design allows many of the possible sources of measurement error to be at least equalized between a pair of samples, but gives no information on the reliability of the result. This has led to a widespread understanding of microarrays as a survey method, whose indicative results must be checked by other methods. As new uses for microarrays are conceived and costs come down, this is changing and more attention is being paid to statistical methods. This is particularly true for the highly standardized, industrially produced arrays such as the Affymetrix GeneChips.

The Affymetrix Array

I will discuss the Affymetrix product in this article, though many aspects of the discussion are applicable to other kinds of microarrays. I will also limit the discussion to arrays used for gene expression analysis, although once again much of the discussion will also be relevant to microarrays designed for genotyping and other purposes. I will start with a brief description of the method, going through the steps of data analysis from the raw image upwards, addressing experimental design last.

The principle of measurement in microarrays is nucleic acid hybridization. Immobilized on the array, are known sequences of DNA (normally, although RNA or synthetic NA analogues would also be possible) known as *probes*. Reacted with these in solution is a mixture of unknown (labeled RNA or DNA) fragments to be analyzed, known as *targets* (traditionally these terms were used the other way around). The targets bind to the probes in (ideally) a sequence-specific manner and fill the available probe sites to an extent that depends on

the target concentration. After the unbound target is washed away, the quantity of target bound to each probe is determined by a fluorescence method.

In the case of Affymetrix arrays, the probes are synthesized *in situ* using a proprietary photolithography method, which generates an unknown but small quantity of the desired sequence on an ever-smaller feature size. On current products the feature size is $11 - \mu\text{m}$ square, allowing nearly 300 000 features on a 6-mm square chip. With arrays made by mechanical spotting or similar methods, locating features and generating a representative intensity value can be demanding, but grid location and adjustment are relatively straightforward using the physical anchorage and fluorescent guidespots in the Affymetrix system. Each of the grid squares (representing a feature) contains approximately 25 pixels after an outer rim has been discarded; these are averaged. With current protocols, saturated ($\text{SD}=0$) cells are rare.

Normalization and Outliers

Because the amplified fluorescence assay used in the Affymetrix system is in a single color, experimental fluctuations in labeling, hybridization, and scanning need to be taken into account by normalizing across arrays of an experiment. This has been an extensively studied topic in array studies of all kinds and there are many methods in use [10]. A related issue is the identification of outlying intensity values likely to be due to production flaws, imperfect hybridization, or dust specks. The Affymetrix software performs a simple scaling using a chip-specific constant chosen to make the trimmed mean of intensities across the entire chip a predetermined value. This is probably adequate for modest differences in overall intensity between experiments done in a single series, but it does assume linearity in the response of signal to target concentration, which is unlikely to be true across the entire intensity range. Several excellent third party programs that are free, at least for academic use, offer other options. *dChip* [9], for example, finds a subset of probes whose intensity ranking does not vary across an experimental set of chips and uses this to fit a normalization curve. An excellent toolbox for exploring these issues is provided by packages from the Bioconductor project, particularly *affy* [5].

Summarizing the Probeset

The high density made possible by the photolithographic method allows multiple probes to be used for each transcript (also called *probeset*), which means that the (still not fully understood) variation in the binding characteristics of different oligonucleotide sequences can be buffered or accounted for. For each 25-nucleotide probe (perfect match (PM)), a control probe with a single base in the central (13th) position is synthesized. This was conceived as a control for the sequence specificity of the hybridization or a background value, but this is problematic. Each probeset consists (in current Affymetrix products) of 11 such perfect match-mismatch probe pairs. The next level of data reduction is the distillation of the probeset into a single intensity value for the transcript. Methods of doing this remain under active discussion. Early Affymetrix software used a proprietary method to generate signal values from intensities, but Microarray Suite (MAS) versions 5 and later use a relatively well-documented robust mean approach [7]. This is a robust mean (Tukey's biweight) for the perfect match PM probes and separately for the mismatch (MM) probes. The latter is subtracted and the result is set to zero where $MM > PM$. Other summary methods use model-fitting with or without MM values or thermodynamic data on nucleic acid annealing, to take into account the different performance of different probe sequences. These may have effects on the variability of measurements [8, 13, 12]. It may, under some circumstances, make sense to treat the measurements from individual probes as separate measurements [1].

Statistical Analysis of Replicated Arrays

Once a list of signal values (in some arbitrary units) is available, the interesting part begins and the user is largely on his own. By this I mean that specialized microarray software does not offer very extensive support for statistical evaluation of the data. An honorable exception is dChip which offers **analysis of variance** (ANOVA) through a link with R, and the *affy* package from Bioconductor [5] which is implemented in the same R statistical computing environment. It may be a wise design decision to limit the statistical features of specialized microarray software in that there is nothing unique about the data,

which really should be thought about and analyzed like other experimental data. The signal value for a given probeset is like any other single measurement, and variance and replication need to be considered just as in other types of experiments. Of course the issue of **multiple testing** requires particular attention, since tens of thousands of transcripts are measured in parallel.

Having worked with the experimental data and reduced it to a value for each probeset (loosely, transcript), the analyst can transfer the data to the statistical package he likes best. Because statistical tests are done on each of thousands of transcripts, a highly programmable and flexible environment is needed. R is increasingly popular for this purpose and there are several microarray related packages available. The *affy* package (part of BioConductor) can be installed from a menu in recent versions of R, and offers methods and data structures to do this quite conveniently. Most of the available methods for summarizing probesets are available and are practical on a personal computer (PC) with one gigabyte (GB) of random access memory (RAM). Whether processing of the data is done in R or with other software, it is a real time saver, especially for the beginner, to have the data in a structure that is not unnecessarily complicated – normally a dataframe with one row for each probeset (and the probeset ID as row name), and columns for each chip with brief names identifying the sample.

In advance of testing hypotheses, it may be good to *filter* the data, removing from consideration datasets that have no usable information, which may also aid in thinking about how much multiple testing is really being done. It is also the time to consider whether the data should be transformed. Because outlying signals should have been dealt with in the process of normalization and probeset summarization, the chief criterion for filtration is expression level. Traditional molecular biology wisdom is that roughly half of all genes are expressed in a given tissue. Whether a qualitative call of expression/no expression actually corresponds to a biological reality is open to question, and there is the added complication that any tissue will consist of multiple cell types with differing expression patterns. In practice, differences between weak expression signals will be dominated by noise. MAS 5 provides a presence/absence call for each probeset which can be used as a filtering criterion. One complication is that the fraction

of probesets ‘present’ varies from one hybridization to the next, and is used as an experimental quality measure. An alternative would be to consider those probesets whose signal values exceed the mean (or some other quantile) across the whole array.

It is common to log-transform signal data; this is convenient in terms of informal fold-change criteria, although the statistical rationale is usually unclear. The distribution of signal values across a single array is clearly not normally distributed, but the distribution that is generally of interest is for a given probeset across multiple arrays, where there are fewer data.

The natural starting point in analyzing the data from a set of microarrays is to look for differences between experimental groups. This will generally be a more or less straightforward ANOVA, applied to each of the probesets on the array used. Depending on the design of the experiment, it may be useful to use a **generalized linear mixed model** to reflect the different sources of variance. Some additional power should also be available from the fact that the dispersion of measurements in the thousands of probesets across the set of microarrays is likely to be similar. A reasonable approach to this may be ‘variance shrinking’ [4]. The most difficult and controversial aspect of the analysis is the extent to which the assumptions of ANOVA, in particular, independence within experimental groups, are violated. Peculiarly extreme views on this (especially with regard to inbred strains), as well as bogus inflated significance, are encountered. Pragmatically there will be some violation, but as long as the design and analysis take careful and fair account of the most important sources of variance, we are unlikely to be fooled.

The resulting vector of P values, or more to the point, of probesets ranked by P values, then needs to be evaluated in light of the experimental question and your own opinions on multiple testing. As a starting point, Bonferroni correction (see **Multiple Comparison Procedures**) for the number of probesets is extremely, perhaps absurdly, conservative, but in many experiments there will be probesets that will differ even by this criterion. Many probesets will have background levels of signal for all samples, and among those that present something to measure, there will be many that are highly correlated. False-discovery-rate (FDR) methods [2, 11] offer a sensible

alternative, but in general, experimental results will be a ranked list of candidates, and where this list is cut off may depend on practicalities, or on other information. On the Affymetrix arrays, there is a substantial amount of duplication (multiple probesets for the same transcript), which may give additional support to some weak differences. In other cases, multiple genes of the same pathway, gene family, or genomic region might show coordinate changes for which the evidence may be weak when considered individually. This argues for considerable exploration of the data, and also public archiving of complete data sets.

Hierarchical clustering of expression patterns across chips using a distance measure such as $1 - |r|$, where r is the Pearson correlation between the patterns of two transcripts, is an often used example of a way of exploring the data for groups of transcripts with coordinate expression patterns. This is best done with a filtered candidate list as described earlier, because the bulk of transcripts will not differ, and little will be learned by clustering noise at high computational cost. There is a growing amount of annotation on each transcript and a particularly convenient set of tools for filtering, clustering, and graphic presentation with links to annotation is provided by dChip.

In some experiments, the objective is not so much the identification of interesting genes as classification of samples (targets), such as diagnostic samples [6]. Here, classification techniques such as **discriminant analysis** or supervised **neural networks** can be used to allocate samples to prespecified groups.

Experimental Design

In developing an experimental design, initially scientists tend to be mesmerized by the high cost of a single determination, which is currently about £500/\$1000 in consumables for one RNA on one array. The cost of spotted cDNA arrays (especially the marginal cost for a laboratory making their own) can be much lower, but because of lower spotting densities and lower information per spot (because of greater technical variation), it is not simple to make a fair comparison. The Affymetrix software is designed to support the side-by-side comparison so loved by molecular biologists. A ‘presence’ call is produced complete with a P value for each probeset, and there is provision for pairwise analysis in

which a 'difference' call with a P value is produced. These treat the separate probes of each probeset as separate determinations in order to provide a semblance of statistical support to the values given. This is not a fair assessment of even the purely technical error of the experimental determination, because it knows nothing of the differences in RNA preparation, amplification, and labeling between the two samples. Overall intensity differences between chips resulting from differences in labeling efficiency should be largely dealt with by normalization, but it is easy to see how there could be transcript-specific differences. RNA species differ in stability (susceptibility to degradation by nucleases) and length, for example. Nonetheless, the protocols are quite well standardized and the two chip side-by-side design does give a reasonably robust indication of large differences between samples, which is also what has traditionally been sought from the two color spotted cDNA array experiment. The need for replicates starts to become obvious when the resulting list of candidate differences is reviewed – it may be fairly long, and the cost of follow up of the levels of many individuals by other measurement methods such as quantitative polymerase chain reaction (PCR) mounts quickly.

The greatest source of variation that needs to be considered in the experimental design is the biological variation in the experimental material. This is particularly so when dealing with human samples where it has to be remembered that every person (except for identical twins) is a unique genetic background that has experienced a unique and largely uncontrolled environment. Further variation comes from tissue collection, and it is often uncertain whether tissue samples are precisely anatomically comparable. This biological variation can be physically averaged using pooling. The temptation to do this should be resisted as much as is practical because it does not give any information on variance. Realistically, experimental designs are decided after dividing the funds available by the cost of a chip. There is no rule of thumb for how many replicates are necessary, since this depends on the true effect size and the number of groups to be compared. Even with eight experimental groups and four replicates of each, there is only 80% power to detect between-group variability of one within-group SD. Whether this corresponds to a 10% difference or a twofold difference depends on the variability of

the biological material (*see Sample Size and Power Calculation*).

Rather than minimizing cost by minimizing array replicates, value for money can ideally be maximized by using materials which can be mined many times. An outstanding example of this is the WebQTL database [3], where microarray data on a growing number of tissues are stored along with other phenotype data for the BxD mouse recombinant inbred panel. Here, the value of the data for repeated mining is particularly great, because the data derives from a genetically reproducible biological material (inbred mouse strains).

References

- [1] Barrera, L., Benner, C., Tao, Y.-C., Winzeler, E. & Zhou, Y. (2004). Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays, *BMC Bioinformatics* **5**, 42.
- [2] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B* **57**, 289–300.
- [3] Chesler, E.J., Lu, L., Wang, J., Williams, R.W. & Manly, K.F. (2004). WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior, *Nature Neuroscience* **7**, 485–486.
- [4] Churchill, G. (2004). Using ANOVA to analyze microarray data, *BioTechniques* **37**, 173–175.
- [5] Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. (2004). Affy-analysis of affymetrix genechip data at the probe level, *Bioinformatics* **20**, 307–315.
- [6] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. & Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**, 531–537.
- [7] http://www.affymetrix.com/support/technical/technotes/statistical-reference_guide.pdf Affymetrix Part No. 701110 Rev 1 (2001)
- [8] Li, C. & Wong, W.H. (2001a). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Sciences of the United States of America* **98**, 31–36.
- [9] Li, C. & Wong, W.H. (2001b). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology* **2**(8), research 0032.1–0032.11.
- [10] Quackenbush, J. (2001). Microarray data normalization and transformation, *Nature Genetics* **32**, Suppl, 496–501.

- [11] Storey, J.D. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.
- [12] Wu, Z. & Irizarry, R. (2004). Preprocessing of oligonucleotide array data, *Nature Biotechnology* **22**, 656–657.
- [13] Zhang L., Miles M.F. & Adalpe F.D. (2003). A model of molecular interactions on short oligonucleotide arrays, *Nature Biotechnology* **21**, 818–821.

LEONARD C. SCHALKWYK

Mid-*P* Values

VANCE W. BERGER

Volume 3, pp. 1221–1223

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Mid- P Values

One of the most important statistical procedures is the hypothesis test, in which a formal analysis is conducted to produce a P value, which is used to summarize the strength of evidence against the null hypothesis. There are a few complicated testing procedures that are not based on a test statistic [3]. However, these procedures are the exception, and the general rule is that a test statistic is the basis for a hypothesis test. For example, if one were to test the fairness of a coin, one could toss the coin a given number of times, say ten times, and record the number of heads observed. This number of heads would serve as the test statistic, and it would be compared to a known null reference distribution, constructed under the assumption that the coin is, in fact, fair. The logic is an application of *modus tollens* [1], which states that if A implies B, then not B implies not A. If the coin is fair, then we expect (it is likely that we will see) a certain number of heads. If instead we observe a radically different number of heads, then we conclude that the coin was not fair.

How many heads do we need to observe to conclude that the coin is not fair? The distance between the observed data and what would be predicted by the null hypothesis (in this case, that the coin is fair) is generally measured not by the absolute magnitude of the deviation (say the number of heads minus the null expected value, five), but rather by a P value. This P value is the probability, computed under the assumption that the null hypothesis is true, of observing a result as extreme as, or more extreme than, the one we actually observed. For example, if we are conducting a one-sided test, we would like to conclude that the coin is not fair if, in fact, it is biased toward producing too many heads. We then observe eight heads out of the ten tosses. The P value is the probability of observing eight, nine, or ten heads when flipping a fair coin. Using the binomial distribution (see **Binomial Distribution: Estimating and Testing Parameters**), we can compute the P value, which is $[45 + 10 + 1]/1024$, or 0.0547. It is customary to test at the 0.05 significance level, meaning that the P value would need to be less than 0.05 to be considered significant.

Clearly, eight of ten heads is not significant at the 0.05 significance level, and we cannot rule out that the coin is fair. One may ask if we could have rejected

the null hypothesis had we observed instead nine heads. In this case, the P value would be $11/1024$, or 0.0107. This result would be significant at the 0.05 level and gives us a decision rule: reject the null hypothesis if nine or ten heads are observed in ten flips of the coin. Generally, for continuous distributions, the significance level that is used to determine the decision rule, 0.05 in this case, is also the null probability of rejection, or the probability of a Type I error. But in this case, we see that a Type I error occurs if the coin is fair and we observe nine or ten heads, and this outcome occurs with probability 0.0107, not 0.05. There is a discrepancy between the intended Type I error rate, 0.05, and the actual Type I error rate, 0.0107.

It is unlikely that there would be any serious objection to having an error probability that is smaller than it was intended to be. However, a consequence of this conservatism is that the power to detect the alternative also suffers, and this would lead to objections. Even if the coin is biased toward heads, it may not be so biased that nine or ten heads will typically be observed. The extent of conservatism will be decreased with a larger sample size, but there is an approach to dealing with conservatism without increasing the sample size. Specifically, consider these two probabilities, $P\{X > k\}$ and $P\{X \geq k\}$, where X is the test statistic expressed as a random variable (in our example, the number of heads to be observed) and k is the observed value of X (eight in our example). If these two quantities, $P\{X > k\}$ and $P\{X \geq k\}$, were the same, as they would be with a continuous reference distribution for X , then there would be no discreteness, no conservatism, no associated loss of power, and no need to consider the mid- P value.

But these two quantities are not the same when dealing with a discrete distribution, and using the latter makes the P value larger than it ought to be, because it includes the null probability of all outcomes with test statistic equal to k . That is, there are 45 outcomes (ordered sets of eight heads and two tails), and all have the same value of the test statistic, eight. What if we used just half of these outcomes? More precisely, what if we used half the probability of these outcomes, instead of all of it? We could compute $P\{X > k\} = 0.0107$ and $P\{X = k\} = 0.04395$, so $P\{X \geq k\} = 0.0107 + 0.04395 = 0.0547$ (as before), or we could use instead $0.0107 + (0.04395)/2 = 0.0327$. This latter

computation leads to the mid- P value. In general, the mid- p is $P\{X > k\} + (1/2)P\{X = k\}$. See [4–7] for more details regarding the development of the mid- P value.

Certainly, the mid- P value is smaller than the usual P value, and so it is less conservative. However, it does not allow one to recover the basic quantities, $P\{X > k\}$ and $P\{X \geq k\}$. Moreover, it is not a true P value, as it is anticonservative, in the sense that it does not preserve the true Type I error rate [2]. One modification of the mid- P value is based on recognizing the mid- P value as being the midpoint of the interval $[P\{X > k\}, P\{X \geq k\}]$. Instead of presenting simply the midpoint of this interval, why not present the entire interval, in the form $[P\{X > k\}, P\{X \geq k\}]$? This is the P value interval [2], and, as mentioned, it has as its midpoint the mid- P value. But it also tells us the usual P value as its upper endpoint, and it shows us the ideal P value, in the absence of any conservatism, as its lower endpoint. For the example, the P value interval would be (0.0107, 0.0547).

References

- [1] Asogawa, M. (2000). A connectionist production system, which can perform both *modus ponens* and *modus tollens* simultaneously, *Expert Systems* **17**, 3–12.
- [2] Berger, V.W. (2001). The p value interval as an inferential tool, *Journal of the Royal Statistical Society. Section D (The Statistician)* **50**, 79–85.
- [3] Berger, V.W. & Sackrowitz, H. (1997). Improving tests for superior treatment in contingency tables, *Journal of the American Statistical Association* **92**, 700–705.
- [4] Berr, G. & Armitage, P. (1995). Mid- p confidence intervals: a brief review, *Journal of the Royal Statistical Society. Series D (The Statistician)* **44**(4), 417–423.
- [5] Lancaster, H.O. (1949). The combination of probabilities arising from data in discrete distributions, *Biometrika* **36**, 370–382.
- [6] Lancaster, H.O. (1952). Statistical control of counting experiments, *Biometrika* **39**, 419–422.
- [7] Lancaster, H.O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association* **56**, 223–234.

VANCE W. BERGER

Minimum Spanning Tree

GEORGE MICHAILIDIS

Volume 3, pp. 1223–1229

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Minimum Spanning Tree

Introduction

The minimum spanning tree (MST) problem is one of the oldest problems in graph theory, dating back to the early 1900s [13]. MSTs find applications in such diverse areas as least cost electrical wiring, minimum cost connecting communication and transportation networks, network reliability problems, minimum stress networks, clustering and numerical taxonomy (see **Cluster Analysis: Overview**), algorithms for solving traveling salesman problems, and multiterminal network flows, among others. At the theoretical level, its significance stems from the fact that it can be solved in polynomial time – the execution time is bounded by a polynomial function of the problem size – by greedy type algorithms. Such algorithms always take the best immediate solution while searching for an answer. Because of their myopic nature, they are relatively easy to construct, but in many optimization problems they lead to suboptimal solutions. Problems solvable by greedy algorithms have been identified with the class of matroids. Furthermore, greedy algorithms have been extensively studied for their complexity structure and MST algorithms have been at the center of this endeavor.

In this paper, we present the main computational aspects of the problem and discuss some key applications in statistics, probability, and data analysis.

Some Useful Preliminaries

In this section, we introduce several concepts that prove useful for the developments that follow.

Definition 1: An *undirected graph* $G = (V, E)$ is a structure consisting of two sets V and E .

The elements of V are called *vertices* (nodes) and the elements of E are called *edges*. An edge $e = (u, v)$, $u, v \in V$ is an unordered pair of vertices u and v .

An example of an undirected graph is shown in Figure 1. Notice that there is a self-loop (an edge whose endpoints coincide) for vertex E and a multiedge between nodes B and E.

Definition 2: A *path* $p = \{v_0, v_1, \dots, v_k\}$ in a graph G from vertex v_0 to vertex v_k is a sequence of vertices such that (v_i, v_{i+1}) is an edge in G for $0 \leq i \leq k$. Any edge may be used only once in a path.

Definition 3: A *cycle* in a graph G is a path whose end vertices are the same; that is, $v_0 = v_k$.

Notice the cycle formed by the edges connecting nodes D, F, and G.

Definition 4: A graph G is said to be *connected* if there is a path between every pair of vertices.

Definition 5: A *tree* T is a *connected* graph that has *no cycles* (acyclic graph).

Definition 6: A *spanning tree* T of a graph G is a subgraph of G that is a tree and contains all the vertices of G .

A spanning tree is shown in Figure 2.

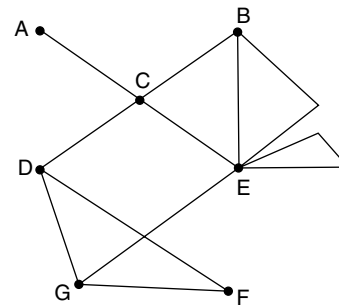


Figure 1 Illustration of an undirected graph

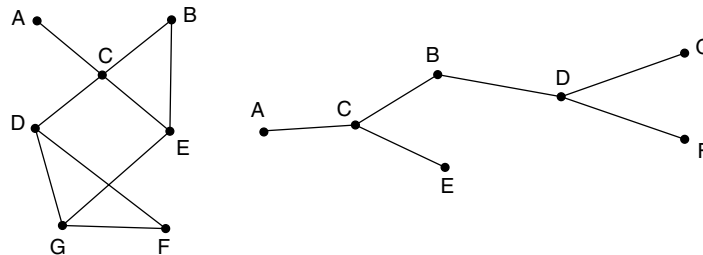


Figure 2 A simple undirected graph and a corresponding spanning tree

2 Minimum Spanning Tree

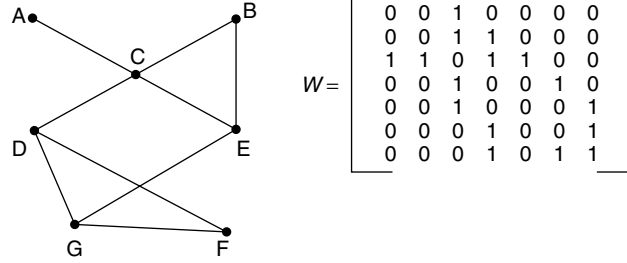


Figure 3 A simple undirected graph and its adjacency matrix W

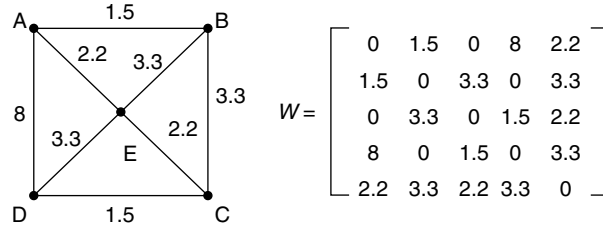


Figure 4 A weighted undirected graph and its adjacency matrix W

A useful mathematical representation of an undirected graph $G = (V, E)$ is through its *adjacency* matrix W , which is a $|V| \times |V|$ matrix with $W_{ij} = 1$, if an edge exists between vertices i and j and $W_{ij} = 0$, otherwise. An example of an undirected graph and its adjacency matrix is given in Figure 3. It can easily be seen that W is a binary symmetric matrix with zero diagonal elements.

In many cases, the edges are associated with nonnegative weights that capture the strength of the relationship between the end vertices. This gives rise to a *weighted* undirected graph that can be represented by a symmetric matrix with nonnegative entries, as shown in Figure 4.

The Minimum Spanning Tree Problem

The MST problem is defined as follows:

Let $G = (V, E)$ be a connected weighted graph. Find a spanning tree of G whose total edge-weight is a minimum.

Remark: When V is a subset of a metric space (e.g., V represents points on the plane equipped with some distance measure), then a solution T to the MST problem represents the shortest network connecting all points in V .

As most graph theory problems, the MST one is very simple to state, and has attracted a lot of interest for finding its solution. A history of the problem and the algorithmic approaches proposed for its solution can be found in Graham and Hell [13] with an update in Nesetril [20]. As the latter author observes ‘(the MST) is a cornerstone of combinatorial optimization and in a sense its cradle’.

We present next two classical (textbook) algorithms due to Prim [21] and Kruskal [16], respectively. For each algorithm, the solution T is initialized with the minimum weight edge and its two endpoints. Furthermore, let $v(T)$ denote the number of vertices in T and $v(F)$ the number of vertices in a collection of trees; that is a *forest*.

Prim’s algorithm: While $v(T) < |V|$ do:

- interrogate edges (in increasing order of their weights) until one is found that has one of its endpoints in T and its other endpoint in $V - T$ and has the minimum weight among all edges that satisfy the above requirement.
- Add this edge and its endpoint to T and increase $v(T)$ by 1.

Kruskal’s algorithm: While $v(F) < |V|$ do:

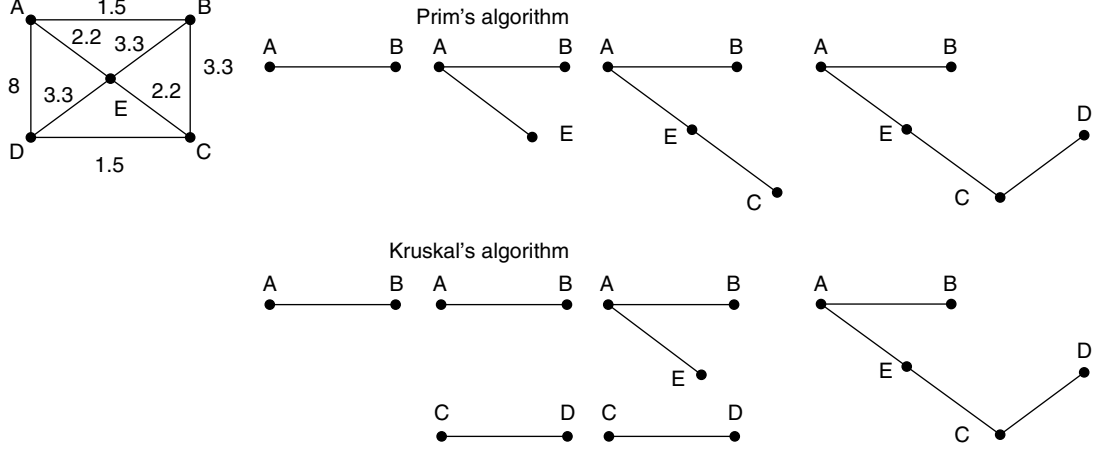


Figure 5 The progression of the two classical MST algorithms with a total cost of 7.4

- interrogate edges (in increasing order of their weights) until one is found that does not generate a cycle in the current forest F .
- Add this edge and its endpoints to F and increase $v(F)$ by 1 or 2.

An illustration of the progression of the two algorithms on a toy graph is shown in Figure 5. Notice that whereas Prim's algorithm grows a single tree, Kruskal's algorithm grows a forest. For a proof of the optimality of these two algorithms, see [23].

These two algorithms take in the worst case scenario approximately $|E| \times \log |V|$ step to compute the solution, which is mainly dominated by the sorting step. Over the last two decades better algorithms with essentially a number of steps proportional to the number of edges have been proposed by several authors; see [5, 7, 11]. However, the work of Moret et al. [19] suggests that in practice Prim's algorithm outperforms on average its competitors.

Applications of the MST

In this section, we provide a brief review of some important applications and results of the MST in statistics, data analysis, and probability.

Multivariate Two-sample Tests

In a series of papers, Friedman and Rafsky [8, 9] proposed the following generalization of classical

nonparametric two-sample tests for multivariate data (see **Multivariate Analysis: Overview**): suppose we have two multivariate samples on $\mathbb{R}^d$, $\mathcal{X}_n = \{X_1, X_2, \dots, X_n\}$ from a distribution F_X and $\mathcal{Y}_m = \{Y_1, Y_2, \dots, Y_m\}$ from another distribution F_Y . The null hypothesis of interest is $H_0: F_X = F_Y$ and the alternative hypothesis is $H_1: F_X \neq F_Y$. The proposed test procedure is as follows:

1. Generate a weighted graph G whose $n + m$ nodes represent the data points of the pooled samples, with the edge weights corresponding to Euclidean distances between the points.
2. Construct the MST, T of G .
3. Remove all the edges in T for which the endpoints come from different samples.
4. Define the test statistic $R_{n,m} = \#$ of disjoint subtrees.

Using the asymptotic normality of $R_{n,m}$ under H_0 , the null hypothesis is rejected at the significance level α , if

$$\frac{R_{n,m} - E(R_{n,m})}{\sqrt{\text{Var}(R_{n,m}|C_{n,m})}} < \Phi^{-1}(\alpha), \quad (1)$$

where $\Phi^{-1}(\alpha)$ is the α -quantile of the standard normal distribution and $C_{n,m}$ the number of edge pairs of T that share a common node. Expressions for $E(R_{n,m})$ and $\text{Var}(R_{n,m}|C_{n,m})$ are given in [8]. In [14] it is further shown that $R_{m,n}$ is asymptotically distribution-free under H_0 , and that the above test is universally consistent.

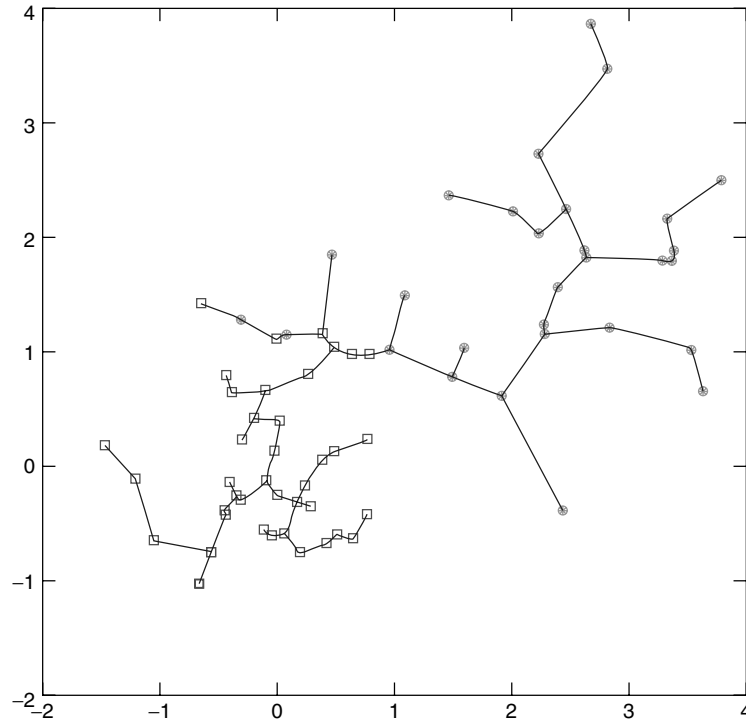


Figure 6 Demonstration of the multivariate two-sample runs test

A demonstration of the two-sample test is given in Figure 6. The 40 $\square$ points come from a two-dimensional multivariate mean zero normal distribution with $\sigma_1^2 = \sigma_2^2 = .36$ and $\sigma_{12} = 0$. The 30 $\star$ points come from a two-dimensional multivariate normal distribution with mean vector $(2, 2)$ and $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{12} = 0$ (see **Catalogue of Probability Density Functions**). The value of the observed test statistic is $R_{40,30} = 6$ and the z -score is around -6.1 , which suggests that H_0 is rejected in favor of H_1 .

Remark: In [8], a Smirnov type of test using a rooted MST is also presented for the multivariate two-sample problem.

MST and Multivariate Data Analysis

The goal of many multivariate data analytic techniques is to uncover interesting patterns and relationships in multivariate data (see **Multivariate Analysis: Overview**). In this direction, the MST has proved a useful tool for grouping objects into homogeneous groups, but also in visualizing the structure of multivariate data.

The presentation of Prim's algorithm in the section titled 'Applications of the MST' basically shows that the MST is essentially identical to the single linkage agglomerative hierarchical clustering algorithm [12] (see **Hierarchical Clustering**). By removing the $K - 1$ highest weight edges, one obtains a clustering solution with K groups [25]. An illustration of the MST as a clustering tool is given in Figure 7, where the first cluster corresponds to objects along a circular pattern and the second cluster to a square pattern inside the circle. It is worth noting that many popular clustering algorithms such as K-means and other agglomerative algorithms are going to miss the underlying group structure as the bottom right panel in Figure 7 shows.

The MST has also been used for identifying influential multivariate observations [15] (see **Multivariate Outliers**), for highlighting inaccuracies of low-dimensional representations of high-dimensional data through **multidimensional scaling** [3] and for visualizing structure in high dimensions [17] (see **k-means Analysis**).

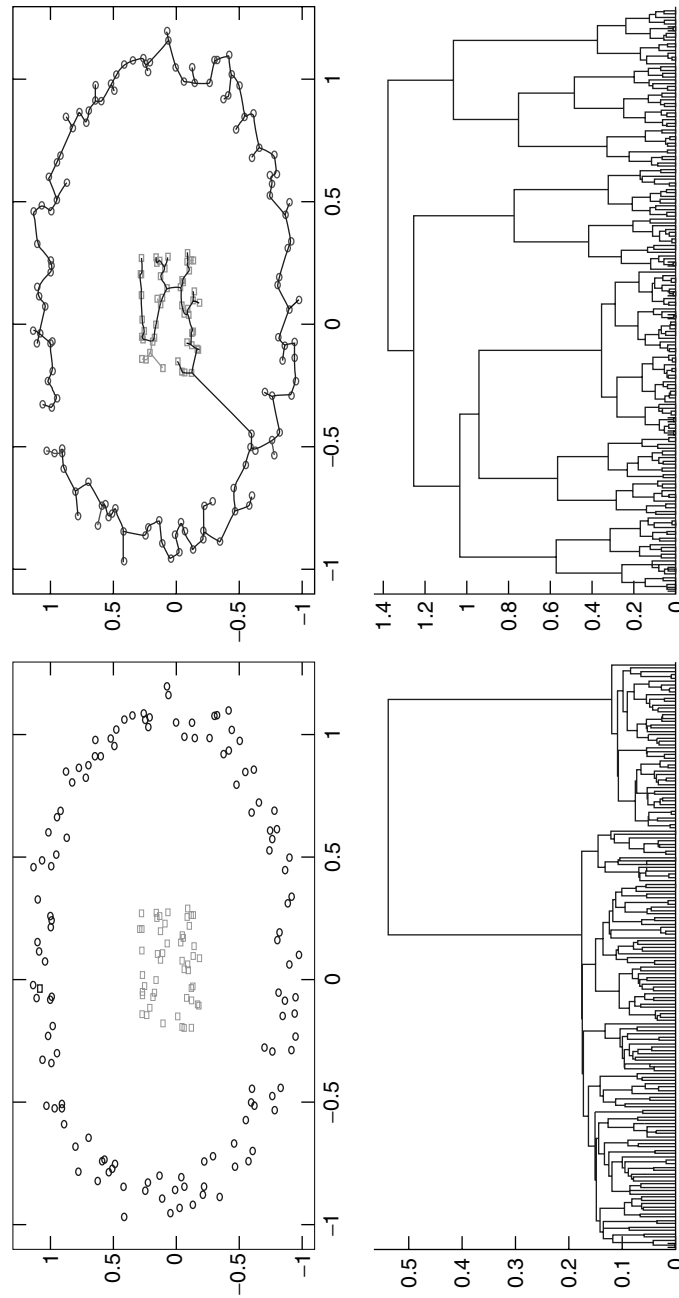


Figure 7 A two-dimensional data example with two underlying clusters (top left panel), the corresponding MST (top right panel), the single linkage dendrogram (bottom left panel) and the average linkage dendrogram (bottom right panel)

The MST in Geometric Probability

The importance of the MST in many practical situations is that it determines the dominant skeleton structure of a point set by outlining the shortest path between nearest neighbors. Specifically, given a set of points $\mathcal{X}_n = \{x_1, x_2, \dots, x_n\}$ in $\mathbb{R}^d$ the MST $T(\mathcal{X}_n)$ connects all the points in the set by using as the weight function on the edges of the underlying complete graph the Euclidean distance. Steele [22] established that if the distances d_{ij} , $1 \leq i, j \leq n$ are independent and identically distributed with common cdf $F(d_{ij})$, then

$$\lim_{n \rightarrow \infty} \frac{T(\mathcal{X}_n)}{n^{\frac{d-1}{d}}} = \beta(d) \left(\frac{d-1}{d} \right)^{1/d}, \text{ a.s.} \quad (2)$$

where $\beta(d)$ is a constant that depends only on the dimension d . The above result is a variation on a theme in geometric probability pioneered by the celebrated paper of Bearwood et al. [2] on the length of the traveling salesman tour in Euclidean space. The exact value of the constant $\beta(d)$ is not known in general (see [1]).

The above result has provided a theoretical basis for using MSTs for image registration [18], for pattern recognition problems [24] and in assignment problems in wireless networks [4].

Concluding Remarks

An MST is an object that has attracted a lot of interest in graph and network theory. In this paper, we have also shown several of its uses in various areas of statistics and probability. Some other applications of MSTs have been in defining measures of multivariate association [10] (an extension of Kendall's τ measure) and more recently in estimating the intrinsic dimensionality of a nonlinear manifold from sparse data sampled from it [6].

Acknowledgments

The author would like to thank Jan de Leeuw and Brian Everitt for many useful suggestions and comments. This work was supported in part by NSF under grants IIS-9988095 and DMS-0214171 and NIH grant 1P41RR018627-01.

References

[1] Avram, F. & Bertsimas, D. (1992). The minimum spanning tree constant in geometric probability and

under the independent model: a unified approach, *Annals of Applied Probability* **2**, 113–130.

- [2] Bearwood, J., Halton, J.H. & Hammersley, J.M. (1959). The shortest path through many points, *Proceedings of the Cambridge Philosophical Society* **55**, 299–327.
- [3] Bienfait, B. & Gasteiger, J. (1997). Checking the projection display of multivariate data with colored graphs, *Journal of Molecular Graphics and Modeling* **15i**, 203–215.
- [4] Blough, D.M., Leoncini, M., Resta, G. & Santi, P. (2002). On the symmetric range assignment problem in wireless ad hoc networks, 2nd IFIP International Conference on Theoretical Computer Science, Montreal, Canada.
- [5] Chazelle, B. (2000). A minimum spanning tree algorithm with inverse Ackermann type complexity, *Journal of the ACM* **47**, 1028–1047.
- [6] Costa, J. & Hero, A.O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning, *IEEE Transactions on Signal Processing* **52**(8), 2210–2221.
- [7] Fredman, M.L. & Tarjan, R.E. (1987). Fibonacci heaps and their use in improved network optimization algorithms, *Journal of the ACM* **34**, 596–615.
- [8] Friedman, J. & Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, *Annals of Statistics* **7**, 697–717.
- [9] Friedman, J. & Rafsky, L. (1981). Graphics for the multivariate two-sample tests, *Journal of the American Statistical Association* **76**, 277–295.
- [10] Friedman, J. & Rafsky, L. (1981). Graph-theoretic measures of multivariate association and prediction, *Annals of Statistics* **11**, 377–391.
- [11] Gabow, H.N., Galil, Z., Spencer, T.H. & Tarjan, R.E. (1986). Efficient algorithms for minimum spanning trees on directed and undirected graphs, *Combinatorica* **6**, 109–122.
- [12] Gower, J.C. & Ross, G.J.S. (1969). Minimum spanning trees and single linkage clustering, *Applied Statistics* **18**, 54–64.
- [13] Graham, R.L. & Hell, P. (1985). On the history of the minimum spanning tree problem, *Annals of the History of Computing* **7**, 43–57.
- [14] Henze, N. & Penrose, M.D. (1999). On the multivariate runs test, *Annals of Statistics* **27**, 290–298.
- [15] Jolliffe, I.T., Jones, B. & Morgan, B.J.T. (1995). Identifying influential observations in hierarchical cluster analysis, *Journal of Applied Statistics* **22**, 61–80.
- [16] Kruskal, J.B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem, *Proceedings of the American Mathematical Society* **7**, 48–50.
- [17] Kwon, S. & Cook, D. (1998). Using a grand tour and minimal spanning tree to detect structure in high dimensions, *Computing Science and Statistics* **30**, 224–228.
- [18] Ma, B., Hero, A.O., Gorma, J. & Olivier, M. (2000). Image registration with minimum spanning tree algorithms, *Proceedings of IEEE International Conference on Image Processing* **1**, 481–484.

-
- [19] Moret, B.M.E. & Shapiro, H.D. (1994). An empirical assessment of algorithms for constructing a minimum spanning tree, *DIMACS Monographs in Discrete Mathematics and Theoretical Computer Science* **15**, 99–117.
 - [20] Nešetřil, J. (1997). Some remarks on the history of the MST problem, *Archivum Mathematicum* **33**, 15–22.
 - [21] Prim, R.C. (1957). The shortest connecting network and some generalizations, *Bell Systems Technical Journal* **36**, 1389–1401.
 - [22] Steele, M.J. (1988). Growth rates of Euclidean minimal spanning trees with power weighted edges, *Annals of Probability* **16**, 1767–1787.
 - [23] Tarjan, R.E. (1983). *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia.
 - [24] Toussaint, G. (1980). The relative neighborhood graph of a finite planar set, *Pattern Recognition* **13**, 261–268.
 - [25] Zahn, C.T. (1971). Graph theoretic methods for detecting and describing gestalt clusters, *IEEE Transactions on Computers* **1**, 68–86.

GEORGE MICHAILIDIS

Misclassification Rates

WOJTEK J. KRZANOWSKI

Volume 3, pp. 1229–1234

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Misclassification Rates

Many situations arise in practice in which individuals have to be allocated into one of a number of prespecified classes. For example, applicants to a bank for a loan may be segregated into two groups: 'repayers' or 'defaulters'; a psychiatrist may have to decide whether a patient is suffering from depression or not, and if yes, then which of a number of types of depressive illness it is most likely to be; and applicants for employment in a company may be classified as being either 'suitable' or 'not suitable' for the position by an industrial psychologist. Rather than merely relying on subjective expert assessment in such cases, use is often made of quantitative predictions via mathematical or computational formulae, termed *allocation rules* or *classifiers*, which are based on a battery of relevant measurements taken on each individual. Thus, the bank manager might provide the classifier with information on each individual's financial circumstances as given, say, by his/her income, savings, number and size of existing loans, mortgage commitments, and so on, as well as concomitant information on such factors as nature of employment, marital status, educational level attained, and similar. The psychiatrist would perhaps supply information on the patient's responses to questions about levels of concentration, loss of sleep, feelings of inadequacy, stress at work, and ability to make decisions; while the industrial psychologist might use numerical scores assessing the subject's letter of application, previous educational attainments, performance at interview, level of ambition, and so on.

The construction of such classifiers has a long history, dating from the 1930s. Original application areas were in the biological and agricultural sciences, but the methods spread rapidly to the behavioral and social sciences and are now used widely in most areas of human activity. Many different forms of classifier have been suggested and currently exist, including simple combinations of measured variables as in the *linear discriminant function* and the *quadratic discriminant function* (see **Discriminant Analysis**), more complicated explicit functions such as the *logistic discriminant function* and *adaptive regression splines* (see **Logistic Regression**; **Scatterplot Smoothers**), implicitly defined functions, and 'black box' routines such as *feed-forward neural networks* (see **Neural Networks**), tree-based methods such

as *decision trees* or **classification and regression trees**, numerically based methods including ones based on *kernels*, *wavelets*, or *nearest neighbors*, and dimensionality-based techniques such as *support vector machines*. The methodology underlying the construction of these classifiers is described elsewhere (see **Discriminant Analysis**). We will simply assume that the form of classifier has been chosen in a particular application, and will focus on the assessment of efficacy of that classifier. Since the objective of classification is to allocate individuals to preexisting classes, and any classification must by its nature be subject to error, the most obvious measure of efficacy of a classifier is the proportion of individuals that it is likely to misclassify. Hence, the estimation of misclassification rates is an important practical consideration.

In any given application, there will be a set of observations available for individuals whose true class membership is known, from which the classifier is to be built; this is known as the *training set*. For example, the bank manager will have data on the chosen variables for a set of known defaulters, as well as for those who are known to have repaid their loans. Likewise, the psychiatrist will have measured the relevant variables for patients known to be suffering from each type of depressive illness as well as individuals who are not ill, and the industrial psychologist will have a similarly relevant data set available. Once the form of classifier has been chosen, these data are used to estimate its parameters; this is known as *training* the classifier.

As a specific example, consider a set of data originally given by Reaven and Miller [9]. These authors were interested in studying the relationship between chemical subclinical and overt nonketotic diabetes in nonobese adult subjects. The three primary variables used in the analysis were glucose intolerance (x_1), insulin response to oral glucose (x_2), and insulin resistance (x_3), and all subjects in the study were assigned by medical diagnosis to one of the three classes 'normal', 'chemical diabetic' and 'overt diabetic'. For illustrative purposes, we will use a simple linear discriminant function to distinguish the first and third classes, so our training set consists of the 76 normal and 33 overt diabetic subjects. The resulting trained classifier is $y = -21.525 + 0.023x_1 + 0.017x_2 + 0.015x_3$, and future patients can be assigned to one or other class according to their observed value of y .

2 Misclassification Rates

The problem now is to evaluate the (future) performance of such a trained classifier, and this can be done by estimating misclassification rates from each of the classes (i.e., the *class-conditional* misclassification rates). If the classifier has been constructed using a theoretical derivation from an assumed probability model, then there may be a theoretical expression available from which the misclassification rates can be estimated. For example, if we are classifying individuals into one of two groups, and we can assume that the measured variables follow multivariate normal distributions in the two groups, with mean vectors μ_1, μ_2 , respectively, and a common dispersion matrix Σ , then standard theory [7, p. 59] shows that the optimal classifier in terms of minimizing costs due to misclassification is the linear discriminant function used in the example above. Moreover, under the assumption of equal costs from each group and equal prior probabilities of each group, the probability of misclassifying an individual from each group by this function is $\Phi\{-1/2[\mu_1 - \mu_2]^T \Sigma^{-1}[\mu_1 - \mu_2]\}$ where $\mathbf{x}^T$ denotes the transpose of the vector $\mathbf{x}$ and $\Phi\{u\}$ is the integral of the standard normal density function between minus infinity and u . The parameters μ_1, μ_2 and Σ can be estimated by the group mean vectors and pooled within-groups **covariance matrix** of the training data, and substituting these values in the expression above will thus yield an estimate of the misclassification rates. Doing this for the diabetes example, we estimate the misclassification rate from each class to be 0.018.

However, estimation of the misclassification rates in this way is heavily dependent on the distributional assumptions made (whereas the classifier itself may be more widely applicable). For example, it turns out that a linear discriminant function is a useful classifier in many situations, even when the data do not come from normal populations, but the assumption of normality is critical to the above estimates of misclassification rates, and these estimates can be wrong when the data are not normally distributed. Moreover, there are many classifiers (e.g., neural networks, classification trees) for which appropriate probability models are very difficult if not impossible to specify. Hence, in general, reliance on distributional assumptions is dangerous and the training data must be used in a more direct fashion to estimate misclassification rates.

The most intuitive approach is simply to apply the classifier to all training set members in turn, and

to estimate the misclassification rates by the proportion of training set members from each group that are misclassified. This is known as the *resubstitution* method of estimation, and the resulting estimates are often called the *apparent* error rates. However, it can be seen easily that this is a biased method that will generally be overoptimistic (i.e., will underestimate the misclassification rates). This is because the classifier has itself been built from the training data and has utilized any differences between the groups in an optimal fashion, thereby minimizing as far as possible any misclassifications. Future individuals for classification may well lie outside the bounds of the training set individuals or have some different characteristics, and so misclassifications are likely to be more frequent. The smaller the available data set, the more will the classifier be tailored to it and, hence, the more extreme will be the difference in performance on future individuals. In this case, we say that the training data have been *overfitted* and the classifier has poor *generalization*.

The only way to ensure unbiased estimates of misclassification rates is to use different sets of data for forming and assessing the classifier. This can be achieved by having a training set for building the classifier, and then an independent *test set* (drawn from the same source) for estimating its misclassification rates. If the original training set is very large, then it can be randomly split into two portions, one of which forms the classifier training set and the other the test set. The proportions of individuals misclassified from each group of the test set then form the required estimates. The final classifier for future use is then obtained from the combined set. However, the process of dividing the data into two sets raises problems. If the training set is not very large, then the classifier being assessed may be poorly estimated, and may then differ considerably from the one finally used on future individuals; while if the test set is not very large, then the estimates of misclassification rates are subject to small-sample volatility and may not be reliable. Of course, if the data sets are huge, then the scope for overfitting the training data is small and resubstitution will give very similar results to test set estimation, and the problems do not arise. However, in many practical situations, the available data set is small to moderate in size, so resubstitution will not be reliable while splitting the set into training, and test sets is not viable. In these cases, we need something a little more ingenious, and **cross-validation**,

jackknifing, and **bootstrapping** are three data-based methods that have been proposed to fill the gap.

Cross-validation

Cross-validation aims to steer a middle ground between splitting the data into just one training and one test set, and using all the data for both building the classifier and assessing it. In this method, the data are divided into a number k of subsets; each subset is used in turn as the test set for a classifier constructed from the remainder of the data, and the proportions of misclassified individuals in each group when results from all k subsets have been combined give the estimates of misclassification rates. The final step is to build the classifier for future individuals from the full data set. This method was first introduced in [6] for the case $k = 1$, and this case is now commonly called the *leave-one-out* estimate of misclassification rates, while the general case is usually termed the *k-fold cross-validation* estimate. Clearly, if $k = 1$ and each individual is left out in turn, then the final classifier for future individuals will not differ much from the classifiers used to allocate each omitted individual, but a lot of computing may be necessary to complete the process. On the other hand, if k is quite large, then there is much less computing but more discrepancy between the classifiers obtained during the assessment procedure and the one used for future classifications. The leave-one-out estimator has been studied quite extensively, and has been shown to be approximately unbiased but to have large variance. Nevertheless, it is popular in applications. When k greater than 1 is preferred, then values around 6 to 10 are commonly used.

Jackknifing

The jackknife was originally introduced in [8] as a general way of reducing the bias of an estimator, and its application to misclassification rate estimation came later. It uses the same operations as the leave-one-out variant of cross-validation, but with a different purpose; namely, to reduce the bias of the resubstitution estimator. Let us denote by $e_{r,i}$ the resubstitution estimate of misclassification rate for group i as obtained from a classifier built from the whole training set. Suppose now that the j th individual is omitted from the data, a new classifier is built from

the reduced data set, and $e_{r,i}^{(j)}$ is the resubstitution estimate of misclassification rate for this classifier in group i . If this process is repeated, leaving out each of the n individuals in the data in turn, the average of the n resubstitution estimates of misclassification rate in group i (each based on a sample of size $n - 1$) can be found as $\bar{e}_{r,i} = 1/n \sum_j e_{r,i}^{(j)}$. Then the jackknife estimate of bias in $e_{r,i}$ is $(n - 1)(e_{r,i} - \bar{e}_{r,i})$ so that the corrected estimate becomes $e_{r,i} + (n - 1)(e_{r,i} - \bar{e}_{r,i})$. For some more details, and some relationships between this estimator and the leave-one-out estimator, see [2].

Bootstrapping

An alternative approach to removal of the bias in the resubstitution estimator is provided by bootstrapping. The basic bootstrap estimator works as follows. Suppose that we draw a sample of size n *with replacement* from the training data. This will have repeats of some of the individuals in the original data while others will be absent from it; it is known as a *bootstrap sample*, and it plays the role of a potential sample from the population from which the training data came. A classifier is built from the bootstrap sample and its resubstitution estimate of misclassification rate for group i is computed; denote this by $e_{r,i}^b$. If this classifier is then applied to the original training data and the proportion of misclassified group i individuals is computed as $e_{r,i}^{cb}$, then this latter quantity is an estimate of the ‘true’ misclassification rate for group i using the classifier from the bootstrap sample. The difference $d_b = e_{r,i}^{cb} - e_{r,i}^b$ is then an estimate of the bias in the resubstitution misclassification rate. To smooth out the vagaries of individual samples, we take a large number of bootstrap samples and obtain the average difference $\bar{d}$ over these samples; this is the bias correction to add to $e_{r,i}$.

The bootstrap has also been used for direct estimation of error rates. A number of ways of doing this have been proposed (see [1] for example), but one of the most successful is the so-called 632 bootstrap. In this method, the estimate of misclassification rate for group i is given by $0.368e_{r,i} + 0.632e_i$, where e_i is the proportion of all points *not* in each bootstrap sample that were misclassified for group i , and $e_{r,i}$ is the resubstitution estimate as before. The value 0.632 arises because it is equal to $1 - e^{-1}$, which is

the approximate probability that a specified individual will be selected for the bootstrap sample.

Tuning

The above methods are the ones currently recommended for estimating misclassification rates, but in some complex situations, they may become embedded in the classifier itself. This can happen for a number of reasons. One typical situation occurs when some parameters of the classifier have to be optimized by finding those values that minimize the misclassification rates; this is known as *tuning* the classifier. An example of this usage is in multilayer perceptron neural networks, where the training process is iterative and some mechanism is necessary for deciding when to stop. The obvious point at which to stop is the one at which the network performance is optimized, that is, the one at which its misclassification rate is minimized. If the training data set is sufficiently large, then the network is trained on one portion of it, the misclassification rates are continuously estimated from a second portion, and the training stops when these rates reach a minimum. Note, however, that quoting this achieved minimum rate as the estimate of future performance of the network would be incorrect, sometimes badly so. In effect, the same mistake would be made here as was made by using the resubstitution estimate in the ordinary classifier; the classifier has been trained to minimize the misclassifications on the second set of available data, so the minimum achieved rate is an overoptimistic assessment of how it will fare on future data. To obtain an unbiased estimate, we need a *third* (independent) set of data on which to assess the trained network. So, ideally, we need to divide our data into three: one set for training the data, one set to decide when training stops, and a third set to assess the final network. Many overoptimistic claims were made in the early days of development of neural networks because this point was not appreciated.

Variable Selection

A related problem occurs when variable selection is an objective of the analysis, whatever classifier is used. Typically, a set of variables is deemed to be optimal if it yields a minimum number of misclassifications. The same comments as above are relevant in

this situation also, and simply quoting this minimum achieved rate will be an overoptimistic assessment of future performance of these variables. However, if data sets are small, then an additional problem frequently encountered is that the estimation of misclassification rates for deciding on variable subsets is effected by means of the leave-one-out process. If this is the case, then an unbiased assessment of the overall procedure will require a *nested* leave-one-out process: each individual is omitted from the data in turn, and then the *whole procedure* is conducted on the remaining individuals. If this procedure involves variable selection using leave-one-out to choose the variables, then this must be done, and only when it is completed is the first omitted individual classified. Then the next individual is omitted and the whole procedure is repeated, and so on. Thus, considerably more computing needs to be done than at first meets the eye, but this is essential if unbiased estimates are to be obtained. Comparative studies showing the bias incurred when these procedures are sidestepped have been described in [3] for tuning and [4] for variable selection.

Note that in all the discussion above we have focused on estimation of the *class-conditional* misclassification rates. The *overall* misclassification rate of a classifier is given by a weighted average of the class-conditional rates, where the weights are the prior probabilities of the classes. Often, these prior probabilities have to be estimated. There is no problem if the training data have been obtained by random sampling of the whole population (i.e., *mixture sampling*), as the proportions of each class in the training data then yield adequate estimates of the prior probabilities. However, the proportions cannot be used in this way whenever the investigator has controlled the class sizes (*separate sampling*, as for example in case-control studies). In such cases, additional external information is generally needed for estimating prior probabilities.

A final matter to note is that all the estimates discussed above are *point* estimates, that is, single-value 'guesses' at the true error rate of a given classifier. Such point estimates will always be subject to sampling variability, so an *interval* estimate might give a more realistic representation. This topic has not received much attention, but a recent study [5] has looked at coverage rates of various intervals generated by standard jackknife, bootstrap, and cross-validation methods. The conclusion was that 632

bootstrap error rates plus jackknife-after-bootstrap estimation of their standard errors gave the most reliable confidence intervals.

To illustrate some of these ideas, consider the results from [5] on discriminating between the 76 normal subjects and the 33 subjects suffering from overt nonketotic diabetes using the linear discriminant function introduced above. All data-based estimates of misclassification rates differed from the normal-based estimates of 0.018 from each class, suggesting that the data deviate from normality to some extent. The apparent error rates were 0.000 and 0.061 for the two groups respectively, while the bootstrap 632 rates for these groups were 0.002 and 0.091. This illustrates the overoptimism typical of the resubstitution process, but the bootstrap error rates, nevertheless, seem to indicate that the classifier is a good one with less than a 10% chance of misallocating to the 'harder' group of diabetics. However, when the variability in estimation is taken into account and 90% confidence intervals are calculated from the bootstrap rates, we find intervals of (0.000, 0.012) for the normal subjects and (0.000, 0.195) for the diabetics. Hence, classification to the diabetic group could be much poorer than first thought.

The interested reader can find more detailed discussion of the above topics, with allied lists of references, in [2, 7].

References

- [1] Efron, B. (1983). Estimating the error of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association* **78**, 316–331.
- [2] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*, John Wiley & Sons, Chichester.
- [3] Jonathan, P., Krzanowski, W.J. & McCarthy, W.V. (2000). On the use of cross-validation to assess performance in multivariate prediction, *Statistics and Computing* **10**, 209–229.
- [4] Krzanowski, W.J. (1995). Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis, *Computational Statistics and Data Analysis* **19**, 419–431.
- [5] Krzanowski, W.J. (2001). Data-based interval estimation of classification error rates, *Journal of Applied Statistics* **28**, 585–595.
- [6] Lachenbruch, P.A. & Mickey, M.R. (1968). Estimation of error rates in discriminant analysis, *Technometrics* **10**, 1–11.
- [7] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York.
- [8] Quenouille, M. (1956). Notes on bias in estimation, *Biometrika* **43**, 353–360.
- [9] Reaven, G.M. & Miller, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia* **16**, 17–24.

WOJTEK J. KRZANOWSKI

Missing Data

RODERICK J. LITTLE

Volume 3, pp. 1234–1238

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Missing Data

Introduction

Missing values arise in behavioral science data for many reasons: **dropouts in longitudinal studies** (see **Longitudinal Data Analysis**), unit nonresponse in sample surveys where some individuals are not contacted or refusal to respond, refusal to answer particular items in a questionnaire, missing components in an index variable constructed by summing values of particular items. Missing data can also arise by design. For example, suppose one objective in a study of obesity is to estimate the distribution of a measure Y_1 of body fat in the population, and correlate it with other factors. Suppose Y_1 is expensive to measure but a proxy measure Y_2 , such as body mass index, which is a function of height and weight, is inexpensive to measure. Then it may be useful to measure Y_2 and covariates, X , for a large sample and Y_1 , Y_2 and X for a smaller subsample. The subsample allows predictions of the missing values of Y_1 to be generated for the larger sample, yielding more efficient estimates than are possible from the subsample alone.

Unless missing data are a deliberate feature of the study design, the most important step in dealing with missing data is to try to avoid it during the data-collection stage. Since data are still likely to be missing despite these efforts, it is important to try to collect covariates that are predictive of the missing values, so that an adequate adjustment can be made. In addition, the process that leads to missing values should be determined during the collection of data if possible, since this information helps to model the missing-data mechanism when an adjustment for the missing values is performed [2].

We distinguish three major approaches to the analysis of missing data:

1. Discard incomplete cases and analyze the remainder (complete-case analysis), as discussed in the section titled Complete-case, Available-case, and Weighting Analysis”;
2. Impute or fill in the missing values and then analyze the filled-in data, as discussed in the sections titled Single Imputation and Multiple Imputation;
3. Analyze the incomplete data by a method that does not require a complete (that is, a rectangular) data set, as discussed in the sections

titled Maximum Likelihood for Ignorable Models, Maximum Likelihood for Nonignorable Models, and Bayesian Simulation Methods.

A basic assumption in all our methods is that missingness of a particular value hides a true underlying value that is meaningful for analysis. This may seem obvious but is not always the case. For example, in a study of a behavioral intervention for people with heart disease, it is not meaningful to consider a quality of life measure to be missing for subjects who die prematurely during the course of the study. Rather, it is preferable to restrict the analysis to the quality of life measures of individuals who are alive.

Let $Y = (y_{ij})$ denote an $(N \times p)$ rectangular dataset without missing values, with i th row $y_i = (y_{i1}, \dots, y_{ip})$ where y_{ij} is the value of variable Y_j for subject i . With missing values, the *pattern* of missing data is defined by the *missing-data indicator matrix* $M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. An important example of a special pattern is *univariate* nonresponse, where missingness is confined to a single variable. Another is *monotone* missing data, where the variables can be arranged so that $Y_{j+1}, \dots, Y_p$ are missing for all cases where Y_j is missing, for all $j = 1, \dots, p - 1$. This pattern arises commonly in longitudinal data subject to attrition, where once a subject drops out, no more data are observed. Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume a special pattern.

The missing-data mechanism addresses the reasons why values are missing, and in particular, whether these reasons relate to values in the data set. For example, subjects in a longitudinal intervention may more likely to drop out of a study because they feel the treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin [5] treated M as a random matrix, and characterized the missing-data mechanism by the conditional distribution of M given Y , say $f(M|Y, \phi)$, where ϕ denotes unknown parameters. Assume independent observations, and let m_i and y_i denote the rows of M and Y corresponding to individual i . When missingness for case i does not depend on the values of the data, that is,

$$f(m_i|y_i, \phi) = f(m_i|\phi) \quad \text{for all } y_i, \phi, \quad (1)$$

the data are called *missing completely at random* (MCAR). An MCAR mechanism is plausible in

2 Missing Data

planned missing-data designs, but is a strong assumption when missing data do not occur by design, because missingness often does depend on recorded variables. A less restrictive assumption is that missingness depends only on data that are observed, say $y_{\text{obs},i}$, and not on data that are missing, say $y_{\text{mis},i}$; that is,

$$f(m_i|y_i, \phi) = f(m_i|y_{\text{obs},i}, \phi) \quad \text{for all } y_{\text{mis},i}, \phi. \quad (2)$$

The missing-data mechanism is then called *missing at random* (MAR). Many methods for handling missing data assume the mechanism is MCAR, and yield biased estimates when the data are not MCAR. Better methods rely only on the MAR assumption.

Complete-case, Available-case, and Weighting Analysis

A common default approach is complete-case (CC) analysis, also known as *listwise* deletion, where incomplete cases are discarded and standard analysis methods applied to the complete cases [3, Section 3.2]. In many statistical packages, this is the default analysis. When the missing data are MCAR, the complete cases are a random subsample of the original sample, and CC analysis results in valid (but often inefficient) inferences. If the data are not MCAR then the complete cases are a biased sample, and CC analysis is often (though not always) biased, with a bias that depends on the degree of departure from MCAR, the amount of missing data, and the specifics of the analysis. The potential for bias is why sample surveys with high rates of unit nonresponse (say 30% or more) are often considered unreliable for making inferences to the whole population.

A modification of CC analysis, commonly used to handle unit nonresponse in surveys, is to weight respondents by the inverse of an estimate of the probability of response [3, Section 3.3]. A simple approach to estimation is to form adjustment cells (or subclasses) on the basis of background variables and weight respondents in an adjustment cell by the inverse of the response rate in that cell. This method eliminates bias if respondents within each adjustment cell respondents can be regarded as a random subsample of the original sample in that cell (i.e., the data are MAR given indicators for the

adjustment cells). A useful extension with extensive background information is *response propensity* stratification (see **Propensity Score**), where adjustment cells are based on similar values of estimates of probability of response, computed by a logistic regression of the indicator for missingness on the background variables. Although weighting methods can be useful for reducing nonresponse bias, they are potentially inefficient.

Available-case (AC) analysis [3, Section 3.4] is a straightforward attempt to exploit the incomplete information by using all the cases available to estimate each individual parameter. For example, suppose the objective is to estimate the correlation matrix of a set of continuous variables $Y_1, \dots, Y_p$. AC analysis uses all the cases with both Y_j and Y_k observed to estimate the correlation of Y_j and Y_k , $1 \leq j, k \leq p$. Since the sample base of available cases for measuring each correlation includes the set of complete cases, the AC method appears to make better use of available information. The sample base changes from correlation to correlation, however, creating potential problems when the missing data are not MCAR. In the presence of high correlations, AC can be worse than CC analysis, and there is no guarantee that the AC correlation matrix is even positive definite.

Single Imputation

Methods that impute or fill in the missing values have the advantage that, unlike CC analysis, observed values in the incomplete cases are retained. A simple version imputes missing values by their unconditional sample means based on the observed data, but this method often leads to biased inferences and cannot be recommended in any generality [3, Section 4.2.1]. An improvement is *conditional mean* imputation [3, Section 4.2.2], in which each missing value is replaced by an estimate of its conditional mean given the values of observed values. For example, in the case of univariate nonresponse with $Y_1, \dots, Y_{p-1}$ fully observed and Y_p sometimes missing, regression imputation estimates the regression of Y_p on $Y_1, \dots, Y_{p-1}$ from the complete cases, and the resulting prediction equation is used to impute the estimated conditional mean for each missing value of Y_p . For a general pattern of missing data, the missing values for each case can be imputed from the regression of the missing variables on the observed

variables, computed using the set of complete cases. Iterative versions of this method lead (with some important adjustments) to maximum likelihood estimates under multivariate normality [3, Section 11.2].

Although conditional mean imputation yields best predictions of the missing values in the sense of mean squared error, it leads to distorted estimates of quantities that are not linear in the data, such as percentiles, variances, and correlations. A solution to this problem is to impute random draws from the predictive distribution rather than best predictions. An example is *stochastic regression* imputation, in which each missing value is replaced by its regression prediction plus a random error with variance equal to the estimated residual variance. Another approach is *hot-deck* imputation, which classifies respondents and nonrespondents into adjustment cells based on similar values of observed variables and then imputes values for nonrespondents from randomly chosen respondents in the same cell. A more general approach to hot-deck imputation matches nonrespondents and respondents using a distance metric based on the observed variables. For example, *predictive mean matching* imputes values from cases that have similar predictive means from a regression of the missing variable on observed variables. This method is somewhat robust to misspecification of the regression model used to create the matching metric [3, Section 4.3].

The imputation methods discussed so far assume the missing data are MAR. In contrast, models that are not missing at random (NMAR) assert that even if a respondent and nonrespondent to Y_p appear identical with respect to observed variables $Y_1, \dots, Y_{p-1}$, their Y_p -values differ systematically. A crucial point about the use of NMAR models is that there is never direct evidence in the data to address the validity of their underlying assumptions. Thus, whenever NMAR models are being considered, it is prudent to consider several NMAR models and explore the sensitivity of analyses to the choice of model [3, Chapter 15].

Multiple Imputation

A serious defect with imputation is that a single imputed value cannot represent the uncertainty about, which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled

correctly. **Multiple imputation** (MI, [6, 7]) fixes this problem by creating a set of Q (say $Q = 5$ or 10) draws for each missing value from a predictive distribution, and thence Q datasets, each containing different sets of draws of the missing values. We then apply the standard complete-data analysis to each of the Q datasets and combine the results in a simple way. In particular, for scalar estimands, the MI estimate is the average of the estimates from the Q datasets, and the variance of the estimate is the average of the variances from the five datasets plus $1 + 1/Q$ times the sample variance of the estimates over the Q datasets (The factor $1 + 1/Q$ is a small- Q correction). The last quantity here estimates the contribution to the variance from imputation uncertainty, missed (i.e., set to zero) by single imputation methods. Another benefit of multiple imputation is that the averaging over datasets results in more efficient point estimates than does single random imputation. Often, MI is not much more difficult than doing a single imputation – the additional computing from repeating an analysis Q times is not a major burden and methods for combining inferences are straightforward. Most of the work is in generating good predictive distributions for the missing values.

Maximum Likelihood for Ignorable Models

Maximum likelihood (ML) avoids imputation by formulating a statistical model and basing inference on the likelihood function of the incomplete data [3, Section 6.2]. Define Y and M as above, and let X denote an $(n \times q)$ matrix of fixed covariates, assumed fully observed, with i th row $x_i = (x_{i1}, \dots, x_{iq})$ where x_{ij} is the value of covariate X_j for subject i . Covariates that are not fully observed should be treated as random variables and modeled with the set of Y_j 's.

Two likelihoods are commonly considered. The *full* likelihood is obtained by integrating the missing values out of the joint density of Y and M given X . The *ignorable* likelihood is obtained by integrating the missing values out of the joint density of Y given X , ignoring the distribution of M . The ignorable likelihood is easier to work with than the full likelihood, since it is easier to handle computationally, and more importantly because it avoids the need to specify a model for the missing-data mechanism, about

which little is known in *many situations*. Rubin [5] showed that valid inferences about parameters can be based on the ignorable likelihood when the data are MAR, as defined above. (A secondary condition, *distinctness*, is needed for these inferences to be fully efficient). Large-sample inferences about parameters can be based on standard ML theory [3, Chapter 6].

In many problems, maximization of the likelihood requires numerical methods. Standard optimization methods such as Newton–Raphson or Scoring can be applied. Alternatively, we can apply the *Expectation–Maximization (EM)* algorithm [1] or one of its extensions [3, 4]. Reference [3] includes many applications of EM to particular models, including normal and t models for multivariate continuous data, **log-linear models** for multiway contingency tables, and the general location model for mixtures of continuous and categorical variables. Asymptotic standard errors are not readily available from EM, unlike numerical methods like scoring or Newton–Raphson. A simple approach is to use the **bootstrap** or the **jackknife** method [3, Chapter 5]. An alternative is to switch to a Bayesian simulation method that simulates the posterior distribution of the parameters (see the section titled Bayesian Simulation Methods).

Maximum Likelihood for Nonignorable Models

Nonignorable, non–MAR models apply when missingness depends on the missing values (see the section titled Introduction). A correct likelihood analysis must be based on the full likelihood from a model for the joint distribution of Y and M . The standard likelihood asymptotics apply to nonignorable models providing the parameters are identified, and computational tools such as EM also apply to this more general class of models. However, often information to estimate both the parameters of the missing-data mechanism and the parameters of the complete-data model is very limited, and estimates are sensitive to misspecification of the model. Often, a **sensitivity analysis** is needed to see how much the answers change for various assumptions about the missing-data mechanism [[3], Chapter 15].

Bayesian Simulation Methods

Maximum likelihood is most useful when sample sizes are large, since then the log-likelihood is nearly

quadratic and can be summarized well using the ML estimate θ and its large sample variance–covariance matrix. When sample sizes are small, a useful alternative approach is to add a prior distribution for the parameters and compute the posterior distribution of the parameters of interest. Since the posterior distribution rarely has a simple analytic form for incomplete-data problems, simulation methods are often used to generate draws of θ from the posterior distribution $p(\theta|Y_{obs}, M, X)$. Data augmentation [10] is an iterative method of simulating the posterior distribution of θ that combines features of the EM algorithm and multiple imputation, with Q imputations of each missing value at each iteration. It can be thought of as a small-sample refinement of the EM algorithm using simulation, with the imputation step corresponding to the E-step (random draws replacing expectation) and the posterior step corresponding to the M-step (random draws replacing MLE). When Q is set equal to one, data augmentation is a special case of the Gibbs’ sampler. This algorithm can be run independently Q times to generate M iid draws from the approximate joint posterior distribution of the parameters and the missing data. The draws of the missing data can be used as multiple imputations, yielding a direct link with the methods in the section titled Multiple Imputation.

Conclusion

In general, the advent of modern computing has made more principled methods of dealing with missing data practical, such as multiple imputation (e.g. PROC MI in [8]) or ignorable ML for repeated measures with dropouts (see **Dropouts in Longitudinal Studies: Methods of Analysis**). See [3] or [9] to learn more about missing-data methodology.

Acknowledgments

This research was partially supported by National Science Foundation Grant DMS 9803720.

References

- [1] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society. Series B* **39**, 1–38.

-
- [2] Little, R.J.A. (1995). Modeling the drop-out mechanism in longitudinal studies, *Journal of the American Statistical Association* **90**, 1112–1121.
- [3] Little, R.J.A. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd Edition, John Wiley, New York.
- [4] McLachlan, G.J. & Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.
- [5] Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581–592.
- [6] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- [7] Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489; with discussion, 507–515; and rejoinder, 515–517.
- [8] SAS. (2003). *SAS/STAT Software, Version 9*, SAS Institute, Inc., Cary.
- [9] Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.
- [10] Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association* **82**, 528–550.
- (See also **Dropouts in Longitudinal Studies: Methods of Analysis**)

RODERICK J. LITTLE

Model Based Cluster Analysis

BRIAN S. EVERITT

Volume 3, pp. 1238–1238

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Model Based Cluster Analysis

Many methods of cluster analysis are based on *ad hoc* ‘sorting’ algorithms rather than some sound statistical model (see **Hierarchical Clustering; *k*-means Analysis**). This makes it difficult to evaluate and assess solutions in terms of a set of explicit assumptions. More statistically respectable are model-based methods in which an explicit model for the clustering process is specified, containing parameters defining the characteristics of the assumed clusters. Such an approach replaces ‘sorting’ with estimation and may lead to formal

tests of cluster structure, particularly for **number of clusters**.

The most well-known type of model-based clustering procedure uses **finite mixture distributions**. Other possibilities are described in [2] and include the classification likelihood approach [1] and **latent class analysis**.

References

- [1] Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**, 803–822.
- [2] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th Edition, Arnold Publishing, London.

BRIAN S. EVERITT

Model Evaluation

DANIEL J. NAVARRO AND JAY I. MYUNG

Volume 3, pp. 1239–1242

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Model Evaluation

Introduction

The main reason for building models is to link theoretical ideas to observed data, and the central question that we are interested in is ‘Is the model any good’? When dealing with quantitative models, we can at least partially answer this question using statistical tools. Before going into detail, there is a touchy, even philosophical, issue that one cannot ignore. A naive view of modeling is to identify the underlying process (truth) that has actually generated the data. This is an ill-posed problem, meaning that the solution is nonunique. The finite data sample rarely contains sufficient information to lead to a single process and also, is corrupted by unavoidable random noise, blurring the identification. An implication of noise-corrupted data is that it is not in general possible to determine with complete certainty that what we are fitting is the regularity, which we are interested in, or the noise, which we are not. A model that assumes a certain amount of error is present may be worse than a yet to be postulated model that can explain more of what we thought of as error in the first model. In short, identifying the true model based on data samples is an unachievable goal. Furthermore, the ‘truth’ of any phenomenon is likely to be rather different from any proposed model. Ultimately, it is crucial to recognize that *all* models are wrong, and a realistic goal of modeling is to find a model that represents a ‘good’ approximation to the truth in a statistically defined sense.

In what follows, we assume that we have a model $\mathcal{M}$ with k free parameters $\theta = (\theta_1, \dots, \theta_k)$, and a data set that consists of n observations $X = (X_1, \dots, X_n)$. Quantitative models generally come in two main types: They either assign some probability to the observed data $f(X|\theta)$ (probabilistic models), or they produce a single predicted data set $X^{prd}(\theta)$ (deterministic models). We should note that most model testing and model selection methods require a probabilistic formulation, so it is commonplace to define a model as $\mathcal{M} = \{f(X|\theta) | \theta \in \Omega\}$ where Ω is the parameter space. When written in this form, a model can be conceptualized as a family of probability distributions.

Model Fitting

At a minimum, any reasonable model needs to be able to mimic the structure of the data: It needs to be able to ‘fit’ the data. When measuring the goodness of a model’s fit, we find the parameter values that allow the model to best mimic the data, denoted $\hat{\theta}$. The two most common methods for this are **maximum likelihood estimation** (for probabilistic models) and **least squares estimation** (for deterministic models). In the maximum likelihood approach, introduced by **Sir Ronald Fisher** in the 1920s, $\hat{\theta}$ is the set of parameter values that maximizes $f(X|\theta)$, and is referred to as the maximum likelihood estimate (MLE). The corresponding measure of fit is the maximized log-likelihood $\hat{L} = \ln f(X|\hat{\theta})$. See [3] for a tutorial on maximum likelihood estimation with example applications in cognitive psychology.

Alternatively, the least squares estimate of $\hat{\theta}$ is the set of parameters that minimizes the *sum of squared errors* (SSE), and the minimized SSE value is denoted by $\hat{E}$:

$$\hat{E} = \sum_{i=1}^n (X_i - X_i^{prd}(\hat{\theta}))^2. \quad (1)$$

When this approach is employed, there are several commonly used measures of fit. They are *mean-squared error* $MSE = \hat{E}/n$, *root mean-squared deviation* $RMSD = \sqrt{(\hat{E}/n)}$, and *squared correlation* (also known as proportion of variance accounted for) $r^2 = 1 - \hat{E}/SST$. In the last formula, SST stands for the *sum of squares total* defined as $SST = \sum_{i=1}^n (X_i - \bar{X})^2$, where $\bar{X}$ denotes the mean of X . There is a nice correspondence between maximum likelihood and least squares, in that for a model with independent, identically and normally distributed errors, the same set of parameters is obtained as the one that maximizes the log-likelihood L but also minimizes the sum of squared errors SSE .

Model fitting yields goodness-of-fit measures, such as $\hat{L}$ or $\hat{E}$, that tell us how well the model fits the observed data sample but by themselves are not particularly meaningful. If our model has a minimized sum squared error of 0.132, should we be impressed or not? In other words, a goodness-of-fit measure may be useful as a purely descriptive measure, but by itself is not amenable to statistical inference. This is because the measure does not address the relevant

2 Model Evaluation

question: ‘Does the model provide an *adequate* fit to the data, in a defined sense’?. This question is answered in model testing.

Model Testing

Classical null hypothesis testing is a standard method of judging a model’s goodness of fit. The idea is to set up a null hypothesis that ‘the model is correct’, obtain the P value, and then make a decision about rejecting or retaining the hypothesis by comparing the resulting P value with the alpha level.

For discrete data such as frequency counts, the two most popular methods are the Pearson chi-square (χ^2) test and the log-likelihood ratio test (G^2), which have test statistics given by

$$\begin{aligned}\chi^2 &= \sum_{i=1}^n \frac{(X_i - X_i^{prd}(\hat{\theta}))^2}{X_i^{prd}(\hat{\theta})}; \\ G^2 &= -2 \sum_{i=1}^n X_i \ln \frac{X_i^{prd}(\hat{\theta})}{X_i},\end{aligned}\quad (2)$$

where $\ln$ is the natural logarithm of base e . Both of these statistics have the nice property that they are always nonnegative, and are equal to zero when the observed and predicted data are in full agreement. In other words, the larger the statistic, the greater the discrepancy. Under the null hypothesis, both are approximately distributed as a χ^2 distribution with $(n - k - 1)$ degrees of freedom, so we would reject the null if the obtained P value is larger than some critical value obtained by setting an appropriate α level.

For continuous data such as response time, goodness-of-fit tests are a little more complicated, since there are no general-purpose methods available for testing the validity of a single model, unlike the discrete case. Instead, we rely on the *generalized likelihood ratio test* that involves two models. In this test, in addition to the theoretically motivated model, denoted by $\mathcal{M}_r$ (reduced model), we create a second model, $\mathcal{M}_f$ (full model), such that the reduced model is obtained as a special case of the full model by fixing one or more of $\mathcal{M}_f$ ’s parameters. Then, the goodness of fit of the reduced model is assessed by the following G^2 statistic:

$$G^2 = 2(\ln \hat{L}_f - \ln \hat{L}_r), \quad (3)$$

recalling that $\hat{L}$ denotes the maximized log-likelihood. Under the null hypothesis that the theoretically motivated, reduced model is correct, the above statistic is approximately distributed as χ^2 with degrees of freedom given by the difference in the number of parameters ($k_f - k_r$). If the hypothesis is retained (rejected), then we conclude that the reduced model $\mathcal{M}_r$ provides (does not provide) an adequate description of the data (see [4] for an example application of this test).

Model Selection

What does it mean that a model provides an adequate fit of the data? One should not jump to the conclusion that one has identified the underlying regularity. A good fit merely puts the model on a list of candidate models worthy of further consideration. It is entirely possible that there are several distinct models that fit the data well, all passing goodness of fit tests. How should we then choose among such models? This is the problem of model selection.

In model selection, the goal is to select the one, among a set of candidate models, that represents the closest approximation to the underlying process in some defined sense. Choosing the model that best fits a particular set of observed data will not accomplish the goal. This is because a model can achieve a superior fit to its competitors for reasons unrelated to the model’s exactness. For instance, it is well known that a complex model with many parameters and highly nonlinear form can often fit data better than a simple model with few parameters even if the latter generated the data. This is called *overfitting*.

Avoiding overfitting is what every model selection method is set to accomplish. The essential idea behind modern model selection methods is to recognize that, since data are inherently noisy, an ideal model is one that captures only the underlying phenomenon, not the noise. Since noise is idiosyncratic to a particular data set, a model that captures noise will make poor predictions about future events. This leads to the present-day ‘gold standard’ of model selection, **generalizability**. Generalizability, or predictive accuracy, refers to a model’s ability to predict the statistics of future, as yet unseen, data samples from the same process that generated the observed data sample.

The intuitively simplest way to measure generalizability is to estimate it directly from the data, using

cross-validation (CV; [10]). In cross-validation, we split the data set into two samples, the calibration sample X_c and the test sample X_t . We first estimate the best-fitting parameters by fitting the model to X_c which we denote $\hat{\theta}(X_c)$. The generalizability estimate is obtained by measuring the fit of the model to the test sample at those original parameters, that is, $\hat{\theta}(X_c)$,

$$CV = \ln f(X_t | \hat{\theta}(X_c)). \quad (4)$$

The main attraction of CV is its ease of implementation (see [4] for its application example for psychological models). All that is required is a model fitting procedure and a resampling scheme. One concern with CV is that there is a possibility that the test sample is not truly independent of the calibration sample: Since both were produced in the same experiment, systematic sources of error variation are likely to induce correlated noise across the two samples, artificially inflating the CV measure.

An alternative approach is to use theoretical measures of generalizability based on a single sample. In most of these theoretical approaches, generalizability is measured by suitably combining goodness-of-fit with model complexity. The practical difference between them is the way in which complexity is measured. One of the earliest measures of this kind was the **Akaike information criterion** (AIC; [1]), which treats complexity as the number of parameters k :

$$AIC = -2 \ln f(X | \hat{\theta}) + 2k, \quad (5)$$

The method prescribes that the model minimizing AIC should be chosen. AIC seeks to find the model that lies ‘closest’ to the true distribution, as measured by the Kullback–Leibler [8] discrepancy. As shown in the above criterion equation, this is achieved by trading the first, minus goodness-of-fit (lack of fit) term of the right hand side for the second complexity term. As such, a complex model with many parameters, having a large value of the complexity term, will not be selected unless its fit justifies the extra complexity. In this sense, AIC represents a formalization of the principle of Occam’s razor, which states ‘Entities should not be multiplied beyond necessity’ (William of Occam, ca. 1290–1349) (see **Parsimony/Occam’s Razor**).

Another approach is given by the much older notion of **Bayesian statistics**. In the Bayesian approach, we assume that *a priori* uncertainty about

the value of model parameters is represented by a prior distribution $\pi(\theta)$. Upon observing the data X , this prior is updated, yielding a posterior distribution $\pi(\theta | X) \propto f(X | \theta) \pi(\theta)$. In order to make inferences about the model (rather than its parameters), we integrate across the posterior distribution. Under the assumption that all models are *a priori* equally likely (because the Bayesian approach requires model priors as well as parameter priors), Bayesian model selection chooses the model $\mathcal{M}$ with highest marginal likelihood defined as:

$$f(X | \mathcal{M}) = \int f(X | \theta) \pi(\theta) d\theta. \quad (6)$$

The ratio of two marginal likelihoods is called a *Bayes factor* (BF; [2]), which is a widely used method of model selection in Bayesian inference. The two integrals in the Bayes factor are nontrivial to compute unless $f(X | \theta)$ and $\pi(\theta)$ form a conjugate family. Monte Carlo methods are usually required to compute BF, especially for highly parameterized models (see **Markov Chain Monte Carlo and Bayesian Statistics**). A large sample approximation of BF yields the easily computable Bayesian information criterion (BIC; [9])

$$BIC = -2 \ln f(X | \hat{\theta}) + k \ln n. \quad (7)$$

The model minimizing BIC should be chosen. It is important to recognize that the BIC is based on a number of restrictive assumptions. If these assumptions are met, then the difference between two BIC values approaches twice the logarithm of the Bayes factor as n approaches infinity.

A third approach is minimum description length (MDL; [5]), which originates in algorithmic coding theory. In MDL, a model is viewed as a code that can be used to compress the data. That is, data sets that have some regular structure can be compressed substantially if we know what that structure is. Since a model is essentially a hypothesis about the nature of the regularities that we expect to find in data, a good model should allow us to compress the data set effectively. From an MDL standpoint, we choose the model that permits the greatest compression of data in its total description: That is, the description of data obtainable with the help of the model plus the description of the model itself. A series of papers by Rissanen expanded on and refined this idea, yielding a number of different model selection criteria (one

of which was essentially identical to the BIC). The most complete MDL criterion currently available is the *stochastic complexity* (SC; [7]) of the data relative to the model,

$$SC = -\ln f(X|\hat{\theta}) + \ln \int f(Y|\hat{\theta}(Y)) dY. \quad (8)$$

Note that the second term of SC represents a measure of model complexity. Since the integral over the sample space is generally nontrivial to compute, it is common to use the Fisher-information approximation (FIA; [6]): Under regularity conditions, the stochastic complexity asymptotically approaches

$$FIA = -\ln f(X|\hat{\theta}) + \frac{k}{2} \ln \left(\frac{n}{2\pi} \right) + \ln \int_{\Theta} \sqrt{\det I(\theta)} d\theta, \quad (9)$$

where $I(\theta)$ is the expected Fisher **information matrix** of sample size one, consisting of the covariances between the partial derivatives of L with respect to the parameters. Once again, the integral can still be intractable, but it is generally easier to calculate than the exact SC. As in AIC and BIC, the first term of FIA is the lack of fit term and the second and third terms together represent a complexity measure. From the viewpoint of FIA, complexity is determined by the number of free parameters (k) and sample size (n) but also by the ‘functional form’ of the model equation, as implied by the Fisher information $I(\theta)$, and the range of the parameter space Θ .

When using generalizability measures, it is important to recognize that AIC, BIC, and FIA are all *asymptotic* criteria, and are only guaranteed to work as n becomes arbitrarily large, and when certain regularity conditions are met. The AIC and BIC, in particular, can be misleading for small n . The FIA is safer (i.e., the error level generally falls faster as n increases), but it too can still be misleading in some cases. The SC and BF criteria are more sensitive, since they are exact rather than asymptotic criteria, and can be quite powerful even when presented with very similar models or small samples. However, they can be difficult to employ, and often need to be approximated numerically. The status of CV is a little more complicated, since it is not always clear what CV is doing, but its performance in practice is often better than AIC or BIC, though it is not usually as good as SC, FIA, or BF.

Conclusion

When evaluating a model, there are a number of factors to consider. Broadly speaking, statistical methods can be used to measure the descriptive adequacy of a model (by fitting it to data and testing those fits), as well as its generalizability and simplicity (using model selection tools). However, the strength of the underlying theory also depends on its interpretability, its consistency with other findings, and its overall plausibility. These things are inherently subjective judgments, but they are no less important for that. As always, there is no substitute for thoughtful evaluations and good judgment. After all, statistical evaluations are only one part of a good analysis.

References

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, B.N. Petrov & F. Csaki, eds, Akademiai Kiado, Budapest, pp. 267–281.
- [2] Kass, R.E. & Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**(430), 773–795.
- [3] Myung, I.J. (2003). Tutorial on maximum likelihood estimation, *Journal of Mathematical Psychology* **47**, 90–100.
- [4] Myung, I.J. & Pitt, M.A. (2002). Mathematical modeling, in *Stevens' Handbook of Experimental Psychology*, 3rd Edition, J. Wixten, ed., John Wiley & Sons, New York, pp. 429–459.
- [5] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore.
- [6] Rissanen, J. (1996). Fisher information and stochastic complexity, *IEEE Transactions on Information Theory* **42**, 40–47.
- [7] Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data, *IEEE Transactions on Information Theory* **47**, 1712–1717.
- [8] Schervish, M.J. (1995). *Theory of Statistics*, Springer, New York.
- [9] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**(2), 461–464.
- [10] Stone, M. (1974). Cross-validated choice and assessment of statistical predictions, *Journal of Royal Statistical Society Series B* **36**, 111–147.

DANIEL J. NAVARRO AND JAY I. MYUNG

Model Fit: Assessment of

CEES A.W. GLAS

Volume 3, pp. 1243–1249

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Model Fit: Assessment of

Introduction

Item response theory (IRT) models (*see* **Item Response Theory (IRT) Models for Dichotomous Data**) provide a useful and well-founded framework for measurement in the social sciences. However, the applications of these models are only valid if the model fits the data. IRT models are based on a number of explicit assumptions, so methods for the evaluation of model fit focus on these assumptions. Model fit can be viewed from two perspectives: the items and the respondents. In the first case, for every item, residuals (differences between predictions from the estimated model and observations) and item fit statistics are computed to assess whether the item violates the model. In the second case, residuals and person-fit statistics are computed for every person to assess whether the responses to the items follow the model.

The most important assumptions evaluated are subpopulation invariance (the violation is often labeled **differential item functioning**), the form of the item response function, and local stochastic independence. The first assumption entails that the item responses can be described by the same parameters in all possible subpopulations. Subpopulations are defined on the basis of background variables that should not be relevant in a specific testing situation. One might think of gender, race, age, or socioeconomic status. The second assumption addressed is the form of the item response function that describes the relation between a **latent variable**, say proficiency, and observable responses to items. Evaluation of the appropriateness of the item response function is usually done by comparing observed and expected item response frequencies given some measure of the latent trait level. The third assumption targeted is local stochastic independence. The assumption entails that responses to different items are independent given the latent trait value. So the proposed latent variables completely describe the responses and no additional variables are necessary to describe the responses.

A final remark in this introduction pertains to the relation between formal tests of model fit and residual analyses. A well-known problem with formal tests of model fit is that they tend to reject the model even

for moderate sample sizes. That is, their **power** (the probability of rejection when the model is violated) grows very fast as a function of the sample size. As a result, small deviations from the IRT model may cause a rejection of the model while these deviations may hardly have practical consequences in the foreseen application of the model. Inspection of the magnitude of the residuals can shed light on the severity of the model violation. The reason for addressing problem of evaluation of model fit in the framework of formal model tests is that the alternative hypotheses in these model tests clarify which model assumptions are exactly targeted by the residuals. This will be explained further below.

We will start with describing evaluation of fit to IRT models for dichotomous items, then a general approach to evaluation of model fit for a general class of IRT models will be outlined, and finally, a small example will be given. The focus will be on parameterized IRT models in a likelihood-based framework. The relation with other approaches to IRT will be briefly sketched in the conclusion.

Assessing Model Fit for Items with Dichotomous Responses

In the 1-, 2-, and 3-parameter logistic models (1PLM, 2PLM and 3PLM), [2] it is assumed that the proficiency level of a respondent (indexed i) can be represented by a one-dimensional proficiency parameter θ_i . In the 3PLM, the probability of a correct response as a function of θ_i is given by

$$\begin{aligned} \Pr(Y_{ik} = 1 | \theta_i, a_k, b_k, c_k) \\ = P_k(\theta_i) = c_k + (1 - c_k) + \frac{\exp(a_k(\theta_i - b_k))}{1 + \exp(a_k(\theta_i - b_k))}, \end{aligned} \quad (1)$$

where Y_{ik} is a random variable assuming a value equal to one if a correct response was given to item k , and zero otherwise. The three item parameters, a_k , b_k , and c_k are called the *discrimination*, *difficulty* and *guessing parameter*, respectively. The 2PLM follows upon setting the guessing parameter c_k equal to zero, and the 1PLM follows upon introducing the additional constraint $a_k = 1$.

2 Model Fit: Assessment of

Testing the Form of the Item Characteristic Function

Ideally, a test of the fit of the item response function $P_k(\theta_i)$ would be based on assessing whether the proportion of correct responses to item k of respondents with a proficiency level θ^* matches $P_k(\theta^*)$. In the 2PLM and 3PLM this has the problem that the estimates of the proficiency parameters are virtually unique, so for every available θ -value there is only one observed response on item k available. As a result, we cannot meaningfully compute proportions of correct responses given a value of θ . However, the number-correct scores and the estimates of θ are usually highly correlated, and, therefore, tests of model fit can be based on the assumption that groups of respondents with the same number-correct score will probably be quite homogeneous with respect to their proficiency level. So a broad class of test statistics has the form

$$Q_1 = \sum_{r=1}^{K-1} N_r \frac{(O_{rk} - E_{rk})^2}{E_{rk}(1 - E_{rk})}, \quad (2)$$

where K is the test length, N_r is the number of respondents obtaining a number-correct score r , O_{rk} is the proportion of respondents with a score r and a correct score on item k , and E_{rk} is the analogous probability under the IRT model. Examples of statistics of this general form are a statistic proposed by Van den Wollenberg [25] for the 1PLM in the framework of (conditional maximum likelihood estimation) CML estimation and a test statistic for the 2PLM and 3PLM proposed by Orlando and Thissen [13] in the framework of (marginal maximum likelihood) MML estimation. For details on the computation of these statistics see [13] and [25] (*see Maximum Likelihood Estimation*).

Simulation studies show that the Q_1 -statistic is well approximated by a χ^2 -distribution [7, 22].

For long tests, it would be practical when a number of adjacent scores could be combined, say to obtain 4 to 6 score-level groups. So the sum over number-correct scores r would be replaced by a sum over score-levels g ($g = 1, \dots, G$) and the test would be based on the differences $O_{gk} - E_{gk}$. Unfortunately, it turns out that the variances of these differences cannot be properly estimated using an expression analogous to the denominator of Formula (2). The problem of weighting the differences $O_{gk} - E_{gk}$ with a proper

estimate of their variance (and covariance) will be returned to below.

Differential Item Functioning

Differential item functioning (DIF) is a difference in item responses between equally proficient members of two or more groups. One might think of a test of foreign language comprehension, where items referring to football might impede girls. The poor performance of the girls on the football-related items must not be attributed to their low proficiency level but to their lack of knowledge of football. Since DIF is highly undesirable in fair testing, methods for the detection of DIF are extensively studied. Several techniques for detection of DIF have been proposed. Most of them are based on evaluation of differences in response probabilities between groups conditional on some measure of proficiency. The most generally used technique is based on the **Mantel-Haenszel (MH) statistic** [9]. In this approach, the respondent's number-correct score is used as a proxy for proficiency, and DIF is evaluated by testing whether the response probabilities differ between the score groups. In an IRT model, proficiency is represented by a latent variable θ , and DIF can be assessed in the framework of IRT by evaluating whether the same item parameters apply in subgroups (say subgroups $g = 1, \dots, G$) that are homogeneous with respect to θ . This is achieved by generalizing the test statistic in Formula (2) to

$$Q_1 = \sum_{g=1}^G \sum_{r=1}^{K-1} N_{gr} \frac{(O_{grk} - E_{grk})^2}{E_{grk}(1 - E_{grk})}, \quad (3)$$

where N_{gr} is the number of respondents obtaining a number-correct score r in subgroup g , and O_{grk} and E_{grk} are the observed proportion and estimated probability for that combination of h and r .

Combination of number-correct scores has similar complications as discussed above, so this problem will also be addressed in the last section.

Testing Local Independence and Multidimensionality

The statistics of the previous section can be used for testing whether the data support the form of the item response functions. Another assumption underlying the IRT models presented above is unidimensionality.

Suppose unidimensionality is violated. If the respondent's position on one latent trait is fixed, the assumption of local stochastic independence requires that the association between the items vanishes. In the case of more than one dimension, however, the respondent's position in the latent space is not sufficiently described by one one-dimensional proficiency parameter and, as a consequence, the association between the responses to the items given this one proficiency parameter will not vanish. Therefore, tests for unidimensionality are based on the association between the items.

Van den Wollenberg [25] and Yen [26, 27] show that violation of local independence can be tested using a test statistic based on the evaluation of the association between items in a 2-by-2 table. Applying this idea to the 3PLM in an MML framework, a statistic can be based on the difference between observed and expected frequencies given by

$$\begin{aligned} d_{kl} &= n_{kl} - E(N_{kl}) \\ &= n_{kl} - \sum_{r=2}^{K-1} n_r P(Y_k = 1, Y_l = 1 | R = r), \end{aligned} \quad (4)$$

where n_{kl} is the observed number of respondents making item k and item l correct, in the group of respondents obtaining a score between 2 and $K-2$, and $E(N_{kl})$ is its expectation. Only scores between 2 and $K-2$ are considered, because respondents with a score less than 2 cannot make both items correct, and respondents with a score greater than $K-2$ cannot make both items incorrect. So these respondents contribute no information to the 2-by-2 table. Using Pearson's X^2 -statistic for association in a 2-by-2 table results in

$$S_{3kl} = \frac{d_{kl}^2}{E(N_{kl})} + \frac{d_{kl}^2}{E(N_{k\bar{l}})} + \frac{d_{kl}^2}{E(N_{\bar{k}l})} + \frac{d_{kl}^2}{E(N_{\bar{k}\bar{l}})}, \quad (5)$$

where $E(N_{kl})$ is the expectation of making item k correct and l wrong, and $E(N_{k\bar{l}})$ and $E(N_{\bar{k}l})$ are defined analogously. Simulation studies by Glas and Suárez-Falcón (2003) show that this statistic is well approximated by a χ^2 -distribution with one degree of freedom.

Person Fit

In the previous sections, item fit was investigated across respondents. Analogously, fit of respondents

can be investigated across items. Usually, investigation of item fit precedes the investigation of person fit, but evaluation of item and person fit can also be an iterative process. Person fit is used to check whether a person's pattern of item responses is unlikely given the model. Unlikely response patterns may occur because respondents are unmotivated or unable to give proper responses that relate to the relevant proficiency variable, or because they have preknowledge of the correct answers, or because they are cheating.

As an example we will consider a person fit statistic proposed by Smith [19]. The set of test items is divided into G nonoverlapping subtests denoted A_g ($g = 1, \dots, G$) and the test is based on the discrepancies between the observed scores and the expected scores under the model summed within subsets of items. That is, the statistic is defined as

$$UB = \frac{1}{G-1} \sum_{g=1}^G \frac{\left[\sum_{k \in A_g} (y_k - P_k(\theta)) \right]^2}{\sum_{k \in A_g} P_k(\theta)(1 - P_k(\theta))} \quad (6)$$

Since the statistic is computed only for individual students, the index i was dropped. One of the problems of this statistic is that its distribution under the null-hypothesis cannot be derived because the effects of the estimation of the parameters are not taken into account. Snijders [20] proposed a correction-factor that solves this problem.

Polytomous Items and Multidimensional Models

In the previous section, assessment of fit to IRT models was introduced by considering one-dimensional IRT models for dichotomously scored items. In this section, we will consider a more general framework, where items are scored polytomously (with dichotomous scoring as a special case) and where the latent proficiency can be multidimensional. In polytomous scoring the response to an item k can be in one of the categories $j = 0, \dots, M_k$. In one-dimensional IRT models for polytomous items, it is assumed that the probability of scoring in an item category depends on a one-dimensional latent variable θ . An example of the category response functions for an item with

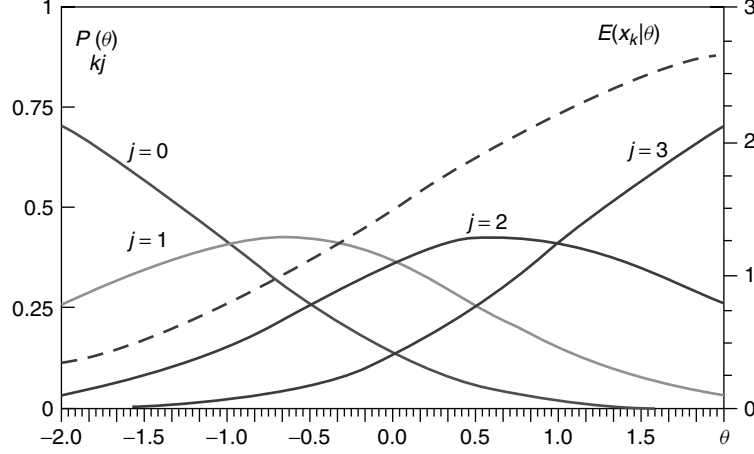


Figure 1 Response functions of a polytomously scored item

four ordered response categories is given in Figure 1. Note that the forms of the response functions are plausible for proficiency items: the response function of the zero-category decreases as a function of proficiency, the response function for the highest category increases with proficiency, and the response functions for the intermediate categories are single-peaked.

Item response models giving rise to response functions as in Figure 1 fall into three classes [11]: adjacent-category models [10], continuation-ratio models [24] and cumulative probability models [18]. The rationales underlying the models are very different but the shapes of their response functions are very similar. As an example, we give a model in the first class, the generalized partial credit model (GPCM) by Muraki [12]. The probability of a student i scoring in category j on item k given by

$$p(Y_{ikj} = 1|\theta_i) = \frac{\exp(ja_k\theta_i - b_{kj})}{1 + \sum_{h=1}^{M_k} \exp[ha_k\theta_i - b_{kh}]}, \quad (7)$$

where Y_{ikj} is a random variable assuming a value one if a response was given in category j ($j = 1, \dots, M_k$), and assuming a value zero otherwise.

Usually, in IRT models it is assumed that there is one (dominant) latent variable that explains test performance. However, it may be clear *a priori* that multiple latent variables are involved or the dimensionality of the latent variable structure might not be clear at all. In these cases, **multidimensional**

IRT models can serve confirmatory and exploratory purposes. A multidimensional version of the GPCM is obtained by assuming a Q -dimensional latent variable $(\theta_{i1}, \dots, \theta_{iq}, \dots, \theta_{iQ})$ and by replacing $a_k\theta_i$ in (7) by

$$\sum_{q=1}^Q a_{kq}\theta_{iq}.$$

Usually, it is assumed that $(\theta_{i1}, \dots, \theta_{iq}, \dots, \theta_{iQ})$ has a multivariate normal distribution [16]. Since the model can be viewed as a factor-analysis model (see **Factor Analysis: Exploratory**) [23], the latent variables $(\theta_{i1}, \dots, \theta_{iq}, \dots, \theta_{iQ})$ are often called *factor-scores* while the parameters $(a_{k1}, \dots, a_{iq}, \dots, \theta_{iQ})$ are referred to as *factor-loadings*.

A General Framework for Assessing Model Fit

In this section, we will describe a general approach to assessing evaluation of fit to IRT models based on the Lagrange multiplier test [15, 1]. Let η_1 be a vector of the parameters of some IRT model, and let η_2 be a vector of parameters added to this IRT model to obtain a more general model. Let $\mathbf{h}(\eta_1)$ and $\mathbf{h}(\eta_2)$ be the first-order derivatives of the log-likelihood function. The parameters η_1 of the IRT model are estimated by maximum likelihood, so $\mathbf{h}(\eta_1) = \mathbf{0}$. The hypothesis $\eta_2 = \mathbf{0}$ can be tested using the statistic $LM = \mathbf{h}(\eta_2)^t \Sigma^{-1} \mathbf{h}(\eta_2)$, where Σ is the covariance matrix of $\mathbf{h}(\eta_2)$. Details on the computation of $\mathbf{h}(\eta_2)$ and Σ for IRT models in an MML framework can

be found in Glas [4, 5]. It can be proved that the LM-statistic has an asymptotic χ^2 -distribution with degrees of freedom equal to the number of parameters in η_2 . In the next sections, it will be shown how this can be applied to construct tests of fit to IRT models based on residuals.

Fit of the Response Functions

As above, the score range is partitioned into G subsets to form subgroups that are homogeneous with respect to θ . A problem with polytomously scored items is that there are often too few observations on low item categories in high number-correct score groups and in high item categories in low number-correct score groups. Therefore, we will consider the expectation of the item score, defined by

$$X_{ik} = \sum_{j=1}^{M_k} jY_{ikj}, \quad (8)$$

rather than the category scores themselves. In Figure 1, the expected item score function given θ for the GPCM is drawn as a dashed line.

To define the statistic, an indicator function $w(\mathbf{y}_i^{(k)}, g)$ is introduced that is equal to one if the number-correct score on the response pattern without item k , say $\mathbf{y}_i^{(k)}$, falls in subrange g , and equal to zero if this is not the case. We will now detail the approach further for the GPCM. The alternative model on which the LM test is based, is given by

$$\begin{aligned} P(Y_{ikj} = 1 | \theta_i, \delta_{kg}, w(\mathbf{y}_i^{(k)}, g) = 1) \\ = \frac{\exp(ja_k\theta_i + j\delta_g - b_{kj})}{1 + \sum_{h=1}^{M_k} \exp(ha_k\theta_i + h\delta_g - b_{kh})}, \end{aligned} \quad (9)$$

for $j = 1, \dots, M_k$. Under the null model, which is the GPCM model, the additional parameter δ_g is equal to zero. In the alternative model, δ_g acts as a shift in ability for subgroup g . If we define $\eta_2 = (\delta_1, \dots, \delta_g, \dots, \delta_G)$, the hypothesis $\eta_2 = \mathbf{0}$ can be tested using the LM-statistic defined above. It can be shown that $\mathbf{h}(\eta_2)$ is a vector of the differences between the observed and expected item scores in the subgroups $g = 1, \dots, G$. In an MML framework, the statistic is based on the differences

$$\sum_{i|g} x_{ik} - \sum_{i|g} E[X_{ik} | \mathbf{y}_i], \quad (10)$$

for $g = 1, \dots, G$, where the summations are over all respondents in subgroup g , and the expectation is the posterior expectation given response pattern $\mathbf{y}_i$ (for details, see [5]).

Evaluation of Local Independence

Also local independence can be evaluated using the framework of LM-tests. For the GPCM, dependency between the items k and item l can be modeled as

$$\begin{aligned} P(Y_{ikj} = 1, Y_{ilp} = 1 | \theta_i, \delta_{kl}) \\ = \frac{\exp(m\theta_i - b_{kj} + p\theta_i - b_{lp} + \delta_{kjl})}{1 + \sum_g \sum_h \exp(g\theta_i - b_{kg} + h\theta_i - b_{lh} + \delta_{kglh})}. \end{aligned} \quad (11)$$

Note the parameter δ_{kjl} models the association between the two items. The LM test can be used to test the special model, where $\delta_{kjl} = 0$, against the alternative model, where $\delta_{kjl} \neq 0$. In an MML framework, the statistic is based on the differences

$$n_{kjl} - E(N_{kjl} | \mathbf{y}_i), \quad (12)$$

where n_{kjl} is the observed number of respondents scoring in category j of item k and in category p of item l , and $E(N_{kl})$ is its posterior expectation (see, [5]). The statistic has an asymptotic χ^2 -distribution with $(M_k - 1)(M_l - 1)$ degrees of freedom.

Person-fit

The fact that person-fit statistics are computed using estimates of the proficiency parameters θ can lead to serious problems with respect to their distribution under the null-hypothesis [20]. However, person-fit statistics can also be redefined as LM-statistics. For instance, the UB-test can be viewed as a test whether the same proficiency parameter θ can account for the responses in all partial response patterns. For the GPCM, we model the alternative that this is not the case by

$$\begin{aligned} P(Y_{kj} = 1 | \theta, k \in A_g) \\ = \frac{\exp[m(\theta + \theta_g) - b_{kj}]}{1 + \sum_{h=1}^{M_k} \exp[h(\theta + \theta_g) - b_{kh}]}. \end{aligned} \quad (13)$$

One subtest, say the first, should be used as a reference. Further, the test length is too short to consider too many subtests, so usually, we only consider two subtests. Defining $\eta_2 = \theta_2$ leads to a test statistic for assessing whether the total score on the second part of the test is as expected from the first part of the test.

An Example

Part of a school effectiveness study serves as a very small illustration. In a study of the effectiveness of Belgium secondary schools, several cognitive and noncognitive tests were administered to 2207 pupils at the end of the first, second, fourth, and sixth school year. The ultimate goal of the analyses was to estimate a correlation matrix between all scales over all time points using concurrent MML estimates of a multidimensional IRT model. Here we focus on the first step of these analyses, which was checking whether a one-dimensional GPCM held for each scale at each time point. The example pertains to the scale for ‘Academic Self Esteem’ which consisted of 9 items with five response categories each. The item parameters were estimated by MML assuming a standard normal distribution for θ . To compute the LM-statistics, the score range was divided into four sections in such a way that the numbers of respondents scoring in the sections were approximately equal. Section 1 contained the (partial) total-scores $r \leq 7$, Section 2 contained the scores $8 \leq r \leq 10$, Section 3 the scores $11 \leq r \leq 13$, and Section 4 $r \geq 14$. The results are given in Table 1. The column labeled ‘LM’ gives the values of the LM-statistics; the column labeled ‘Prob’ gives the significance probabilities. The statistics have three degrees of freedom.

Note that 6 of the 9 LM-tests were significant at a 5% significance level. To assess the seriousness of the misfit, the observed and the expected average item scores in the subgroups are shown under the headings ‘Obs’ and ‘Exp’, respectively. Note that the observed average scores increased with the score level of the group. Further, it can be seen that the observed and expected values were quite close: the largest absolute difference was .09 and the mean absolute difference was approximately .02. So from the values of the LM-statistics, it must be concluded that the observed item scores are definitely outside the confidence intervals of their expectations, but the precision of the predictions from the model is good enough to accept the model from the intended application.

Conclusion

The principles sketched above pertain to a broad class of parameterized IRT models, both in a logistic or normal-ogive formulation, and their multidimensional generalizations. The focus was on a likelihood-based framework. Recently, however, a Bayesian estimation framework for IRT models has emerged [14, 3] (see **Bayesian Item Response Theory Estimation**). Evaluation of model fit can be based on the same rationales and statistics as outlined above, only here test statistics are implemented as so-called posterior predictive check. Examples of this approach are given by Hoijsink [8] and Glas and Meijer [6], who show how item- and person-fit statistics can be used as posterior predictive checks.

Another important realm of IRT are the so-called nonparametric IRT models [17, 21]. Also here

Table 1 Results of the LM test to evaluate fit of the item response functions for the scale ‘academic self esteem’

Item	LM	Prob	Group 1		Group 2		Group 3		Group 4	
			Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
1	7.95	0.05	1.37	1.37	1.68	1.71	1.90	1.87	2.10	2.07
2	11.87	0.01	0.45	0.50	0.85	0.87	1.23	1.16	1.61	1.64
3	9.51	0.02	0.70	0.77	1.21	1.19	1.53	1.49	1.91	1.92
4	0.64	0.89	0.98	0.97	1.36	1.36	1.60	1.62	1.99	1.98
5	10.62	0.01	0.32	0.33	0.73	0.71	0.99	0.98	1.31	1.34
6	16.90	0.00	0.34	0.33	0.67	0.67	0.98	0.94	1.26	1.31
7	15.17	0.00	0.77	0.77	1.19	1.19	1.48	1.45	1.78	1.85
8	2.37	0.50	0.73	0.72	1.12	1.10	1.35	1.36	1.76	1.77
9	2.41	0.49	1.56	1.54	1.81	1.85	2.04	2.01	2.21	2.23

fit to IRT models can be evaluated by comparing observed proportions and expected probabilities of item responses conditional on number-correct scores, response probabilities in subpopulations, and responses to pairs of items. So the principles are the same as in parametric IRT models, but the implementation differs between applications.

References

- [1] Aitchison, J. & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints, *Annals of Mathematical Statistics* **29**, 813–828.
- [2] Birnbaum, A. (1968). Some latent trait models, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading.
- [3] Bradlow, E.T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets, *Psychometrika* **64**, 153–168.
- [4] Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests, *Statistica Sinica* **8**, 647–667.
- [5] Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model, *Psychometrika* **64**, 273–294.
- [6] Glas, C.A.W. & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models, *Applied Psychological Measurement* **27**, 217–233.
- [7] Glas, C.A.W. & Suárez-Falcón, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model, *Applied Psychological Measurement* **27**, 87–106.
- [8] Hoijsink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model, in *Essays on Item Response Theory*, A. Boomsma, M.A.J. van Duijn & T.A.B. Snijders, eds, Springer, New York, pp. 109–130.
- [9] Holland, P.W. & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure, in *Test Validity*, H. Wainer & H.I. Braun, eds, Lawrence Erlbaum, Hillsdale.
- [10] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.
- [11] Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses, *Applied Psychological Measurement* **19**, 91–100.
- [12] Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm, *Applied Psychological Measurement* **16**, 159–176.
- [13] Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models, *Applied Psychological Measurement* **24**, 50–64.
- [14] Patz, R.J. & Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics* **24**, 146–178.
- [15] Rao, C.R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.
- [16] Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data, in *Handbook of Modern Item Response Theory*, W.J. van der Linden & R.K. Hambleton, eds, Springer, New York, pp. 271–286.
- [17] Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory, *Psychometrika* **49**, 425–436.
- [18] Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores, *Psychometrika Monograph Supplement*, (17) 100–114.
- [19] Smith, R.M. (1986). Person fit in the Rasch model, *Educational and Psychological Measurement* **46**, 359–372.
- [20] Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter, *Psychometrika* **66**, 331–342.
- [21] Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality, *Psychometrika* **52**, 589–617.
- [22] Suarez-Falcon, J.C. & Glas, C.A.W. (2003). Evaluation of global testing procedures for item fit to the Rasch model, *British Journal of Mathematical and Statistical Psychology* **56**, 127–143.
- [23] Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables, *Psychometrika* **52**, 393–408.
- [24] Tutz, G. (1990). Sequential item response models with an ordered response, *British Journal of Mathematical and Statistical Psychology* **43**, 39–55.
- [25] Van den Wollenberg, A.L. (1982). Two new tests for the Rasch model, *Psychometrika* **47**, 123–140.
- [26] Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, *Applied Psychological Measurement* **8**, 125–145.
- [27] Yen, W. (1993). Scaling performance assessments: strategies for managing local item dependence, *Journal of Educational Measurement* **30**, 187–213.

CEES A.W. GLAS

Model Identifiability

GUAN-HUA HUANG

Volume 3, pp. 1249–1251

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Model Identifiability

In some statistical models, different parameter values can give rise to identical probability distributions. When this happens, there will be a number of different parameter values associated with the **maximum likelihood** of any set of observed data. This is referred to as the model identifiability problem. For example, suppose someone attempts to compute the regression equation predicting Y from three variables X_1 , X_2 , and their sum $(X_1 + X_2)$, the program will probably crash or give an error message because it cannot find a unique solution. The model is the same if $Y = 0.5X_1 + 1.0X_2 + 1.5(X_1 + X_2)$, $Y = 1.0X_1 + 1.5X_2 + 1.0(X_1 + X_2)$, or $Y = 2.0X_1 + 2.5X_2 + 0.0(X_1 + X_2)$; indeed there are an infinite number of equally good possible solutions. Model identifiability is a particular problem for the latent class model, a statistical method for finding the underlying traits from a set of psychological tests, because, by postulating latent variables, it is easy to introduce more parameters into a model than can be fitted from the data.

A model is identifiable if the parameter values uniquely determine the probability distribution of the data and the probability distribution of the data uniquely determines the parameter values. Formally, let ϕ be the parameter value of the model, y be the observed data, and $F(y; \phi)$ be the probability distribution of the data. A model is identifiable if for all $(\phi_0, \phi) \in \Phi$ and for all $y \in S_Y$:

$$F(y; \phi_0) = F(y; \phi) \text{ if and only if } \phi_0 = \phi, \quad (1)$$

where Φ denotes the set of all possible parameter values, and S_Y is the set of all possible values of the data.

The most common cause of model nonidentifiability is a poorly specified model. If the number of unique model parameters exceeds the number of independent pieces of observed information, the model is not identifiable. Consider the example of a latent class model that classifies people into three states (severely depressed/mildly depressed/not depressed) and that is used to account for the responses of a group of people to three psychological tests with binary (positive/negative) outcomes. Let (Y_1, Y_2, Y_3) denote the test results and let each take the value

1 when the outcome is positive and 0 when it is negative. S specifies the unobservable states where $S = 1$ where there is no depression, 2 where the depression is mild, and 3 where the depression is severe. The probability of the test results is then

$$\begin{aligned} & \Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \\ &= \sum_{j=1}^3 \Pr(S = j) \prod_{m=1}^3 \Pr(Y_m = 1|S = j)^{y_m} \\ & \quad \times \Pr(Y_m = 0|S = j)^{1-y_m}. \end{aligned} \quad (2)$$

The test results have $2^3 - 1 = 7$ independent patterns and the model requires 11 unique parameters (Two probabilities for depression status $\Pr(S = 3)$, $\Pr(S = 2)$, and one conditional probability $\Pr(Y_m = 1|S = j)$ for each depression status j and test m); therefore, the model is not identifiable.

If the model is not identifiable, one can make it so by imposing various constraints upon the parameters. When there appears to be sufficient total observed information for the number of estimated parameters, it is also necessary to specify the model unambiguously. For the above latent class model, suppose that, for the second and third tests, the probabilities of observing a positive test result are the same for people with severe, mild, or no depression (i.e., $\Pr(Y_m = 1|S = 3) = \Pr(Y_m = 1|S = 2) = \Pr(Y_m = 1|S = 1) = p_m$ for $m = 2, 3$). In other words, only the first test discriminates between the unobservable states of depression. The model now has only seven parameters, which is equal to the number of independent test result patterns. The probability distribution of test results becomes

$$\begin{aligned} & \Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) \\ &= \Theta \prod_{m=2}^3 (p_m)^{y_m} (1 - p_m)^{1-y_m}, \end{aligned} \quad (3)$$

where

$$\begin{aligned} \Theta &= (1 - \eta_2 - \eta_3)(p_{11})^{y_1}(1 - p_{11})^{1-y_1} \\ & \quad + \eta_2(p_{12})^{y_1}(1 - p_{12})^{1-y_1} \\ & \quad + \eta_3(p_{13})^{y_1}(1 - p_{13})^{1-y_1}, \end{aligned} \quad (4)$$

$\eta_2 = \Pr(S = 2)$, $\eta_3 = \Pr(S = 3)$, $p_{11} = \Pr(Y_1 = 1|S = 1)$, $p_{12} = \Pr(Y_1 = 1|S = 2)$, and $p_{13} = \Pr(Y_1 = 1|S = 3)$. Θ imposes two restrictions on parameters

(i.e., for $y_1 = 1$ or 0), and there are five parameters to consider (i.e., η_2 , η_3 , p_{11} , p_{12} and p_{13}). Because the number of restrictions is less than the number of parameters of interest, Θ and the above latent class model are not identifiable – the same probability distributions could be generated by supposing that there was a large chance of being in a state with a small effect on the probability of being positive on test 1 or by supposing that there was a small chance of being in this state but it was associated with a large probability of responding positively.

Sometimes it is difficult to find an identifiable model. A weaker form of identification, called *local identifiability*, may exist, namely, it may be that other parameters generate the same probability distribution as ϕ_0 does, but one can find an open neighborhood of ϕ_0 that contains none of these parameters [3]. For example, we are interested in β in the regression $Y = \beta^2 X$ (the square root of the association between Y and X). $\beta = 1$ and $\beta = -1$ result in the same Y prediction; thus, the model is not (globally) identifiable. However, the model is locally identifiable because one can easily find two nonoverlapping intervals $(0.5, 1.5)$ and $(-1.5, -0.5)$ for 1 and -1 , respectively. A locally but not globally identifiable model does not have a unique interpretation, but one can be sure that, in the neighborhood of the selected solution, there exist no other equally good solutions; thus, the problem is reduced to determining the regions where local identifiability applies. This concept is especially useful in models containing nonlinearities as the above regression example, or models with complex structures, for example, **factor analysis**, latent class models and **Markov Chain Monte Carlo**.

It is difficult to specify general conditions that are sufficient to guarantee (global) identifiability. Fortunately, it is fairly easy to determine local identifiability. One can require that the columns of the Jacobian matrix, the first-order partial derivative of the

likelihood function with respect to the unique model parameters, are independent [2, 3]. Alternatively, we can examine whether the Fisher information matrix possesses **eigenvalues** greater than zero [4]. Formann [1] showed that these two approaches are equivalent. A standard practice for checking local identifiability involves using multiple sets of initial values for parameter estimation. Different sets of initial values that yield the same likelihood maximum should result in the same final parameter estimates. If not, the model is not locally identifiable.

When applying a nonidentifiable model, different people may draw different conclusions from the same model of the observed data. Before one can meaningfully discuss the estimation of a model, model identifiability must be verified. If researchers come up against identifiability problems, they can first identify the parameters involved in the lack of identifiability from their extremely large asymptotic standard errors [1], and then impose reasonable constraints on identified parameters based on prior knowledge or empirical information.

References

- [1] Formann, A.K. (1992). Linear logistic latent class analysis for polytomous data, *Journal of the American Statistical Association* **87**, 476–486.
- [2] Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**, 215–231.
- [3] McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis, *Psychometrika* **21**, 331–347.
- [4] Rothenberg, T.J. (1971). Identification in parametric models, *Econometrica* **39**, 577–591.

GUAN-HUA HUANG

Model Selection

LILIA M. CORTINA

Volume 3, pp. 1251–1253

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Model Selection

When a behavioral scientist statistically models relationships among a set of variables, the goal is to provide a meaningful and parsimonious explanation for those relationships, ultimately achieving a close approximation to reality. However, given the complexity inherent in social science data and the phenomena they attempt to capture, there are typically multiple plausible explanations for any given set of observations. Even when one model fits well, other models with different substantive interpretations are virtually always possible. The task at hand, then, is to determine ‘which models are the “fittest” to survive’ [1, p. 71]. In this article, I will review issues to consider when selecting among alternative models, using **structural equation modeling** (SEM) as a context. However, the same general principles apply to other types of statistical models (e.g., **multiple regression**, **analysis of variance**) as well.

In light of the existence of multiple viable explanations for a set of relationships, one approach is to formulate multiple models in advance, test each with the same set of data, and determine which model shows superior qualities in terms of fit, parsimony, interpretability, and meaningfulness. Each model should have a strong theoretical rationale. Justifications for such a *model comparison* strategy are many. For example, this approach is quite reasonable when investigating a new research domain – if a phenomenon is not yet well understood, there is typically some degree of uncertainty about how it operates, so it makes sense to explore different alternatives [4]. This exploration should happen at the model-development stage (*a priori*) rather than at the model-fitting stage (*post hoc*), to avoid capitalizing on chance. Even in established lines of research, scientists may have competing theoretical propositions to test, or equivocal findings from prior research could suggest multiple modeling possibilities. Finally, researchers can argue more persuasively for a chosen model if they can demonstrate its statistical superiority over rival, theoretically compelling models. For this reason, some methodologists advocate that consideration of multiple alternatives be standard practice when modeling behavioral phenomena [e.g., [2, 5]].

After one has found strong, theoretically sensible results for a set of competing models, the question of

selection then arises: Which model to retain? **Model selection** guidelines vary depending on whether the alternative models are nested or nonnested. Generally speaking, if Model B is *nested* within Model A, then Model B contains the same variables as Model A but specifies fewer relations among them. In other words, Model B is a special case or a subset of Model A; because it is more restrictive, Model B cannot fit the data as well [6]. However, Model B provides a more parsimonious explanation of the data, so if the decrement in overall fit is trivial, then B is often considered the better model.

To compare the fit of nested structural equation models, the *chi-square difference test* (also termed the Likelihood Ratio or LR test – see **Maximum Likelihood Estimation**) is most often employed. For example, when comparing nested Models A and B, the researcher estimates both models, computing the overall χ^2 fit statistic for each. She or he then calculates the difference between the two χ^2 values ($\chi^2_{\text{ModelA}} - \chi^2_{\text{ModelB}}$). The result is distributed as a χ^2 , with degrees of freedom (df) equal to the difference in df for the two models ($\text{df}_{\text{ModelA}} - \text{df}_{\text{ModelB}}$). If this value is significant (according to a chi-square table), then the restrictions imposed in the smaller Model B led to a significant worsening of fit, so the more comprehensive Model A should be retained. (Another way to characterize this same situation is to say that the additional relationships introduced in Model A led to a significant improvement in fit – again suggesting that Model A should be retained.) However, if this test does not point to a significant difference in the fit of these two models, the simpler, nested Model B is typically retained [e.g., 1, 6].

When comparing alternative models that are *nonnested*, a chi-square difference test is not appropriate. Instead, one can rely on ‘information measures of fit’, such as **Akaike’s Information Criterion** (AIC), the *Corrected Akaike’s Information Criterion* (CAIC), or the single-sample *Expected Cross Validation Index* (ECVI). These measures do not appear often in behavioral research, but they are appropriate for the comparison of alternative nonnested models. The researcher simply estimates the models, computes one of these fit indices for each, and then rank-orders the models according to the chosen index; the model with the lowest index value shows the best fit to the data [3].

In the cases reviewed above, only overall ‘goodness of fit’ is addressed. However, during model

2 Model Selection

selection, researchers should also pay close attention to the substantive implications of obtained results: Are the structure, valence, and magnitude of estimated relationships consistent with theory? Are the parameter estimates interpretable and meaningful? Even if fit indices suggest one model to be superior to others, this model is useless if it makes no sense from a substantive perspective [4]. Moreover, there often exist multiple plausible models that are mathematically comparable and yield identical fit to a given dataset. These *equivalent models* differ only in terms of substantive meaning, so this becomes the primary factor driving model selection [5].

One final *caveat* about model selection bears mention. Even after following all of the practices reviewed here – specifying a parsimonious model based on strong theory, testing it against viable alternatives, and evaluating it to have superior fit and interpretability, a researcher still cannot definitively claim to have captured ‘Truth,’ or even to have identified THE model that BEST approximates reality. Many, many models are always plausible, and selection of an excellent model could artifactually result from failure to consider every possible alternative [4, 5]. With behavioral processes being the complex, messy phenomena that they are, we can only aspire to represent them *imperfectly* in statistical models,

rarely (if ever) knowing the ‘true’ model. Ay, there’s the rub [6].

References

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, John Wiley & Sons, New York.
- [2] Breckler, S.J. (1990). Applications of covariance structure modeling in psychology: cause for concern? *Psychological Bulletin* **107**, 260–273.
- [3] Hu, L. & Bentler, P.M. (1995). Evaluating model fit, in *Structural Equation Modeling: Concepts, Issues, and Applications*, R.H. Hoyle, ed., Sage Publications, Thousand Oaks, pp. 76–99.
- [4] MacCallum, R.C. (1995). Model specification: procedures, strategies, and related issues, in *Structural Equation Modeling: Concepts, Issues, and Applications*, R.H. Hoyle, ed., Sage Publications, Thousand Oaks, pp. 16–36.
- [5] MacCallum, R.C., Wegener, D.T., Uchino, B.N. & Fabrigar, L.R. (1993). The problem of equivalent models in applications of covariance structure analysis, *Psychological Bulletin* **114**, 185–199.
- [6] Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd Edition, Wadsworth Publishers.

(See also **Goodness of Fit**)

LILIA M. CORTINA

Models for Matched Pairs

VANCE W. BERGER AND JIALU ZHANG

Volume 3, pp. 1253–1256

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Models for Matched Pairs

When **randomization** is impractical or impossible, matched studies are often used to enhance the comparability of comparison groups, so as to facilitate the assessment of the association between an exposure and an event while minimizing confounding. Matched pairs designs, which tend to be more resistant to biases than are **historically controlled studies**, are characterized by a particular type of statistical dependence, in which each pair in the sample contains observations either from the same subject or from subjects that are related in some way (*see Matching*). The matching is generally based on subject characteristics such as age, gender, or residential zip code, but could also be on a propensity score based on many such subject characteristics [13]. The standard design for such studies is the **case-control study**, in which each case is matched to either one control or multiple controls. The simplest example would probably be binary data (*see Binomial Distribution: Estimating and Testing Parameters*) with one-to-one matched pairs, meaning that the response or outcome variable is binary and each case is matched to a single control.

There are several approaches to the analysis of matched data. For example, the matched pairs t Test can be used for a continuous response. Stratified analysis, the McNemar test [11], and conditional logistic regression [6, 7] can be used for data with discrete or binary responses. In some cases, it even pays to break the matching [9].

The Matched Pairs t Test

The matched pairs t Test is used to test for a difference between measurements taken on subjects before an intervention or an event versus measurements taken on them after an intervention or an event. In the test, each subject is allowed to serve as his or her own control. The matched paired t Test reduces confounding that could result from comparing one group of subjects receiving one treatment to a different group of subjects receiving a different treatment.

Let $X = (x_1, x_2, \dots, x_n)$ be the first observations from each of the n subjects, and let $Y = (y_1, y_2, \dots, y_n)$ be the second observations from the

same n subjects. The test statistic is a function of the differences $d = (d_1, d_2, \dots, d_n)$, where $d_i = x_i - y_i$. If the d_i can be considered to have a normal distribution (an assumption not to be taken lightly [2, 4]) with mean zero and arbitrary variance, then it is appropriate to use the t Test. The test statistic can be written as follows:

$$t = \frac{\sum_i d_i / n}{\left(\frac{S_d}{\sqrt{n}} \right)}, \quad (1)$$

where S_d is the standard deviation computed from $(d_1, d_2, \dots, d_n)$. t follows a Student's t distribution (*see Catalogue of Probability Density Functions*)(with mean zero) under the null hypothesis. If this test is rejected, then one would conclude that the before and after observations from the same pair are not equivalent to each other.

The McNemar Test

When dealing with discrete data, we denote by π_{ij} the probability of the first observation of a pair having outcome i and the second observation having outcome j . Let n_{ij} be the number of such pairs. Clearly, then, π_{ij} can be estimated by n_{ij}/n , where n is the total number of pairs in the data. If the response is binary, then McNemar's test can be applied to test the marginal homogeneity of the **contingency table** with an exposure and an event. That is, McNemar's test tests the null hypothesis that $H_0: \pi_{+1} = \pi_{1+}$, where $\pi_{+1} = \pi_{11} + \pi_{21}$ and $\pi_{1+} = \pi_{11} + \pi_{12}$. This is equivalent, of course, to testing the null hypothesis that $\pi_{21} = \pi_{12}$. This is called *symmetry*, as in Table 1. The McNemar test statistic z_0 is computed as follows:

$$z_0 = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}. \quad (2)$$

Table 1 The Structure of a Hypothetical 2×2 Contingency Table

	Before		Total
	1	2	
After			
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	π_{++}

2 Models for Matched Pairs

Now asymptotically z_0^2 follows a χ^2 distribution with degree of freedom equal to 1. One can also use a continuity correction, an exact test, or still other variations [10].

Stratified Analysis

Stratified analysis is used to control confounding by covariates that are associated with both the outcome and the exposure [14]. For example, if a study examines the association between smoking and lung cancer among several age groups, then age may have a confounding effect on the apparent association between smoking and lung cancer, because there may be an age imbalance with respect to both smoking and incidence of lung cancer. In such a case, the data would need to be stratified by age (*see Stratification*). One method for doing this is the **Mantel-Haenszel test** [10], which is performed to measure the average conditional association by estimating the common **odds ratio** as the sum of weighted odds ratios from each stratum. Of course, in reality, the odds ratios across the strata need not be common, and so the **Breslow-Day test** [5] can be used to test the homogeneity of odds ratios across the strata. If the strata do not have a common odds ratio, then the association between smoking and lung cancer should be tested separately in each stratum.

Conditional Logistic Regression

If the matching is one-to-one so that there is one control per case, and if the response is binary with outcomes 0 and 1, then within each matched set there are four possibilities for the pair (case, control): (1,0), (0,1), (1,1), (0,0). Denote by (Y_{i1}, Y_{i2}) the two observations of the i th matched set. Then the conditional **logistic regression** model can be expressed as follows:

$$\begin{aligned} \text{logit}[P(Y_{it} = 1)] &= \alpha_i + \beta x_{it} \\ i &= 1, 2, \dots, n; \quad t = 1, 2, \end{aligned} \quad (3)$$

where x_{it} is the explanatory variable of interest and α_i and β are model parameters. In particular, α_i describes the matched-set-specific effect while β is a common effect across pairs. With the above

model, one can assume independence of the observations, both for different subjects and within the same subject.

Hosmer and Lemeshow [8] provide a dataset with 189 observations representing 189 women, of whom 59 had low birth-weight babies and 130 had normal-weight babies. Several risk factors for low birth-weight babies were under investigation. For example, whether the mother smoked or not (sm) and history of hypertension (ht) were considered as possible predictors, among others. A subset containing data from 20 women is used here (Table 2) as an example to illustrate the 1:1 matching. Specifically, 20 women were matched by age (to the nearest year), and divided accordingly into 10 strata. In each age stratum, there is one case of low birth weight ($bwt = 1$) and one control ($bwt = 0$).

The conditional logistic regression model for this data set can be written as follows:

$$\begin{aligned} \text{logit}[P(Y_{it} = 1)] &= \alpha_i + \beta_{\text{sm}} x_{\text{sm},it}, \\ i &= 1, 2, \dots, 10, \quad t = 1, 2, \end{aligned} \quad (4)$$

where Y_{it} represents the outcome for woman t in age group i , α_i is the specific effect in age group i ,

Table 2 An example of a matched data set. A data extracted from Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Edition, Wiley-Interscience

Stratum	<i>bwt</i>	Age	Race	Smoke	ht
1	0	14	1	0	0
1	1	14	3	1	0
2	0	15	2	0	0
2	1	15	3	0	0
3	0	16	3	0	0
3	1	16	3	0	0
4	0	17	3	0	0
4	1	17	3	1	0
5	0	17	1	1	0
5	1	17	1	1	0
6	0	17	2	0	0
6	1	17	1	1	0
7	0	17	2	0	0
7	1	17	2	0	0
8	0	17	3	0	0
8	1	17	2	0	1
9	0	18	1	1	0
9	1	18	3	0	0
10	0	18	1	1	0
10	1	18	2	1	0

and $x_{sm,it}$ represents the smoking status of woman t in age group i , which can be either 1 or 0. In this particular case, β_{sm} is estimated to be 0.55. Therefore, the odds of having low birth-weight babies for woman in group i who smoke is $\exp(\beta_{sm}) = 1.74$ times the odds for woman in the same group who don't smoke.

Conditional logistic regression can also be used in $n:m$ matching scenario. If the complete low birth-weight dataset from Hosmer and Lemeshow [8] is used, then each age stratum, instead of having one case and one control, contains multiple cases and multiple controls. Similar conditional logistic regression model can be applied.

The conditional logistic regression model can also include extra predictors for the covariates that are not controlled by matched pairs. It is also possible to treat the α_i 's as random effects to eliminate the large number of nuisance parameters. The pair-specific effect can be modeled as a parameter following some normal distribution with unknown mean and variance. If, instead of the two levels that would prevail in the binary response case, the response has $J > 2$ levels, then a reference level is chosen for the purpose of comparisons. Without loss of generality, say that level J is considered to be the reference level, to which other levels are compared. Then the model is as follows:

$$\log \left(\frac{P(Y_{it} = k)}{P(Y_{it} = J)} \right) = \alpha_{ik} + \beta_k x_{it},$$

$$k = 1, 2, \dots, J - 1, t = 1, 2. \quad (5)$$

Clearly, this is a generalization of the model for binary data. Specialized methods also exist for analyzing matched pairs in a nonparametric fashion with missing data without assuming that the missing data are missing completely at random (see **Missing Data**) [1]. When considering 2×2 matched pairs and testing for noninferiority, it has been found that asymptotic tests may exceed the claimed nominal Type I error rate [12], and so an exact test (see **Exact Methods for Categorical Data**) would generally be preferred. The very term 'exact test' may appear to be a misnomer in the context of a matched design, because there is neither random sampling nor random allocation, and hence, technically, no basis for formal hypothesis testing [3]. However, the basis for inference in matched designs is distinct from that of randomized designs, which

involve either random sampling or random allocation (see **Randomization Based Tests**). Specifically, while in a randomized design the outcome of a given subject exposed at a given time to a given treatment is generally taken as fixed (not random), the outcome of a matched design is taken as the random quantity. So here the randomness is a within-subject factor, or, more correctly, is random even within the combination of subject, treatment, and time. That such a random component to the outcomes exists needs to be determined on a case-by-case basis. Rubin [13] pointed out that the lack of randomization creates sensitivity to the assignment mechanism, which cannot be avoided simply by using Bayesian methods instead of randomization-based methods.

References

- [1] Akritas, M.G., Kuha, J. & Osgood, D.W. (2002). A nonparametric approach to matched pairs with missing data, *Sociological Methods & Research* **30**(3), 425–454.
- [2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [3] Berger, V.W. & Bears, J. (2003). When can a clinical trial be called 'randomized'? *Vaccine* **21**, 468–472.
- [4] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**(1), 74–82.
- [5] Breslow, N.E. & Day, N.E. (1993). *Statistical Methods in Cancer Research, Vol. I: The Analysis of Case-Control Studies*, IARC Scientific Publications, No. 32, Oxford University Press, New York.
- [6] Cox, D.R. (1958). *Planning of Experiments*, John Wiley, New York.
- [7] Cox, D.R. (1970). *The Analysis of Binary Data*, Methuen, London.
- [8] Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Edition, Wiley-Interscience.
- [9] Lynn, H.S. & McCulloch, C.E. (1992). When does it pay to break the matches for analysis of a matched-pairs design? *Biometrics* **48**, 397–409.
- [10] Mantel, N. & Haenszel, A. (1959). Statistical aspects of the analysis of data from retrospective studies of diseases, *Journal of the National Cancer Institute* **22**, 719–748.
- [11] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–1157.
- [12] Rigby, A.S. & Robinson, M.B. (2000). Statistical methods in epidemiology. IV. Confounding and the matched pairs odds ratio, *Disability and Rehabilitation* **22**(6), 259–265.

4 Models for Matched Pairs

- [13] Rubin, D.B. (1991). Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism, *Biometrics* **47**(4), 1213–1234.
- [14] Sidik, K. (2003). Exact unconditional tests for testing non-inferiority in matched-pairs design, *Statistics in Medicine* **22**, 265–278.

Further Reading

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley-Interscience.

VANCE W. BERGER AND JIALU ZHANG

Moderation

J. MATTHEW BEAUBIEN

Volume 3, pp. 1256–1258

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Moderation

Every field of scientific inquiry begins with the search for universal predictor–criterion correlations. Unfortunately, universal relationships rarely exist in nature [4]. At best, researchers find that their hypothesized correlations are weaker than expected. At worst, the hypothesized correlations are wildly inconsistent from study to study. The resulting cacophony has caused more than a few researchers to abandon promising lines of research for others that are (hopefully) more tractable.

In some cases, however, these counterintuitive or conflicting findings motivate the researchers to reexamine their underlying theoretical models. For example, the researchers may attempt to specify the conditions under which the hypothesized predictor–criterion relationship will hold true. This theoretical-based search for moderator variables – or interaction effects – is one indication of the sophistication or maturity of a field of study [3].

The Moderator-mediator Distinction

Many researchers confuse the concepts of moderation and **mediation** [3, 5]. However, the distinction is relatively simple. Moderation concerns the effect of a third variable (z) on the strength or direction of a predictor–criterion ($x - y$) correlation. Therefore, moderation addresses the issues of ‘when?’, ‘for whom?’, or ‘under what conditions?’ does the hypothesized predictor–criterion correlation hold true. In essence, moderator variables are nothing more than **interaction effects**.

A moderated relationship is represented by a single, nonadditive, linear function where the criterion variable (y) varies as a product of the independent (x) and moderator variables (z) [5]. Algebraically, this function is expressed as $y = f(x, z)$. When analyzed using multiple regression, the function is usually expressed as $y = b_1x + b_2z + b_3xz + e$. Specifically, the predicted value of y is modeled as a function of the independent variable (x), the moderator variable (z), their interaction (xz), and measurement error (e). Graphically, mediated relationships are represented by an arrow from the moderator (z) that intersects the $x - y$ relationship at a 90° angle (see Figure 1). Although

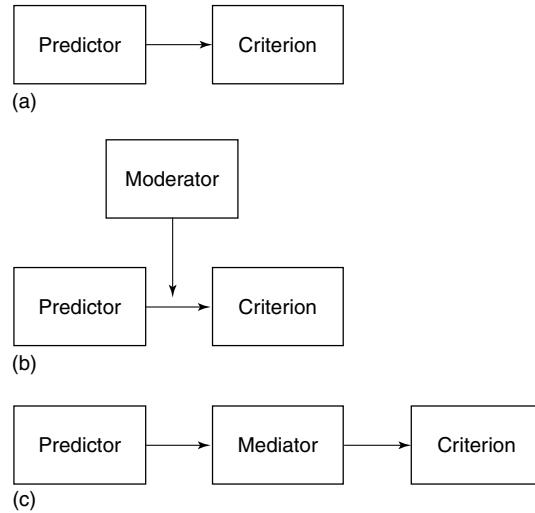


Figure 1 Examples of (a) direct, (b) moderated, and (c) mediated relationships

moderated relationships can help resolve apparently contradictory research findings, they do not imply causality.

By contrast, mediation represents one or more links in the causal chain (z) between the predictor (x) and criterion (y) variables. Therefore, mediation addresses the issues of ‘how?’ or ‘why?’ the predictor variable influences the criterion [3]. In essence, mediator variables are caused by the predictor and, in turn, predict the criterion.

Unlike moderated relationships, mediated relationships are represented by two or more additive, linear functions. Algebraically, these functions are expressed as $y = f(x)$, $z = f(x)$, and $y = f(z)$. When analyzed using **multiple regression**, these functions are usually expressed as $z = bx + e$ and $y = bz + e$. Specifically, the predicted value of the mediator (z) varies as a function of the independent variable (x) and measurement error (e). In addition, the predicted value of the criterion variable (y) varies as a function of the mediator (z) and measurement error (e). Graphically, moderated relationships are represented as a series of arrows in which the predictor variable influences the mediator variable, which in turn influences the criterion variable (see Figure 1). Unlike moderated relationships, mediated relationships specify the chain of causality [3, 5].

Testing Moderated Relationships

Moderated relationships can be tested in a variety of ways. When both the predictor and moderator variables are measured as categorical variables, the moderated relationship can be tested using **analysis of variance** (ANOVA). However, when one or both are measured on a continuous scale, hierarchical regression is preferred (*see Hierarchical Models*). Many researchers favor regression because it is more flexible than ANOVA. It also eliminates the need to artificially dichotomize continuous variables. Regardless of which analytical technique is used, the tests are conducted in very similar ways.

First, the researcher needs to carefully choose the moderator variable. The moderator variable is typically selected on the basis of previous research, theory, or both. Second, the researcher needs to specify the nature of the moderated relationship. Most common are enhancing interactions, which occur when the moderator enhances the effect of the predictor variable, or buffering interactions, which occur when moderator weakens the effect of the predictor variable. Less common are antagonistic interactions, which occur when the predictor and moderator variables have the same effect on the criterion, but their interaction produces an opposite effect [3].

Third, the researcher needs to ensure that the study has sufficient statistical **power** to detect an interaction effect. Previous research suggests that the power to detect interactions is substantially lower than the 0.80 threshold [1]. Researchers should consider several factors when attempting to maximize their study's statistical power. For example, researchers should consider not only the total effect size (R^2) but also the incremental effect size (ΔR^2) when selecting the necessary minimum sample sizes. Other important tasks include selecting reliable and valid measures, taking steps such as oversampling to avoid range restriction, centering predictor variables to reduce collinearity, and ensuring that subgroups have equivalent sample sizes and error variances (i.e., when one or more of the variables is measured on a categorical scale) [3].

Fourth, the researcher needs to create the appropriate product terms. These product terms, which are created by multiplying the predictor and moderator variables, represent the interaction between them.

Fifth, the researcher needs to structure the equation. For example, using hierarchical regression, the researcher would enter the predictor and moderator variables during the first step. After controlling for these variables, the researcher would enter the interaction terms during the second step. The significance of the interaction term is determined by examining the direction of the interaction term's regression weight, the magnitude of the effect (ΔR^2), and its statistical significance [3].

Finally, the researcher should plot the effects to determine the type of effect. For each grouping variable, the researcher should plot the scores at the mean, one standard deviation above the mean, and one standard deviation below the mean. These plots should help the researcher to visualize the form of the moderator effect: an enhancing interaction, a buffering interaction, or an antagonistic interaction. Alternatively, the researcher could test the statistical significance of the simple regression slopes for different values of the moderator variable [2].

Miscellaneous

In this entry, we explored the basics of moderation, the search for interaction effects. The discussion was limited to single-sample studies using analytical techniques such as ANOVA or hierarchical regression. However, moderator analyses can also be assessed in other ways. For example, moderators can be assessed in **structural equation modeling** (SEM) or **meta-analysis** by running the model separately for various subgroups and comparing the two sets of results.

Regardless of how moderators are tested, previous research suggests that tests for moderation tend to be woefully underpowered. Therefore, it should come as no surprise that many researchers have failed to find significant interaction effects, even though they are believed to be the norm, rather than the exception [4].

References

- [1] Aguinis, H., Boik, R. & Pierce, C. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression, *Organizational Research Methods* 4, 291–323.

-
- [2] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [3] Frazier, P.A., Tix, A.P. & Barron, K.E. (2004). Testing moderator and mediator effects in counseling psychology research, *Journal of Counseling Psychology* **51**, 115–134.
- [4] Jaccard, J., Turrissi, R. & Wan, C. (1990). *Interaction Effects in Multiple Regression*, Sage Publications, Thousand Oaks.
- [5] James, L.R. & Brett, J.M. (1984). Mediators, moderators, and tests for mediation, *Journal of Applied Psychology* **69**, 307–321.

J. MATTHEW BEAUBIEN

Moments

REBECCA WALWYN

Volume 3, pp. 1258–1260

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Moments

Moments are an important class of **expectation** used to describe probability distributions. Together, the entire set of moments of a random variable will generally determine its probability distribution exactly.

There are three main types of moments:

1. raw moments,
2. central moments, and
3. factorial moments.

Raw Moments

Where a random variable is denoted by the letter X , and k is any positive integer, the k th raw moment of X is defined as $E(X^k)$, the expectation of the random variable X raised to the power k . Raw moments are usually denoted by μ'_k where $\mu'_k = E(X^k)$, if that expectation exists. The first raw moment of X is $\mu'_1 = E(X)$, also referred to as the mean of X . The second raw moment of X is $\mu'_2 = E(X^2)$, the third $\mu'_3 = E(X^3)$, and so on. If the k th moment of X exists, then all moments of lower order also exist. Therefore, if the $E(X^2)$ exists, it follows that $E(X)$ exists.

Central Moments

Where X again denotes a random variable and k is any positive integer, the k th central moment of X is defined as $E[(X - c)^k]$, the expectation of X minus a constant, all raised to the power k . Where the constant is the mean of the random variable, this is referred to as the k th central moment around the mean. Central moments around the mean are usually denoted by μ_k where $\mu_k = E[(X - \mu_X)^k]$. The first central moment is equal to zero as $\mu_1 = E[(X - \mu_X)] = E(X) - E(\mu_X) = \mu_X - \mu_X = 0$. In fact, if the probability distribution is symmetrical around the mean (e.g., the normal distribution) all odd central moments of X around the mean are equal to zero, provided they exist. The most important central moment is the second central moment of X around the mean. This is $\mu_2 = E[(X - \mu_X)^2]$, the variance of X .

The third central moment about the mean, $\mu_3 = E[(X - \mu_X)^3]$, is sometimes used as a measure of

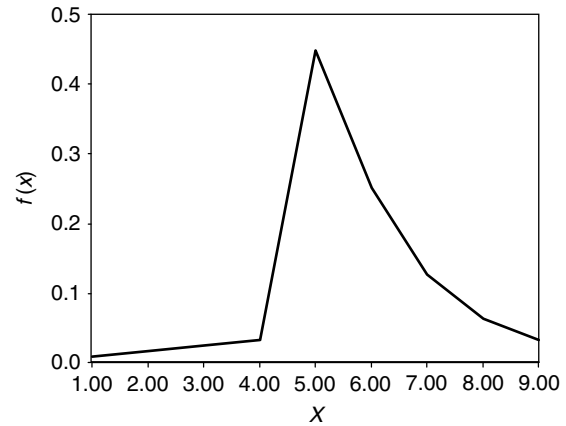


Figure 1 An example of an asymmetrical probability distribution where the third central moment around the mean is equal to zero

asymmetry or **skewness**. As an odd central moment around the mean, μ_3 is equal to zero if the probability distribution is symmetrical. If the distribution is negatively skewed, the third central moment about the mean is negative, and if it is positively skewed, the third central moment around the mean is positive. Thus, knowledge of the shape of the distribution provides information about the value of μ_3 . Knowledge of μ_3 does not necessarily provide information about the shape of the distribution, however. A value of zero may not indicate that the distribution is symmetrical. As an illustration of this, μ_3 is approximately equal to zero for the distribution depicted in Figure 1, but it is not symmetrical. The third central moment is therefore not used much in practice.

The fourth central moment about the mean, $\mu_4 = E[(X - \mu_X)^4]$, is sometimes used as a measure of excess or **kurtosis**. This is the degree of flatness of the distribution near its center. The coefficient of kurtosis $(\mu_4/\sigma^4 - 3)$ is sometimes used to compare an observed distribution to that of a normal curve. Positive values are thought to be indicative of a distribution that is more peaked around its center than that of a normal curve, and negative values are thought to be indicative of a distribution that is more flat around its center than that of a normal curve. However, as was the case for the third central moment around the mean, the coefficient of kurtosis does not always indicate what it is supposed to.

Factorial Moments

Finally, where X denotes a random variable and k is any positive integer, the k th factorial moment of X is defined as the following expectation $E[X(X-1)\dots(X-k+1)]$. The first factorial moment of X is therefore $E(X)$, the second factorial moment of X is $E[(X-1+1)(X-2+1)] = E(X^2 - X)$, and so on. Factorial moments are easier to calculate than raw moments for some random variables (usually discrete). As raw moments can be obtained from factorial moments and vice versa, it is sometimes easier to obtain the raw moments for a random variable from its factorial moments.

Moment Generating Function

For each type of moment, there is a function that can be used to generate all of the moments of a random variable or probability distribution. This is referred to as the 'moment generating function' and denoted by mgf, $m_X(t)$ or $m(t)$. In practice, however, it is often easier to calculate moments directly. The main use of the moment generating function is therefore in

characterizing a distribution and for theoretical purposes. For instance, if a moment generating function of a random variable exists, then this moment generating function uniquely determines the corresponding distribution function. As such, it can be shown that if the moment generating functions of two random variables both exist and are equal for all values of t in an interval around zero, then the two cumulative distribution functions are equal. However, existence of all moments is not equivalent to existence of the moment generating function.

More information on the topic of moments and moment generating functions is given in [1, 2, 3].

References

- [1] Casella, G. & Berger, R.L. (1990). *Statistical Inference*, Duxbury Press, Belmont.
- [2] DeGroot, M.H. (1986). *Probability and Statistics*, 2nd Edition, Addison-Wesley, Reading.
- [3] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, McGraw-Hill, Singapore.

REBECCA WALWYN

Monotonic Regression

JAN DE LEEUW

Volume 3, pp. 1260–1261

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Monotonic Regression

In linear regression, we fit a linear function $y = \alpha + \beta x$ to a scatterplot of n points (x_i, y_i) . We find the parameters α and β by minimizing

$$\sigma(\alpha, \beta) = \sum_{i=1}^n w_i (y_i - \alpha - \beta x_i)^2, \quad (1)$$

where the w_i are known positive weights (*see Multiple Linear Regression*).

In the more general **nonlinear regression** problem, we fit a nonlinear function $\phi_\theta(x)$ by minimizing

$$\sigma(\theta) = \sum_{i=1}^n w_i (y_i - \phi_\theta(x_i))^2 \quad (2)$$

over the parameters θ . In both cases, consequently, we select the minimizing function from a family of functions indexed by a small number of parameters.

In some statistical techniques, low-dimensional parametric models are too restrictive. In nonmetric **multidimensional scaling** [3], for example, we can only use the rank order of the x_i and not their actual numerical values. Parametric methods become useless, but we still can fit the best fitting monotone (increasing) function nonparametrically. Suppose there are no ties in x , and the x_i are ordered such that $x_1 < \dots < x_n$. In monotone regression, we minimize

$$\sigma(z) = \sum_{i=1}^n w_i (y_i - z_i)^2 \quad (3)$$

over z , under the linear inequality restrictions that $z_1 \leq \dots \leq z_n$. If the solution to this problem is $\hat{z}$, then the best fitting increasing function is the set of pairs $(x_i, \hat{z}_i)$. In monotone regression, the number of parameters is equal to the number of observations. The only reason we do not get a perfect solution all the time is because of the order restrictions on z .

Actual computation of the best fitting monotone function is based on the theorem that if $y_i > y_{i+1}$, then $\hat{z}_i = \hat{z}_{i+1}$. In words: if two consecutive values of y are in the wrong order, then the two corresponding consecutive values of the solution $\hat{z}$ will be equal. This basic theorem leads to a simple algorithm, because knowing that two values of $\hat{z}$ must be equal reduces the number of parameters by one. We thus have a monotone regression problem with $n - 1$

parameters. Either the elements are now in the correct order, or there is a violation, in which case we can reduce the problem to one with $n - 2$ parameters, and so on. This process always comes to an end, in the worst possible case when we only have a single parameter left, which is obviously monotone.

We can formalize this in more detail as the up-and-down-blocks algorithm of [4]. It is illustrated in Table 1, in which the first column is y . The first violation we find is $3 > 0$, or 3 is not up-satisfied. We merge the two elements to a block, which contains their weighted average $3/2$ (in our example all weights are one). But now $2 > (3/2)$, and thus the new value $3/2$ is not down-satisfied. We merge all three values to a block of three and find $5/3$, which is both up-satisfied and down-satisfied. We then continue with the next violation. Clearly, the algorithm produces a decreasing number of blocks. The value of the block is computed using weighted averaging, where the weight of a block is the sum of the weights of the elements in the block. In our example, we wind up with only two blocks, and thus the best fitting monotone function $\hat{z}$ is a step function with a single step from $5/3$ to 4.

The result is plotted in Figure 1. The line through the points x and y is obviously the best possible fitting function. The best fitting monotone function, which we just computed, is the step function consisting of the two horizontal lines.

If x has ties, then this simple algorithm does not apply. There are two straightforward adaptations [2]. In the primary approach to ties, we start our monotone regression with blocks of y values corresponding to the ties in x . Thus, we require tied x values to correspond with tied z values. In the secondary approach, we pose no constraints on tied values, and it can be shown that in that case we merely have to order the y values such that they are increasing

Table 1 Computing monotone regression

y	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\hat{z}$
2	2	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$
3	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$
0	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$
6	6	6	6	4
6	6	6	3	4
0	0	0	3	4

2 Monotonic Regression

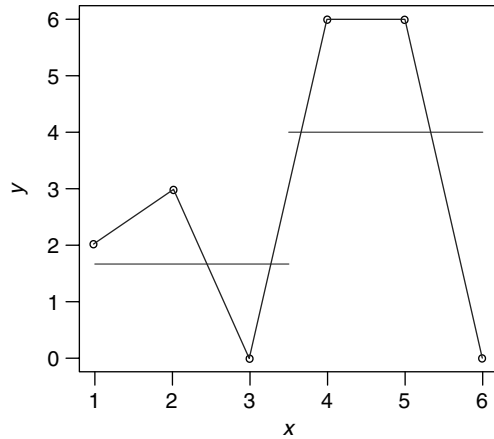


Figure 1 Plotting monotone regression

in blocks of tied x values. And then we perform an ordinary monotone regression.

Monotone regression can be generalized in several important directions. First, basically the same algorithm can be used to minimize any separable function of the form $\sum_{i=1}^n f(y_i - z_i)$, with f any convex function with a minimum at zero. For instance, f

can be the absolute value function, in which case we merge blocks by computing medians instead of means. And second, we can generalize the algorithm from weak orders to partial orders in which some elements cannot be compared; for details, see [1]. Finally, it is sometimes necessary to compute the least squares monotone regression with a nondiagonal weight matrix. In this case, the simple block merging algorithms no longer apply, and more general quadratic programming methods must be used.

References

- [1] Barlow, R.E., Bartholomew, R.J., Bremner, J.M. & Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*, Wiley, New York.
- [2] De Leeuw, J. (1977). Correctness of Kruskal's algorithms for monotone regression with ties, *Psychometrika* **42**, 141–144.
- [3] Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.
- [4] Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: a numerical method, *Psychometrika* **29**, 115–129.

JAN DE LEEUW

Monte Carlo Goodness of Fit Tests

JULIAN BESAG

Volume 3, pp. 1261–1264

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Monte Carlo Goodness of Fit Tests

Monte Carlo P values

It is often necessary, particularly at a preliminary stage of data analysis, to investigate the compatibility between a known multivariate distribution $\{\pi(x) : x \in S\}$ and a corresponding single observation $x^{(1)} \in S$. Here a ‘single observation’ may mean a vector, perhaps corresponding to a random sample from a univariate distribution, or a table or an image or whatever. In exponential families (see **Generalized Linear Models (GLM)**), the requirement that the distribution is known can be achieved by first conditioning on sufficient statistics so as to eliminate the parameters from the original formulation. In frequentist inference, evidence of a conflict between $x^{(1)}$ and π is quantified by the P value obtained by comparing the observed value $u^{(1)}$ of a particular test statistic $u = u(x)$ with its ‘null distribution’ under π . For small datasets, exact calculations can be made, but usually the null distribution of u is intractable analytically and computationally and so asymptotic chi-squared approximations are invoked. However, such approximations are often invalid because the data are too sparse. This is common in analyzing multidimensional contingency tables; see [1, Section 7.1.5] for a 2^5 table in which the conclusion is questionable.

For definiteness, suppose that $x^{(1)}$ is a table and that relatively large values of $u^{(1)}$ indicate a conflict with π . Then an alternative to the above approach is available if a random sample of tables $x^{(2)}, \dots, x^{(m)}$ can be drawn from π , producing values $u^{(2)}, \dots, u^{(m)}$ of the test statistic u . For if $x^{(1)}$ is indeed from π and ignoring for the moment the possibility of ties, the rank of $u^{(1)}$ among $u^{(1)}, \dots, u^{(m)}$ is uniform on $1, \dots, m$. It follows that, if $u^{(1)}$ turns out to be k th largest among all m values, an exact P value k/m can be declared. This procedure, suggested independently in [2] and [8], is called a *Monte Carlo test*, (see **Monte Carlo Simulation**) though there is sometimes confusion with approximate P values obtained by using simulation to estimate the percentiles of the null distribution of u . Both types of P values converge to the P value in the preceding paragraph as $m \rightarrow \infty$.

The choice of m is governed by computational considerations, with $m = 100$ or 1000 or 10000 the

most popular. Note that, if several investigators carry out the same test on the same data $x^{(1)}$, they will generally obtain slightly different P values, despite the fact that marginally each result is exact! Such differences should not be important at a preliminary stage of analysis and disparities diminish as m increases. Ties between ranks can occur with discrete data, in which case one can quote a corresponding range of P values, though one may also eliminate the problem by using a randomized rule. For detailed investigation of Monte Carlo tests when π corresponds to a random sample of n observations from a population, see [11, 13].

A useful refinement is provided by *sequential* Monte Carlo tests [4] (see **Sequential Testing**). First, one specifies a maximum number of simulations $m - 1$, as before, but now additionally a minimum number h , typically 10 or 20. Then $x^{(2)}, \dots, x^{(m)}$ are drawn sequentially from π but with the proviso that sampling is terminated if ever h of the corresponding $u^{(l)}$'s exceed $u^{(1)}$, in which case a P value h/l is declared, where $l \leq m - 1$ is the number of simulations; otherwise, the eventual P value is k/m , as before. See [3] for the validity of this procedure. Sequential tests encourage early termination when there is no evidence against π but continue sampling and produce a finely graduated P value when the evidence against the model is substantial. For example, if the model is correct and one chooses $m = 1000$ and $h = 20$, the expected sample size is reduced to 98.

For more on simple Monte Carlo tests, see, for example [14]. Such tests have been especially useful in the preliminary analysis of spatial data; see, for example, [5] and [7]. The simplest application occurs in assessing whether a spatial point pattern over a perhaps awkwardly shaped study region A is consistent with a homogeneous Poisson process. By conditioning on the observed number of points n , the test is reduced to one of uniformity in which comparisons are made between the data and $m - 1$ realizations of n points placed entirely at random within A , using any choice of test statistic that is sensitive to interesting departures from the Poisson process.

Markov Chain Monte Carlo P values

Unfortunately, it is not generally practicable to generate samples directly from the target distribution π . For example, this holds even when testing for no

three-way interaction in a three-dimensional **contingency table**: here the face totals are the sufficient statistics but it is not known how to generate random samples from the corresponding conditional distribution, except for very small tables. However, it is almost always possible to employ the Metropolis–Hastings algorithm [12, 16] to construct the transition matrix or kernel P of a Markov chain for which π is a stationary distribution. Furthermore, under the null hypothesis, if one seeds the chain by the data $x^{(1)}$, then the subsequent states $x^{(2)}, \dots, x^{(m)}$ are also sampled from π . This provides an advantage over other Markov Chain Monte Carlo (MCMC) applications in which a burn-in phase is required to achieve stationarity. However, there is now the problem that successive states are dependent and so there is no obvious way in which to devise a legitimate P value for the test. Leaving gaps of r steps, where r is large, between each $x^{(t)}$ and $x^{(t+1)}$ or, in other words, replacing P by P^r , reduces the problem but could still lead to serious bias and, in any case, the goal in MCMC methods is to accommodate the dependence rather than effectively eliminate it, which might require prohibitively long runs (*see Markov Chain Monte Carlo and Bayesian Statistics*).

Two remedies that incorporate dependence and yet retain the exact P values of simple Monte Carlo testing are given in [3]. Both involve running the chain *backwards*, as well as forwards, in time. This is possible for any stationary Markov chain via its corresponding backwards transition matrix or kernel Q and is trivial if P is time reversible, because then $Q = P$. Reversibility can always be arranged but we do not assume it here in describing the simpler of two fixes in [3].

Thus, instead of running the chain forwards, suppose we run it backwards from $x^{(1)}$ for r steps, using Q , to obtain a state $x^{(0)}$, say. The value of the integer r is entirely under our control here. We then run the chain forwards from $x^{(0)}$ for r steps, using P , and do this $m-1$ times independently to obtain states $x^{(2)}, \dots, x^{(m)}$ that are contemporaneous with $x^{(1)}$. It is clear that, if $x^{(1)}$ is a draw from π , then so are $x^{(0)}, x^{(2)}, \dots, x^{(m)}$ but not only this: $x^{(1)}, \dots, x^{(m)}$ have an underlying joint distribution that is exchangeable. Moreover, for any choice of test statistic $u = u(x)$, this property must be inherited by the corresponding $u^{(1)}, \dots, u^{(m)}$. Hence, if $x^{(1)}$ is a draw from π , the rank of $u^{(1)}$ among $u^{(1)}, \dots, u^{(m)}$ is once again uniform and provides an exact P value,

just as for a simple Monte Carlo test. The procedure is rigorous because P values are calculated on the basis of a correct model.

Note that $x^{(0)}$ must be ignored and that also it is not permissible to generate separate $x^{(0)}$'s, else $x^{(2)}, \dots, x^{(m)}$, although exchangeable with each other, are no longer exchangeable with $x^{(1)}$. The value of r should be large enough to provide ample scope for mobility around the state space S , so that simulations can reach more probable parts of S when the formulation is inappropriate. That is, larger values of r tend to improve the power of the test, although the P value itself is valid for any value of r , apart from dealing with ties. Note that it is not essential for validity of the exact P value that P be irreducible. However, this may lead to a loss of power in the test. Irreducibility fails or is in question for many applications to multidimensional contingency tables: the search for better algorithms is currently a hot topic in computational algebra. Finally, see [4] for sequential versions of both procedures in [3].

Example Exact P values for the Rasch Model

Consider an $r \times s$ table of binary variables x_{ij} . For example, in educational testing, $x_{ij} = 0$ or 1 corresponds to the correct (1) or incorrect (0) response of candidate i to item j . See [6] for two well-known LSAT datasets, each with 1000 candidates and 5 questions. For such tables, the **Rasch model** [17] asserts that all responses are independent and that the odds of 1 to 0 in cell (i, j) are $\theta_{ij} : 1$, with $\theta_{ij} = \phi_i \psi_j$, where the ϕ_i 's and ψ_j 's are unknown parameters that can be interpreted as measuring the relative aptitude of the candidates and difficulties of the items, respectively. The probability of a table x is then

$$\prod_{i=1}^r \prod_{j=1}^c \frac{\theta_{ij}^{x_{ij}}}{1 + \theta_{ij}} = \frac{\prod_i \phi_i^{x_{i+}} \prod_j \psi_j^{x_{+j}}}{\prod_i \prod_j (1 + \phi_i \psi_j)} \quad (1)$$

and the row and column totals x_{i+} and x_{+j} are sufficient statistics for the ϕ_i 's and ψ_j 's. Thus, if we condition on the row and column totals, the ϕ_i 's and ψ_j 's are eliminated and we obtain a uniform distribution $\pi(x)$ on the space of tables with the same x_{i+} 's and x_{+j} 's. However, this space is generally huge and enumeration is out of the question; nor are simple Monte Carlo tests available.

Binary tables also occur in evolutionary biology, with x_{ij} identifying presence or absence of species i in location j ; see [10, 15], for example. Here the Rasch model accommodates differences between species and differences between locations, but departures from it can suggest competition between the species.

To construct a test for the Rasch model via MCMC, we require an algorithm that maintains a uniform distribution π on the space S of binary tables x with the same row and column totals as in the data $x^{(1)}$. The simplest move that preserves the margins is depicted below, where $a, b = 0$ or 1 . The two row indices and the two column indices are the same on

$$\begin{array}{ccccccc} & \vdots & & \vdots & & \vdots & \\ \cdots & a & \cdots & b & \cdots & & \\ & \vdots & & \vdots & & \vdots & \\ \cdots & b & \cdots & a & \cdots & & \\ & \vdots & & \vdots & & \vdots & \end{array} \rightarrow \begin{array}{ccccccc} & \vdots & & \vdots & & \vdots & \\ \cdots & b & \cdots & a & \cdots & & \\ & \vdots & & \vdots & & \vdots & \\ \cdots & a & \cdots & b & \cdots & & \\ & \vdots & & \vdots & & \vdots & \end{array}$$

the right as on the left. Of course, there is no change in the configuration unless $a \neq b$. It can be shown that any table in S can be reached from any other by a sequence of such switches, so that irreducibility is guaranteed. Among several possible ways in which the algorithm can proceed (see [3]), the simplest is to repeatedly choose two rows and two columns at random and to propose the corresponding swap if this is valid or retain the current table if it is not. This defines a Metropolis algorithm and, since π is uniform, all proposals are accepted.

Closing Comments

We close with some brief additional comments on MCMC tests. As regards the test statistic $u(x)$, the choice should reflect the main alternatives that one has in mind. For example, in educational testing, interest might center on departures from the Rasch model caused by correlation between patterns of correct or incorrect responses to certain items. Then $u(x)$ might be a function of the *coincidence matrix*, whose (j, j') element is the frequency with which candidates provide the same response to items j and j' . Corresponding statistics are easy to define and can provide powerful tools, but note that the total score in the matrix is no use because it is fixed by the row and column totals of the data. For ecologic applications, [10] provides an interesting discussion and advocates using a statistic based on the *co-occurrence matrix*.

Second, it is often natural to apply several different statistics to the same data: there is no particular objection to this at an exploratory stage, provided that all the results are reported. Finally, we caution against some misleading claims. Thus, the Knight's move algorithm [18] for the Rasch model is simply incorrect: see [10]. Also, some MCMC tests are referred to as 'exact' when in fact they do not apply either of the corrections described in [3] and are therefore approximations; see [9], for example.

References

- [1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley.
- [2] Barnard, G.A. (1963). Discussion of paper by M. S. Bartlett, *Journal of the Royal Statistical Society. Series B* **25**, 294.
- [3] Besag, J.E. & Clifford, P. (1989). Generalized Monte Carlo significance tests, *Biometrika* **76**, 633–642.
- [4] Besag, J.E. & Clifford, P. (1991). Sequential Monte Carlo p -values, *Biometrika* **78**, 301–304.
- [5] Besag, J.E. & Diggle, P.J. (1977). Simple Monte Carlo tests for spatial pattern, *Applied Statistics* **26**, 327–333.
- [6] Boch, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika* **35**, 179–197.
- [7] Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.
- [8] Dwass, M. (1957). Modified randomization tests for non-parametric hypotheses, *Annals of Mathematical Statistics* **28**, 181–187.
- [9] Forster, J.J., McDonald, J.W. & Smith, P.W.F. (2003). Markov chain Monte Carlo exact inference for binomial and multinomial logistic regression models, *Statistics and Computing* **13**, 169–177.
- [10] Gotelli, N.J. & Entsminger, G.L. (2001). Swap and fill algorithms in null model analysis: rethinking the knight's tour, *Oecologia* **129**, 281–291.
- [11] Hall, P. & Titterton, D.M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests, *Journal of the Royal Statistical Society. Series B* **51**, 459–467.
- [12] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.
- [13] Jöckel, K.-H. (1986). Finite-sample properties and asymptotic efficiency of Monte Carlo tests, *Annals of Statistics* **14**, 336–347.
- [14] Manly, B.F.J. (1991). *Randomization and Monte Carlo Methods in Biology*, Chapman & Hall, London.
- [15] Manly, B.F.J. (1995). A note on the analysis of species co-occurrences, *Ecology* **76**, 1109–1115.
- [16] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state

4 Monte Carlo Goodness of Fit Tests

- calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.
- [17] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Educational Research Institute, Copenhagen.
- [18] Sanderson, J.G., Moulton, M.P. & Selfridge, R.G. (1998). Null matrices and the analysis of species co-occurrences, *Oecologia* **116**, 275–283.

JULIAN BESAG

Monte Carlo Simulation

JAMES E. GENTLE

Volume 3, pp. 1264–1271

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Monte Carlo Simulation

Introduction

Monte Carlo methods use random processes to estimate mathematical or physical quantities, to study distributions of random variables, to study and compare statistical procedures, and to study the behavior of complex systems. Monte Carlo methods had been used occasionally by statisticians for many years, but with the development of high-speed computers, Monte Carlo methods became viable alternatives to theoretical and experimental methods in studying complicated physical processes. The random samples used in a Monte Carlo method are generated on the computer, and are more properly called “pseudorandom numbers”.

An early example of how a random process could be used to evaluate a fixed mathematical quantity is the Buffon needle problem. The French naturalist Comté de Buffon showed that the probability that a needle of length l thrown randomly onto a grid of parallel lines with distance d ($\geq l$) apart intersects a line is $2l/(\pi d)$. The value of π can, therefore, be estimated by tossing a needle onto a lined grid many times and counting the number of times the needle crosses one of the lines. (See [2], pp. 274, 275 for discussion of the problem and variations on the method.) A key element of the Buffon needle problem is that there is no intrinsic random element; randomness is introduced to study the deterministic problem of evaluating a mathematical constant.

The idea of simulating a random process to study its distributional properties is so basic and straightforward that these methods were used in very early studies of probability distributions. An early documented use of a Monte Carlo method was by the American statistician Erastus Lyman De Forest in 1876 in a study of smoothing a time series (see [4]). Another important early use of Monte Carlo was by “Student” (see **Gosset, William Sealy**) in studying the distributions of the correlation coefficient and of the t statistic (see **Catalogue of Probability Density Functions**). Student used actual biometric data to simulate realizations of normally distributed random variables.

Monte Carlo Evaluation of an Integral

In its simplest form, Monte Carlo simulation is the evaluation of a definite integral

$$\theta = \int_D f(x) \, dx \quad (1)$$

by identifying a random variable Y with support on D and density $p(y)$ and a function g such that the expected value of $g(Y)$ is θ :

$$\begin{aligned} E(g(Y)) &= \int_D g(y)p(y) \, dy \\ &= \int_D f(y) \, dy \\ &= \theta. \end{aligned} \quad (2)$$

In the simplest case, D is the interval $[a, b]$, Y is taken to be a random variable with a uniform density over $[a, b]$; that is, $p(y)$ in (2) is the constant uniform density. In this case,

$$\theta = (b - a)E(f(Y)). \quad (3)$$

The problem of evaluating the integral becomes the familiar statistical problem of estimating a mean, $E(f(Y))$. From a sample of size m , a good estimate of θ is the sample mean,

$$\hat{\theta} = (b - a) \frac{\sum_{i=1}^m f(y_i)}{m}, \quad (4)$$

where the y_i are values of a random sample from a uniform distribution over (a, b) . The estimate is unbiased (see **Estimation**):

$$\begin{aligned} E(\hat{\theta}) &= (b - a) \frac{\sum E(f(Y_i))}{m} \\ &= (b - a)E(f(Y)) \\ &= \int_a^b f(x) \, dx. \end{aligned} \quad (5)$$

The variance is

$$V(\hat{\theta}) = (b - a)^2 \frac{\sum V(f(Y_i))}{m^2}$$

2 Monte Carlo Simulation

$$\begin{aligned}
 &= \frac{(b-a)^2}{m} V(f(Y)) \\
 &= \frac{(b-a)}{m} \int_a^b \left(f(x) - \int_a^b f(t) dt \right)^2 dx. \quad (6)
 \end{aligned}$$

The integral in (6) is a measure of the *roughness* of the function.

Consider again the problem of evaluation of the integral in (1) that has been rewritten as in (2). Now suppose that we can generate m random variates y_i from the distribution with density p . Then our estimate of θ is just

$$\hat{\theta} = \frac{\sum g(y_i)}{m}. \quad (7)$$

Compare this estimator with the estimator in (4).

The use of a probability density as a weighting function allows us to apply the Monte Carlo method to improper integrals (that is, integrals with infinite ranges of integration). The first thing to note, therefore, is that the estimator (7) applies to integrals over general domains, while the estimator (4) applies only to integrals over finite intervals. Another important difference is that the variance of the estimator in (7) is likely to be smaller than that of the estimator in (4).

The square root of the variance (that is, the standard deviation of the estimator) is a good measure of the range within which different realizations of the estimator of the integral may fall. Under certain assumptions, using the standard deviation of the estimator, we can define statistical “**confidence intervals**” for the true value of the integral θ . Loosely speaking, a confidence interval is an interval about an estimator $\hat{\theta}$ that in repeated sampling would include the true value θ a specified portion of the time. (The specified portion is the “level” of the confidence interval and is often chosen to be 90% or 95%. Obviously, all other things being equal, the higher the level of confidence, the wider the interval must be.)

Because of the dependence of the confidence interval on the standard deviation, the standard deviation is sometimes called a “probabilistic error bound”. The word “bound” is misused here, of course, but in any event, the standard deviation does provide some measure of a sampling “error”.

From (6), we note that the order of error in terms of the Monte Carlo sample size is $O(m^{-1/2})$. This results in the usual diminished returns of ordinary

statistical estimators; to halve the error, the sample size must be quadrupled.

An important property of the standard deviation of a Monte Carlo estimate of a definite integral is that the order in terms of the number of function evaluations is independent of the dimensionality of the integral. On the other hand, the usual error bounds for numerical quadrature are $O(m^{-2/d})$, where d is the dimensionality. For one or two dimensions, it is generally better to use one of the standard methods of numerical quadrature, such as Newton–Cotes methods, extrapolation or Romberg methods, and Gaussian quadrature, rather than Monte Carlo quadrature.

Experimental Error in Monte Carlo Methods

Monte Carlo methods are sampling methods; therefore, the estimates that result from Monte Carlo procedures have associated *sampling errors*. The fact that the estimate is not equal to its expected value (assuming that the estimator is unbiased) is not an “error” or a “mistake”; it is just a result of the variance of the random (or pseudorandom) data. Monte Carlo methods are experiments using random data. The variability of the random data results in *experimental error*, just as in other scientific experiments in which randomness is a recognized component.

As in any statistical estimation problem, an estimate should be accompanied by an estimate of its variance. The estimate of the variance of the estimator of interest is usually just the sample variance of computed values of the estimator of interest.

Following standard practice, we could use the square root of the variance (that is, the standard deviation) of the Monte Carlo estimator to form an approximate confidence interval for the integral being estimated.

In reporting numerical results from Monte Carlo simulations, it is mandatory to give some statement of the level of the experimental error. An effective way of doing this is by giving the sample standard deviation. When a number of results are reported, and the standard deviations vary from one to the other, a good way of presenting the results is to write the standard deviation in parentheses beside the result itself, for example,

$$3.147 (0.0051).$$

Notice that if the standard deviation is of order 10^{-3} , the precision of the main result is not greater than 10^{-3} . Just because the computations are done at a higher precision is no reason to write the number as if it had more significant digits.

Variance of Monte Carlo Estimators

The variance of a Monte Carlo estimator has important uses in assessing the quality of the estimate of the integral. The expression for the variance, as in (6), is likely to be very complicated and to contain terms that are unknown. We therefore need methods for estimating the variance of the Monte Carlo estimator.

A Monte Carlo estimate usually has the form of the estimator of θ in (4):

$$\hat{\theta} = c \frac{\sum f_i}{m}. \quad (8)$$

The variance of the estimator has the form of (6):

$$V = (\hat{\theta})^2 c^2 \int \left(f(x) - \int f(t) dt \right)^2 dx.$$

An estimator of the variance is

$$\hat{V}(\hat{\theta}) = c^2 \frac{\sum (f_i - \bar{f})^2}{m-1}. \quad (9)$$

This estimator is appropriate only if the elements of the set of random variables $\{F_i\}$, on which we have observations $\{f_i\}$, are (assumed to be) independent and thus have zero correlations.

Our discussion of variance in Monte Carlo methods that are based on pseudorandom numbers follows the pretense that the numbers are realizations of random variables, and the main concern in pseudorandom number generation is the simulation of a sequence of i.i.d. random variables. In quasirandom number generation, the attempt is to get a sample that is spread out over the sample space more evenly than could be expected from a random sample. Monte Carlo methods based on quasirandom numbers, or “quasi-Monte Carlo” methods, do not admit discussion of variance in the technical sense.

Variance Reduction

An objective in sampling is to reduce the variance of the estimators while preserving other good qualities, such as unbiasedness. Variance reduction results in statistically efficient estimators. The emphasis on efficient Monte Carlo sampling goes back to the early days of digital computing, but the issues are just as important today (or tomorrow) because, presumably, we are solving bigger problems. The general techniques used in statistical sampling apply to Monte Carlo sampling, and there is a mature theory for sampling designs that yield efficient estimators.

Except for straightforward analytic reduction, discussed in the next section, techniques for reducing the variance of a Monte Carlo estimator are called “swindles” (especially if they are thought to be particularly clever). The common thread in variance reduction is to use additional information about the problem in order to reduce the effect of random sampling on the variance of the observations. This is one of the fundamental principles of all statistical design.

Analytic Reduction

The first principle in estimation is to use any known quantity to improve the estimate. For example, suppose that the problem is to evaluate the integral

$$\theta = \int_D f(x) dx \quad (10)$$

by Monte Carlo methods. Now, suppose that D_1 and D_2 are such that $D_1 \cup D_2 = D$ and $D_1 \cap D_2 = \emptyset$, and consider the representation of the integral

$$\begin{aligned} \theta &= \int_{D_1} f(x) dx + \int_{D_2} f(x) dx \\ &= \theta_1 + \theta_2. \end{aligned} \quad (11)$$

Now, suppose that a part of this decomposition of the original problem is known (that is, suppose that we know θ_1). It is very likely that it would be better to use Monte Carlo methods only to estimate θ_2 and take as our estimate of θ the sum of the known θ_1 and the estimated value of θ_2 . This seems intuitively obvious, and it is generally true unless there is some relationship between $f(x_1)$ and $f(x_2)$, where x_1 is in D_1 and x_2 is in D_2 . If there is some known relationship, however, it may be possible to improve

4 Monte Carlo Simulation

the estimate $\hat{\theta}_2$ of θ_2 by using a transformation of the same random numbers used for $\hat{\theta}_1$ to estimate θ_1 . For example, if $\hat{\theta}_1$ is larger than the known value of θ_1 , the proportionality of the overestimate, $(\hat{\theta}_1 - \theta_1)/\theta_1$, may be used to adjust $\hat{\theta}_2$. This is the same principle as ratio or regression estimation in ordinary sampling theory.

Stratified Sampling and Importance Sampling

In stratified sampling (see **Stratification**), certain proportions of the total sample are taken from specified regions (or “strata”) of the sample space. The objective in stratified sampling may be to ensure that all regions are covered. Another objective is to reduce the overall variance of the estimator by sampling more heavily where the function is rough; that is, where the values $f(x_i)$ are likely to exhibit a lot of variability.

Stratified sampling is usually performed by forming distinct subregions with different importance functions in each. This is the same idea as in analytic reduction except that Monte Carlo sampling is used in each region.

Stratified sampling is based on exactly the same principle in sampling methods in which the allocation is proportional to the variance (see [3]). In some of the literature on Monte Carlo methods, stratified sampling is called “geometric splitting”.

In *importance sampling*, just as may be the case in stratified sampling, regions corresponding to large values of the integrand are sampled more heavily. In importance sampling, however, instead of a finite number of regions, we allow the relative sampling density to change continuously. This is accomplished by careful choice of p in the decomposition implied by (2). We have

$$\begin{aligned}\theta &= \int_D f(x) dx \\ &= \int_D \frac{f(x)}{p(x)} p(x) dx,\end{aligned}\quad (12)$$

where $p(x)$ is a probability density over D . The density $p(x)$ is called the *importance function*. Stratified sampling can be thought of as importance sampling in which the importance function is composed of a mixture of densities. In some of the literature on Monte Carlo methods, stratified sampling and importance sampling are said to use “weight windows”.

From a sample of size m from the distribution with density p , we have the estimator,

$$\hat{\theta} = \frac{1}{m} \sum \frac{f(x_i)}{p(x_i)}. \quad (13)$$

Generating the random variates from the distribution with density p weights the sampling into regions of higher probability with respect to p . By judicious choice of p , we can reduce the variance of the estimator.

The variance of the estimator is

$$V(\hat{\theta}) = \frac{1}{m} V\left(\frac{f(X)}{p(X)}\right), \quad (14)$$

where the variance is taken with respect to the distribution of the random variable X with density $p(x)$. Now,

$$V\left(\frac{f(X)}{p(X)}\right) = E\left(\frac{f^2(X)}{p^2(X)}\right) - \left(E\left(\frac{f(X)}{p(X)}\right)\right)^2. \quad (15)$$

The objective in importance sampling is to choose p so that this variance is minimized. Because

$$\left(E\left(\frac{f(X)}{p(X)}\right)\right)^2 = \left(\int_D f(x) dx\right)^2, \quad (16)$$

the choice involves only the first term in the expression for the variance. By Jensen’s inequality, we have a lower bound on that term:

$$\begin{aligned}E\left(\frac{f^2(X)}{p^2(X)}\right) &\geq \left(E\left(\frac{|f(X)|}{p(X)}\right)\right)^2 \\ &= \left(\int_D |f(x)| dx\right)^2.\end{aligned}\quad (17)$$

That bound is obviously achieved when

$$p(x) = \frac{|f(x)|}{\int_D |f(x)| dx}. \quad (18)$$

Of course, if we knew $\int_D |f(x)| dx$, we would probably know $\int_D f(x) dx$ and would not even be considering the Monte Carlo procedure to estimate the integral. In practice, for importance sampling we would seek a probability density p that is nearly proportional to $|f|$; that is, such that $|f(x)|/p(x)$ is nearly constant.

The problem of choosing an importance function is very similar to the problem of choosing a majorizing function for the acceptance/rejection method. Selection of an importance function involves the principles of function approximation with the added constraint that the approximating function be a probability density from which it is easy to generate random variates.

Let us now consider another way of developing the estimator (13). Let $h(x) = f(x)/p(x)$ (where $p(x)$ is positive; otherwise, let $h(x) = 0$) and generate $y_1, \dots, y_m$ from a density $g(y)$ with support D . Compute *importance weights*,

$$w_i = \frac{p(y_i)}{g(y_i)}, \quad (19)$$

and form the estimate of the integral as

$$\hat{\theta} = \frac{1}{m} \frac{\sum w_i h(y_i)}{\sum w_i}. \quad (20)$$

In this form of the estimator, $g(y)$ is a trial density, just as in the acceptance/rejection methods. This form of the estimator has similarities to weighted resampling.

By the same reasoning as above, we see that the trial density should be “close” to f ; that is, optimally, $g(x) = c|f(x)|$ for some constant c .

Although the variance of the estimator in (13) and (20) may appear rather simple, the term $E((f(X)/p(X))^2)$ could be quite large if p (or g) becomes small at some point where f is large. Of course, the objective in importance sampling is precisely to prevent that, but if the functions are not well-understood, it may happen. An element of the Monte Carlo sample at a point where p is small and f is large has an unduly large influence on the overall estimate. Because of this kind of possibility, importance sampling must be used with some care. (See [2], chapter 7, for further discussion the method.)

Use of Covariates

Another way of reducing the variance, just as in ordinary sampling, is to use covariates. Any variable that is correlated with the variable of interest has potential value in reducing the variance of the estimator. Such a variable is useful if it is easy to generate and

if it has properties that are known or that can be computed easily. In the general case in Monte Carlo sampling, covariates are called control variates. Two special cases are called antithetic variates and common variates. We first describe the general case, and then the two special cases. We then relate the use of covariates to the statistical method sometimes called “Rao-Blackwellization”.

Control Variates. Suppose that Y is a random variable, and the Monte Carlo method involves estimation of $E(Y)$. Suppose that X is a random variable with known expectation, $E(X)$, and consider the random variable

$$\tilde{Y} = Y - b(X - E(X)). \quad (21)$$

The expectation of $\tilde{Y}$ is the same as that of Y , and its variance is

$$V(\tilde{Y}) = V(Y) - 2b\text{Cov}(Y, X) + b^2V(X). \quad (22)$$

For reducing the variance, the optimal value of b is $\text{Cov}(Y, X)/V(X)$. With this choice $V(\tilde{Y}) < V(Y)$ as long as $\text{Cov}(Y, X) \neq 0$. Even if $\text{Cov}(Y, X)$ is not known, there is a b that depends only on the sign of $\text{Cov}(Y, X)$ for which the variance of $\tilde{Y}$ is less than the variance of Y .

The variable X is called a *control variate*. This method has long been used in survey sampling, where $\tilde{Y}$ in (21) is called a regression estimator.

Use of these facts in Monte Carlo methods requires identification of a control variable X that can be simulated simultaneously with Y . If the properties of X are not known but can be estimated (by Monte Carlo methods), the use of X as a control variate can still reduce the variance of the estimator.

These ideas can obviously be extended to more than one control variate:

$$\begin{aligned} \tilde{Y} = & Y - b_1(X_1 - E(X_1)) - \dots \\ & - b_k(X_k - E(X_k)). \end{aligned} \quad (23)$$

The optimal values of the b_i s depend on the full variance-covariance matrix. The usual regression estimates for the coefficients can be used if the variance-covariance matrix is not known.

Identification of appropriate control variates often requires some ingenuity, although in some special cases, there may be techniques that are almost always applicable.

Antithetic Variates. Again consider the problem of estimating the integral

$$\theta = \int_a^b f(x) dx \quad (24)$$

by Monte Carlo methods. The standard crude Monte Carlo estimator, (4), is $(b-a) \sum f(x_i)/n$, where x_i is uniform over (a, b) . It would seem intuitively plausible that our estimate would be subject to less sampling variability if, for each x_i , we used its “mirror”

$$\tilde{x}_i = a + (b - x_i). \quad (25)$$

This mirror value is called an antithetic variate, and use of antithetic variates can be effective in reducing the variance of the Monte Carlo estimate, especially if the integral is nearly uniform. For a sample of size n , the estimator is

$$\frac{b-a}{n} \sum_{i=1}^n (f(x_i) + f(\tilde{x}_i)).$$

The variance of the sum is the sum of the variances plus twice the covariance. Antithetic variates have negative covariances, thus reducing the variance of the sum.

Antithetic variates from distributions other than the uniform can also be formed. The linear transformation that works for uniform antithetic variates cannot be used, however. A simple way of obtaining negatively correlated variates from other distributions is just to use antithetic uniforms in the inverse CDF. If the variates are generated using acceptance/rejection, antithetic variates can be used in the majorizing distribution.

Common Variates. Often, in Monte Carlo simulation, the objective is to estimate the differences in parameters of two random processes. The two parameters are likely to be positively correlated. If that is the case, then the variance in the individual differences is likely to be smaller than the variance of the difference of the overall estimates.

Suppose, for example, that we have two statistics, T and S , that are unbiased estimators of some parameter of a given distribution. We would like to know the difference in the variances of these estimators,

$$V(T) - V(S)$$

(because the one with the smaller variance is better). We assume that each statistic is a function of a

random sample: $\{x_1, \dots, x_n\}$. A Monte Carlo estimate of the variance of the statistic T for a sample of size n is obtained by generating m samples of size n from the given distribution, computing T_i for the i th sample, and then computing

$$\hat{V}(T) = \frac{\sum_{i=1}^m (T_i - \bar{T})^2}{m-1}. \quad (26)$$

Rather than doing this for T and S separately, using the unbiasedness, we could first observe

$$V(T) - V(U) = E(T^2) - E(S^2) = E(T^2 - S^2) \quad (27)$$

and hence estimate the latter quantity. Because the estimators are likely to be positively correlated, the variance of the Monte Carlo estimator $\hat{E}(T^2 - S^2)$ is likely to be smaller than the variance of $\hat{V}(T) - \hat{V}(S)$. If we compute $T^2 - S^2$ from each sample (that is, if we use *common variates*), we are likely to have a more precise estimate of the difference in the variances of the two estimators, T and S .

Rao-Blackwellization. As in the discussion of control variates above, suppose that we have two random variables Y and X and we want to estimate $E(f(Y, X))$ with an estimator of the form $T = \sum f(Y_i, X_i)/m$. Now suppose that we can evaluate $E(f(Y, X)|X = x)$. (This is similar to what is done in using (21) above.) Now, $E(E(f(Y, X)|X = x)) = E(f(Y, X))$, so the estimator

$$\tilde{T} = \sum \frac{E(f(Y_i, X)|X = x_i)}{m} \quad (28)$$

has the same expectation as T . However, we have

$$V(f(Y, X)) = V(E(f(Y, X)|X = x)) + E(V(f(Y, X)|X = x)); \quad (29)$$

that is,

$$V(f(Y, X)) \geq V(E(f(Y, X)|X = x)). \quad (30)$$

Therefore, $\tilde{T}$ is preferable to T because it has the same expectation but no larger variance. (The function f may depend on Y only. In that case, if Y and X are independent we can gain nothing.)

The principle of minimum variance unbiased estimation leads us to consider statistics such as $\tilde{T}$ conditioned on other statistics. The Rao-Blackwell

Theorem (see any text on mathematical statistics) tells us that if a sufficient statistic exists, the greatest improvement in variance while still requiring unbiasedness occurs when the conditioning is done with respect to a sufficient statistic. This process of conditioning a given estimator on another statistic is called Rao-Blackwellization. (This name is often used even if the conditioning statistic is not sufficient.)

Applications of Monte Carlo Simulation

Monte Carlo simulation is widely used in many fields of science and business. In the physical sciences, Monte Carlo methods were first employed on a major scale in the 1940s, and their use continues to grow. System simulation has been an important methodology in operations research since the 1960s. In more recent years, simulation of financial processes has become an important tool in the investments industry.

Monte Carlo simulation has two distinct applications in statistics. One is in the study of statistical methods, and the other is as a part of a statistical method for analysis of data.

Monte Carlo Studies of Statistical Methods. The performance of a statistical method, such as a t Test, for example, depends, among other things, on the underlying distribution of the sample to which it is applied. For simple distributions and for simple statistical procedures, it may be possible to work out analytically such things as the power of a test or the exact distribution of a test statistic or estimator. In more complicated situations, however, these properties cannot be derived analytically. The properties can be studied by Monte Carlo, however. The procedure is simple; we merely simulate on the computer many samples from the assumed underlying distribution, compute the statistic of interest from each sample, and use the sample of computed statistics to assess its sampling distribution. There is a wealth of literature and software for the generation of random numbers required in first step in this process (see, for example, [2]).

Monte Carlo simulation to study statistical methods is most often employed in the comparison of methods, frequently in the context of robustness studies. It is relatively easy to compare the relative performance of say a t Test with a sign test under a

wide range of scenarios by generating multiple samples under each scenario and evaluating the t Test and the sign test for each sample in the given scenario.

This application of Monte Carlo simulation is so useful that it is employed by a large proportion of the research articles published in statistics. (In the 2002 volume of the *Journal of the American Statistical Association*, for example, more than 80% of the articles included Monte Carlo (see **Bayesian Statistics**) studies of the performance of the statistical methods.)

Monte Carlo Methods in Data Analysis. In computational statistics, Monte Carlo methods are used as part of the overall methodology of data analysis. Examples include Monte Carlo tests, Monte Carlo **bootstrapping**, and **Markov chain Monte Carlo** for the evaluation of Bayesian posterior distributions.

A Monte Carlo test of an hypothesis, just as any statistical hypothesis test, uses a random sample of observed data. As with any statistical test, a test statistic is computed from the observed sample. In the usual statistical tests, the computed test statistic is compared with the quantiles of the distribution of the test statistic under the null hypothesis, and the null hypothesis is rejected if the computed value is deemed sufficiently extreme. Often, however, we may not know the distribution of the test statistic under the null hypothesis. In this case, if we can simulate random samples from a distribution specified by the null hypothesis, we can simulate the distribution of the test statistic by generating many such samples, and computing the test statistic for each one. We then compare the observed value of the test statistic with the ones computed from the simulated random samples. Just as in the standard methods of statistical hypothesis testing, we reject the null hypothesis if the observed value of the test statistic is deemed sufficiently extreme. This is called a Monte Carlo test. Somewhat surprisingly, a Monte Carlo test needs only a fairly small number of random samples to be relatively precise. In most cases, only 100 or so samples are adequate. See chapters 2 through 4 of [1] for further discussion of Monte Carlo tests and other Monte Carlo methods in statistical data analysis.

References

- [1] Gentle, J.E. (2002). *Elements of Computational Statistics*, Springer, New York.

8 Monte Carlo Simulation

- [2] Gentle, J.E. (2003). *Random Number Generation and Monte Carlo Methods*, 2nd Edition, Springer, New York. (See also **Randomization Based Tests**)
- [3] Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York. JAMES E. GENTLE
- [4] Stigler, S.M. (1978). Mathematical statistics in the early states, *Annals of Statistics* **6**, 239–265.

Multidimensional Item Response Theory Models

TERRY A. ACKERMAN

Volume 3, pp. 1272–1280

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multidimensional Item Response Theory Models

Many educational and psychological tests are inherently multidimensional, meaning these tests measure two or more dimensions or constructs [27]. A construct is a theoretical representation of the underlying trait, concept, attribute, processes, and/or structure that the test is designed to measure [16]. Tests that are composed of items each measuring the same construct, or same composite of multiple constructs, are considered to be unidimensional. If, however, different items are measuring different constructs, or different composites of multiple constructs, the test can be considered to be multidimensional. It is important to distinguish between construct-irrelevant or invalid traits that are being measured versus those traits that are valid and replicable [5].

If a test is unidimensional, then it is appropriate to report examinee performance on the test as a single score. If a test is multidimensional, then reporting examinee results is more problematic. In some cases, if the skills are distinct, a profile of scores may be most appropriate. If the items are measuring similar composites of skills, then a single score may suffice. Problems of how to report results can easily arise. For example, consider a test in which easy items measure skill A and difficult items measure skill B. If the results of this test are reported on a single score scale, comparing results could be impossible because low scores represent differences in skill A and high scores represent differences in skill B.

Response data represent the interaction between a group of examinees and a set of items. Surprisingly, an assessment can be either unidimensional or multidimensional depending on the set of skills inherent in a particular group of examinees who take the test. Consider the following two items:

Item 1: If $500 - 4X = 138$, then $X = ?$

Item 2: Janelle went to the store and bought four pieces of candy. She gave the clerk \$5.00 and received \$1.38 back. How much did one piece of candy cost?

Item 1 requires the examinee to use algebra to solve the linear equation. Item 2, a ‘story problem’, requires the examinee to read the scenario, translate the text to an algebraic expression, and then solve. If examinees vary in the range of reading skill required to read and translate the text in item 2, then the item has the potential to measure a composite of both reading and algebra. If all the examinees vary in reading skill, but in a range beyond the level of reading required by item 2, then this item will most likely distinguish only between levels of examinee proficiency in algebra.

One item is always unidimensional. However, two or more items, each measuring a different composite of, say, algebra and reading, have the potential to yield multidimensional data. Ironically, the same items administered to one group of examinees may result in unidimensional data, yet when given to another group of examinees, may yield multidimensional data.

Determining if Data are Multidimensional

The first step in any multidimensional item response theory (MIRT) analysis (*see Item Response Theory (IRT) Models for Dichotomous Data*) is to determine whether the data are indeed multidimensional. Dimensionality and MIRT analyses should be supported by, and perhaps even preempted with, substantive judgment. A thorough analysis of the knowledge and skills needed to successfully respond to each item should be conducted. It might be helpful to conduct this substantive analysis by referring to the test specifications and the opinions of experts who have extensive knowledge of the content and of the examinees’ cognitive skills. If subsets of items measure different content knowledge and/or cognitive skills, then these items have the potential to represent distinct dimensions if given to a group of examinees that vary on these skills.

Many empirical methods have been proposed to investigate the dimensionality of test data (e.g., [10, 11]). These quantitative methods range from linear factor analysis to several nonparametric methods [9, 12, 17, 18, 25, 32]. Unfortunately, most of the dimensionality tools available to practitioners are exploratory in nature, and many of these tools produce results that contradict substantive dimensionality hypotheses [10].

Factor analysis is a data reduction technique that uses the inter-item correlations to distinguish a small set of underlying skills or factors. A factor can be substantively identified by noting which items load highly on the factor. Additionally, scree plots of **eigenvalues** can be used to determine if more than one factor is necessary to account for the total variance observed in the test scores. Unfortunately, there is no one method that psychometricians can agree upon to determine how large eigenvalues have to be to indicate a set of test data are indeed multidimensional.

A nonparametric approach that some researchers have found useful is **hierarchical cluster analysis** [21]. This approach uses proximity matrices (*see Proximity Measures*) and clustering rules to form homogeneous groups of items. This is an iterative procedure. For an n -item test, the procedure will form n -clusters, then $n - 1$ cluster, and so on until all the items are combined into a single cluster. Researchers often examine the two- or three-cluster solutions to determine if these solutions define identifiable traits. One drawback with this approach is that because the solution for each successive iteration of the algorithm is dependent on the previous solution, the solution attained at one or more levels may not be optimal.

A relatively new approach is a procedure called *DETECT* [32]. *DETECT* is an exploratory nonparametric dimensionality assessment procedure that estimates the number of dominant dimensions present in a data set and the magnitude of the departure from unidimensionality using a genetic algorithm. *DETECT* also identifies the dominant dimension measured by each item. This procedure produces mutually exclusive, dimensionally homogeneous clusters of items.

Perhaps one of the more promising nonparametric approaches for determining if two groups of items are dimensionally distinct is the program *DIMTEST* based on the work of Stout [24]. Hypotheses about multiple dimensions that are formulated using substantive analysis, factor analysis, cluster analysis, or *DETECT* can be tested using *DIMTEST*. This program provides a statistical test of significance, which can verify that test data are not unidimensional. Once it has been determined that the test data are multidimensional, practitioners need to determine which multidimensional model best describes the response process for predicting correct response on the individual items.

MIRT Models

Reckase [19] defined the probability of a correct response for subject j on item i on a compensatory model as

$$P_C(u_{ij} = 1 | \theta_{j1}, \theta_{j2}) = \frac{1}{1 + e^{a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + d_j}} \quad (1)$$

where u_{ij} is the dichotomous score (0 = wrong, 1 = correct),

θ_{j1} and θ_{j2} are the two ability parameters for dimensions 1 and 2,

a_{i1} and a_{i2} are the item discrimination parameters for dimensions 1 and 2,

d_j is the scalar difficulty parameter.

Even though there is a discrimination parameter for each dimension, there is only one difficulty parameter. That is, difficulty parameters for each dimension are indeterminate. Reckase called the model M2PL because it was a two-dimensional version of the two-parameter (discrimination and difficulty) unidimensional item response theory (IRT) model.

In a two-dimensional latent ability space, the a_i vector designates the $\theta_1 - \theta_2$ combination that is being best measured (i.e., the composite for which the item can optimally discriminate). If $a_1 = a_2$, both dimensions are measured equally well. If $a_1 = 0$ and $a_2 > 0$, discrimination only occurs along the θ_1 .

Graphically, for an item, the M2PL model probability of correct response forms an item response surface as opposed to the unidimensional item characteristic curve. Four different perspectives of this surface for an item with $a_1 = 1.0$, $a_2 = 1.0$, and $d = 0.20$ is illustrated in Figure 1.

The reason the M2PL model is denoted as a compensatory model is due to the addition of terms in the logit. This feature makes it possible for an examinee with low ability on one dimension to compensate by having a higher probability on the remaining dimension. Figure 2 illustrates the equiprobability contours for the response surface of the item shown in Figure 1. For the compensatory model, the contours are equally spaced and parallel across the response surface. The contour lines become closer as the slope of the response surface becomes steeper or more discriminating. Notice that Examinee A (high θ_1 , low θ_2) has the same probability of correct response as Examinee B (high θ_2 , low θ_1). Note,

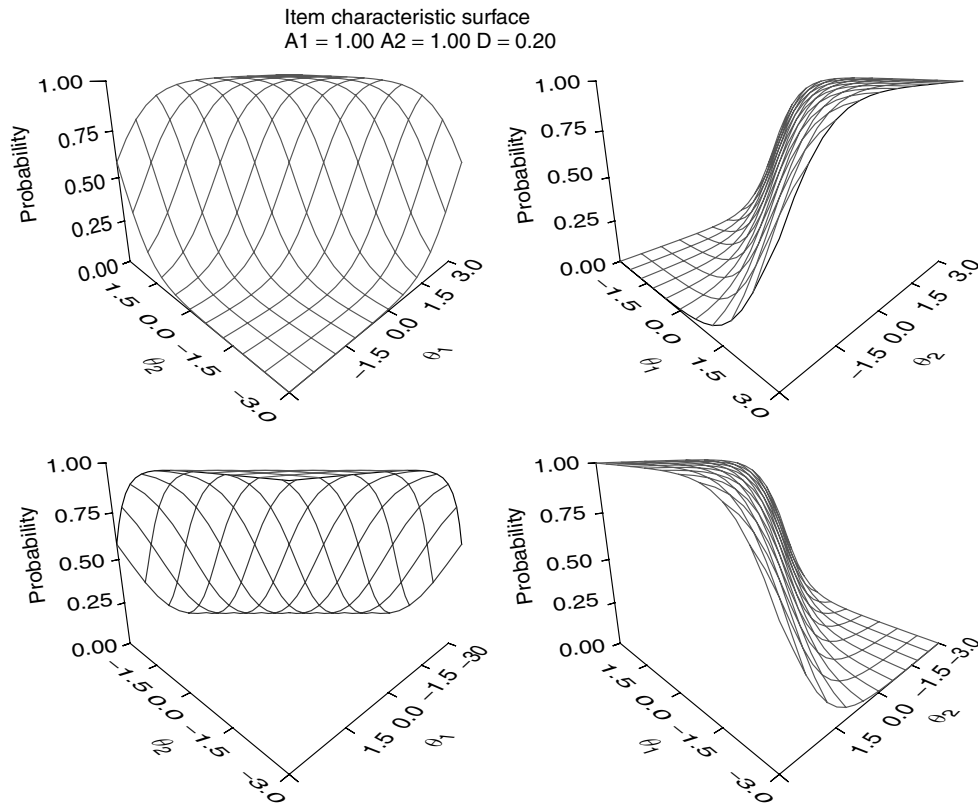


Figure 1 Four perspectives of a compensatory response surface

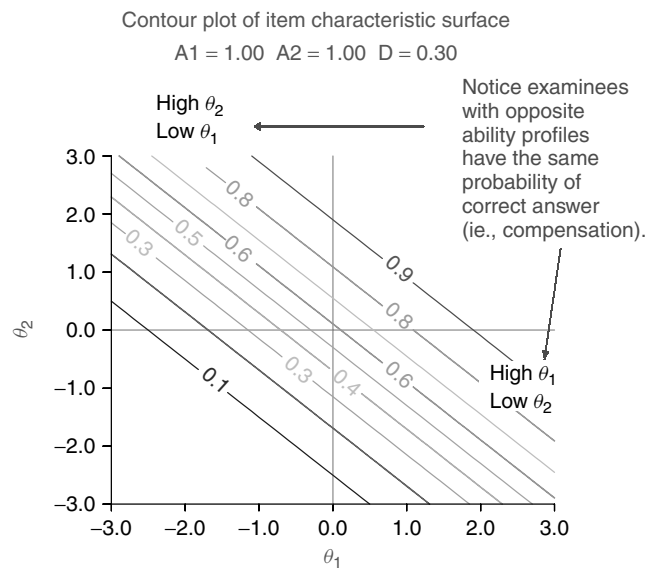


Figure 2 The contour plot of a compensatory item response surface with equal discrimination parameters

4 Multidimensional IRT Models

however, that the degree of compensation is greatest when $a_1 = a_2$. Obviously, the more a_1 and a_2 differ (e.g., $a_1 = 0$ and $a_2 > 0$, or $a_2 = 0$ and $a_1 > 0$), the less compensation can occur. That is, for integration of skills to be possible, an item must require the use of both skills.

Another multidimensional model that researchers have investigated is the noncompensatory model, proposed by Sympson [26]. This model (also known as the partial compensatory model) expresses the probability of correct response for subject j on item i as,

$$P(u_{ij} = 1 | \theta_{j1}, \theta_{j2}) = \left[\frac{1}{1 + e^{-a_{i1}(\theta_{j1} - b_{i1})}} \right] \times \left[\frac{1}{1 + e^{-a_{i2}(\theta_{j2} - b_{i2})}} \right] \quad (2)$$

where u_{ij} is the dichotomous score (0 = wrong, 1 = correct),

θ_{j1} and θ_{j2} are the two ability parameters for dimensions 1 and 2,

a_{i1} and a_{i2} are the item discrimination parameters for dimensions 1 and 2,

b_{i1} and b_{i2} are the item difficulty parameters for dimensions 1 and 2.

This model is essentially the product of two 2PL unidimensional IRT models, the overall probability of a correct response is bounded in the upper limit by the smaller of the two component probabilities. A graph of the item characteristic surface for an item with parameters $a_1 = 1.0$, $a_2 = 1.0$, $b_1 = 0.0$ and $b_2 = 0.0$ is shown in Figure 3.

Examining the contour plot for this item, Figure 4, enables one to see the noncompensatory nature of the model. Note that Examinee A (high θ_1 , low θ_2), and Examinee B (low θ_1 , high θ_2) have approximately the same probability as Examinee C (low θ_1 , low θ_2), indicating that little compensation actually occurs. It should be noted that the noncompensatory surface actually curves around, creating the curvilinear equiprobability contours shown in Figure 4.

Spray, Ackerman, and Carlson [23] extended the work of Reckase and Sympson by formulating a generalized MIRT model that combined both the compensatory and noncompensatory models. Letting

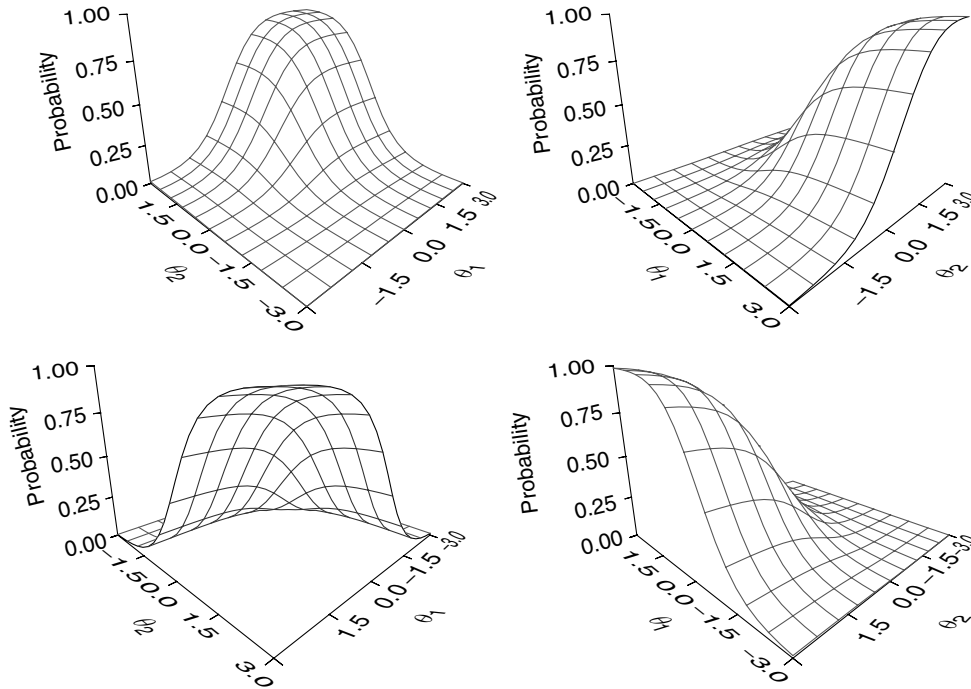


Figure 3 Four perspectives of a noncompensatory response surface

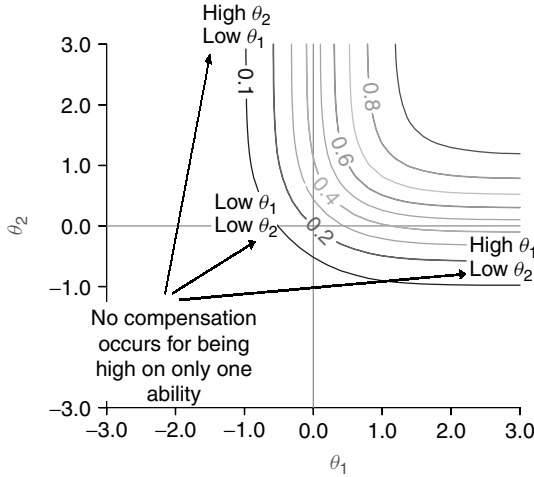


Figure 4 The contour plot of a noncompensatory response surface

$f_1 = a_1(\theta_1 - b_1)$ and $f_2 = a_2(\theta_2 - b_2)$, the compensatory model (1) can be written as

$$P_C = \frac{e^{(f_1+f_2)}}{1 + e^{(f_1+f_2)}} \quad (3)$$

and the noncompensatory model (4) can be written as

$$P_{NC} = \frac{e^{(f_1+f_2)}}{[1 + e^{(f_1+f_2)}] + e^{f_1} + e^{f_2}}. \quad (4)$$

The generalized model can then be expressed as

$$P_G = \frac{e^{(f_1+f_2)}}{[1 + e^{(f_1+f_2)}] + \mu[e^{f_1} + e^{f_2}]} \quad (5)$$

where μ represents a compensation or integration parameter that ranges from 0 to 1. If $\mu = 0$, then P_G is equivalent to the compensatory model (1), and if $\mu = 1$, then P_G is equivalent to the noncompensatory (2). As μ increases from 0, the degree of compensation decreases, and the equiprobability contours become more curvilinear. Contour plots for μ -parameter of 0.1 and 0.3 are shown in Figure 5.

For a more comprehensive development of different types of item response theory models, both unidimensional and multidimensional, readers are referred to van der Linden and Hambleton [28].

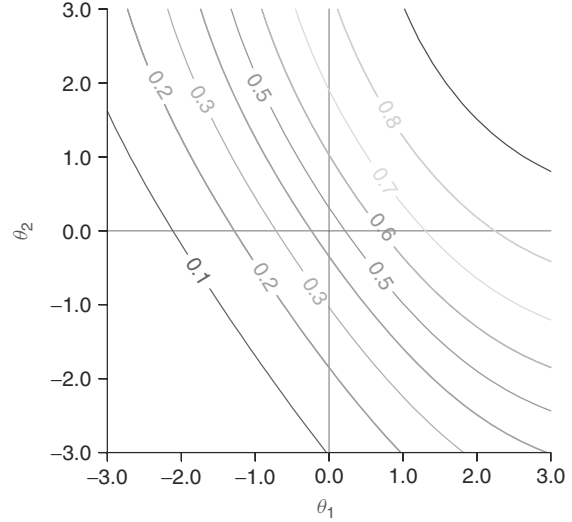


Figure 5 A contour of a generalized model response surface with $\mu = 0.15$

Estimating Item and Ability Parameters

Several software estimation programs were developed to estimate item parameters for the compensatory model, including NOHARM [8] and TESTFACT [30]. NOHARM uses a nonlinear factor analytic approach [14, 15], and can estimate item parameters for models having from one to six dimensions. It has both an exploratory or confirmatory mode (*see Factor Analysis: Confirmatory*). Unfortunately, NOHARM does not have the capability to estimate examinees' ability levels. TESTFACT uses a full-information approach to estimating both item and ability parameters for multidimensional compensatory models.

A relatively new approach using **Markov chain Monte Carlo** procedure was used by Bolt and Lall [6] to estimate parameters for both the compensatory and noncompensatory models. This estimation approach was used to estimate the degree of compensation in the generalized model by Ackerman and Turner [4], though with limited success. Another interesting approach using the genetic algorithm used in the program DETECT has been proposed by Zhang [31] to estimate parameters for the noncompensatory model. To date, there has been very little research on the goodness-of-fit of multidimensional models. One exception is the work of Ackerman,

Hombo, and Neustel [3], which looked at goodness-of-fit measures using the compensatory model.

Graphical Representations of Multidimensional Items and Information

To better understand and interpret MIRT item parameters, practitioners can use a variety of graphical techniques [2]. Item characteristic surface plots and contour plots do not allow the practitioner to examine and compare several items simultaneously. To get around this limitation, Reckase and McKinley [20] developed item vector plots. In this approach, each item is represented by a vector that represents three characteristics: discrimination, difficulty, and location. Using orthogonal axes, *discrimination* corresponds to the length of the item response vector. This length represents the maximum amount of discrimination, and is referred to as MDISC. For item i , MDISC is given by

$$MDISC = \sqrt{a_{i1}^2 + a_{i2}^2}, \quad (6)$$

where a_1 and a_2 are the logistic model discrimination parameters. The tail of the vector lies on the $p = 0.5$ equiprobability contour. If extended, all vectors would pass through the origin of the latent trait plane. Further, the a -parameters are constrained to be positive, as with unidimensional IRT, and, thus, the item vectors will only be located in the first and third quadrants. MDISC is analogous to the a -parameter in unidimensional IRT.

Difficulty corresponds to the location of the vector in space. The signed distance from the origin to the $p = 0.5$ equiprobability contour, denoted by D , is given by Reckase [18] as

$$D = \frac{-d_i}{MDISC}, \quad (7)$$

where d_i is the difficulty parameter for item i . The sign of this distance indicates the relative difficulty of the item. Items with negative D are relatively easy, and are in the third quadrant, whereas items with positive D are relatively hard, and are in first quadrant. D is analogous to the b -parameter in unidimensional IRT.

Location corresponds to the angular direction of each item relative to the positive θ_1 axis. The location

of item i is given by

$$\alpha_i = \arccos \frac{a_{i1}}{MDISC_i}. \quad (8)$$

A vector with location α_i greater than 45 degrees is a better measure of θ_2 than θ_1 , whereas a vector with a location α_i less than 45 degrees is a better measure of θ_1 .

By examining the discrimination, difficulty, and location of each item response vector, the degree of similarity in the two-dimensional composite for all items on the test can be viewed. By color-coding items according to content, practitioners can understand better the different ability composites that are being assessed by their tests. In Figure 6 is a vector plot for 25 items from a mathematics usage test. Note that the composite angle 43.1° represents the average composite direction of the vectors weighted by each item's MDISC value. This direction indicates two-dimensional composite that would be estimated if this test were calibrated using a unidimensional model.

Information. In item response theory, measurement precision is evaluated using information. The reciprocal of the information function is the asymptotic variance of the maximum likelihood estimate of ability. This relationship implies that the larger the

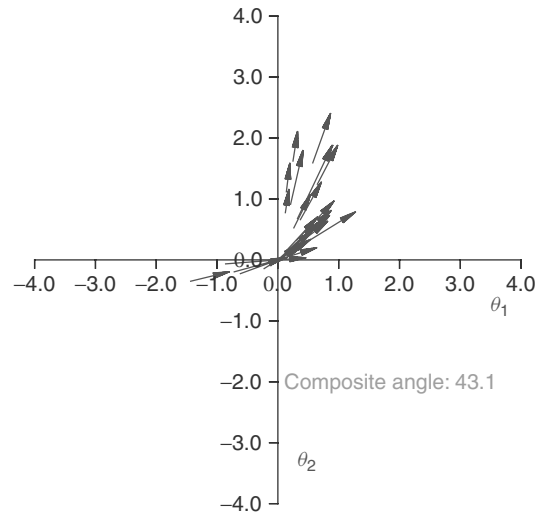


Figure 6 A vector plot of 25 ACT mathematics usage items

information function, the smaller the asymptotic variance and the more measurement precision. Multidimensional information (MINF) serves as one measure of precision. MINF is computed in a manner similar to its unidimensional IRT counterpart, except that the direction of the information is also considered, as shown in the formula

$$MINF = P_i(\theta)[1 - P_i(\theta)] \left(\sum_{k=1}^m \alpha_{ik} \cos \alpha_{ik} \right)^2. \quad (9)$$

MINF provides a measure of information at any point on the latent ability plane (i.e., measurement precision relative to the θ_1, θ_2 composite). MINF can be computed at the item level or at the test level (where the test information is the sum of the item information functions).

Reckase and McKinley [20] developed a *clamshell* plot to represent information with MINF (the representation was said to resemble clamshells, hence the term). To create the clamshells, the amount of information is computed at 49 uniformly spaced points on a 7×7 grid in the θ_1, θ_2 space. At each of the 49 points, the amount of information is computed for 10 different directions or ability composites from 0 to 90 degrees in 10-degree increments, and represented as the length of the 10 lines in each clamshell. Figure 7 contains the clamshell plot for the 25 items

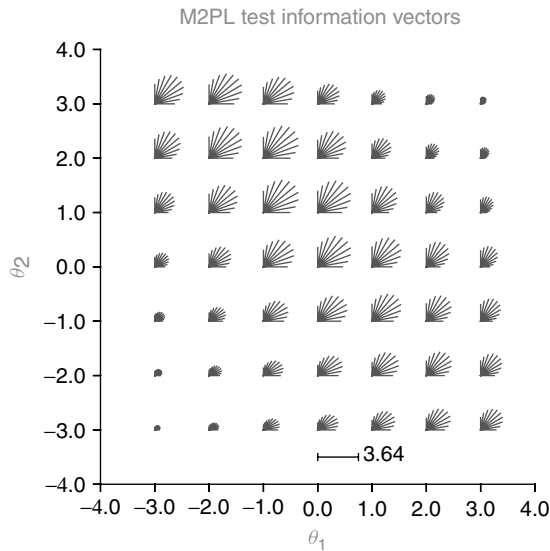


Figure 7 A clamshell information plot of 25 ACT mathematics usage items

whose vectors are displayed in Figure 6. At the ability (0,0), the clamshell vectors are almost of equal length, indicating that most of the composites tend to be measured with equal accuracy and much more accurately than would subjects located at the ability (3,3).

Ackerman [1] expanded upon the work of Reckase and McKinley and provided several other ways to graphically examine the information of two-dimensional tests. One example of his work was a number plot that was somewhat similar to the clamshell plot.

However, at each of the 49 points, the direction of maximum information is given as a numeric value on the grid, while the amount of information is represented by the size of the font for each numeric value (the larger the font, the greater the information). Figure 8 shows the number plot corresponding to the clamshell plot in Figure 7. Note that for examinees located at (0,0), the composite direction that is being best measured is at 40° . Such plots help to determine if two forms of a test are truly parallel.

Future Research Directions in MIRT

Multidimensional item response theory holds a great deal of promise for future psychometric research.

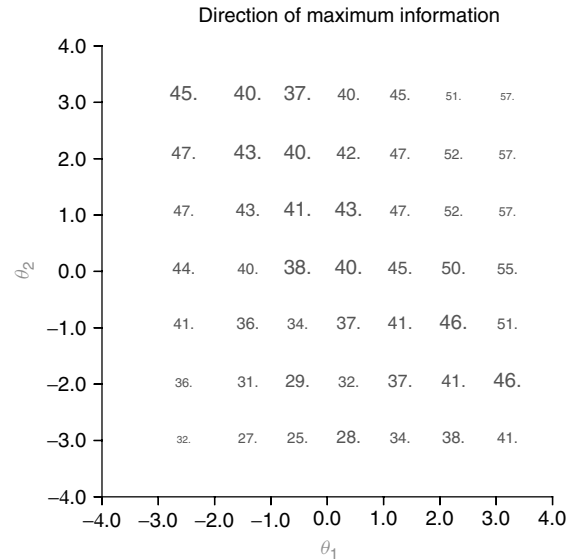


Figure 8 A number plot indicating the direction of maximum information at 49 ability locations

Directions in which MIRT research needs to be headed include computer adaptive testing [13], differential item functioning [22, 29], and diagnostic testing [7]. Other areas that need more development include expanding MIRT interpretations beyond two dimensions and to polytomous or Likert-type data [12, 17].

References

- [1] Ackerman, T.A. (1994). Creating a test information profile in a two-dimensional latent space, *Applied Psychological Measurement* **18**, 257–275.
- [2] Ackerman, T.A. (1996). Graphical representation of multidimensional item response theory analyses, *Applied Psychological Measurement* **20**, 311–330.
- [3] Ackerman, T.A., Hombo, C. & Neustel, S. (2002). Evaluating indices used to assess the goodness-of-fit of the compensatory multidimensional item response theory model, in *Paper Presented at the Meeting of the National Council of Measurement in Education*, New Orleans.
- [4] Ackerman, T.A. & Turner, R. (2003). Estimation and application of a generalized MIRT model: assessing the degree of compensation between two latent abilities, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, Chicago.
- [5] AERA, APA, NCME. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.
- [6] Bolt, D.M. & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo, *Applied Psychological Measurement* **27**(6), 395–414.
- [7] DiBello, L., Stout, W. & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques, in *Cognitive Diagnostic Assessment*, P. Nichols, S. Chipman and R. Brennan, eds, Erlbaum, Hillsdale, pp. 361–389.
- [8] Fraser, C. (1988). *NOHARM: An IBM PC Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory*, The University of New England, Armidale.
- [9] Gessaroli, M.E. & De Champlain, A.F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items, *Journal of Educational Measurement* **33**, 157–179.
- [10] Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items, *Applied Psychological Measurement* **10**, 287–302.
- [11] Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9**, 139–164.
- [12] Hattie, J., Krakowski, K., Rogers, H.J. & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality, *Applied Psychological Measurement* **20**, 1–14.
- [13] Luecht, R.M. (1996). Multidimensional computer adaptive testing in a certification or licensure context, *Applied Psychological Measurement* **20**, 389–404.
- [14] McDonald, R.P. (1967). Nonlinear factor analysis. Psychometric Monograph No. 15.
- [15] McDonald, R.P. (1999). *Test Theory: A Unified Approach*, Erlbaum, Hillsdale.
- [16] Messick, S. (1989). Validity, in *Educational Measurement*, 3rd Edition, R.L. Linn, eds, American Council on Education, Macmillan, New York, pp. 13–103.
- [17] Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality, *Journal of Educational Measurement* **28**, 99–117.
- [18] Nandakumar, R. & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality, *Journal of Educational Statistics* **18**, 41–68.
- [19] Reckase, M.D. (1985). The difficulty of test items that measure more than one ability, *Applied Psychological Measurement* **9**, 401–412.
- [20] Reckase, M.D. & McKinley, R.L. (1991). The discrimination power of items that measure more than one dimension, *Applied Psychological Measurement* **14**, 361–373.
- [21] Roussos, L. (1992). *Hierarchical Agglomerative Clustering Computer Programs Manual*, University of Illinois at Urbana-Champaign, Department of Statistics, (Unpublished manuscript).
- [22] Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm, *Applied Psychological Measurement* **20**, 355–371.
- [23] Spray, J.A., Ackerman, T.A. & Carlson, J. (1986). An analysis of multidimensional item response theory models, in *Paper Presented at the Meeting of the Office of Naval Research Contractors*, Gatlinburg.
- [24] Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika* **52**, 589–617.
- [25] Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L. & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment, *Applied Psychological Measurement* **20**, 331–354.
- [26] Sympton, J.B. (1978). A model for testing multidimensional items, in *Proceedings of the 1977 Computerized Adaptive Testing Conference*, D.J. Weiss, eds University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, pp. 82–98.
- [27] Traub, R.E. (1983). A priori consideration in choosing an item response theory model, in *Applications of Item Response Theory*, R.K. Hambleton, eds Educational Research Institute of British Columbia, Vancouver, pp. 57–70.
- [28] van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*, Springer, New York.
- [29] Walker, C.M. & Beretvas, S.N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF, *Journal of Educational Measurement* **38**, 147–163.

-
- [30] Wilson, D., Wood, R. & Gibbons, R. (1987). *TESTFACT*, [Computer Program]. Scientific Software, Mooresville.
- [31] Zhang, J. (2001). Using multidimensional IRT to analyze item response data, in *Paper Presented at the Meeting of the National Council on Measurement in Education*, Seattle.
- [32] Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure, *Psychometrika* **64**, 231–249.
- Kelderman, H. & Rijkes, C. (1994). Loglinear multidimensional IRT models for polytomously scored items, *Psychometrika* **2**(59), 149–176.
- Nandakumar, R., Yu, F., Li, H. & Stout, W. (1998). Assessing unidimensionality of polytomous data, *Applied Psychological Measurement* **22**, 99–115.

TERRY A. ACKERMAN

Further Reading

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring, *Applied Measurement in Education* **7**, 255–278.

Multidimensional Scaling

PATRICK J.F. GROENEN AND MICHEL VAN DE VELDEN

Volume 3, pp. 1280–1289

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multidimensional Scaling

Introduction

Multidimensional scaling is a statistical technique originating in psychometrics. The data used for multidimensional scaling (MDS) are dissimilarities between pairs of objects (*see Proximity Measures*). The main objective of MDS is to represent these dissimilarities as distances between points in a low dimensional space such that the distances correspond as closely as possible to the dissimilarities.

Let us introduce the method by means of a small example. Ekman [7] collected data to study the perception of 14 different colors. Every pair of colors was judged by a respondent from having ‘no similarity’ to being ‘identical’. The obtained scores can be scaled in such a way that identical colors are denoted by 0, and completely different colors by 1. The averages of these dissimilarity scores over the 31 respondents are presented in Table 1. Starting from wavelength 434, the colors range from bluish-purple, blue, green, yellow, to red. Note that the dissimilarities are symmetric: the extent to which colors with wavelengths 490 and 584 are the same is equal to that of colors 584 and 490. Therefore, it suffices to only present the lower triangular part of the data in Table 1. Also, the diagonal is not of interest in MDS because the distance of an object with itself is necessarily zero.

MDS tries to represent the dissimilarities in Table 1 in a map. Figure 1 presents such an MDS

map in 2 dimensions. We see that the colors, denoted by their wavelengths, are represented in the shape of a circle. The interpretation of this map should be done in terms of the depicted interpoint distances. Note that, as distances do not change under rotation, a rotation of the plot does not affect the interpretation. Similarly, a translation of the solution (that is, a shift of all coordinates by a fixed value per dimension) does not change the distances either, nor does a reflection of one or both of the axes.

Figure 1 should be interpreted as follows. Colors that are located close to each other are perceived as being similar. For example, the colors with wavelengths 434 (violet) and 445 (indigo) or 628 and 651 (both red). In contrast, colors that are positioned far away from each other, such as 490 (green) and 610 (orange) indicate a large difference in perception. The circular form obtained in this example is in accordance with theory on the perception of colors.

Summarizing, MDS is a technique that translates a table of dissimilarities between pairs of objects into a map where distances between the points match the dissimilarities as well as possible. The use of MDS is not limited to psychology but has applications in a wide area of disciplines, such as sociology, economics, biology, chemistry, and archaeology. Often, it is used as a technique for exploring the data. In addition, it can be used as a technique for dimension reduction. Sometimes, as in chemistry, the objective is to reconstruct a 3D model for large DNA-molecules for which only partial information of the distances between atoms is available.

Table 1 Dissimilarities of colors with wavelengths from 434 to 674 nm [7]

nm	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434	–													
445	0.14	–												
465	0.58	0.50	–											
472	0.58	0.56	0.19	–										
490	0.82	0.78	0.53	0.46	–									
504	0.94	0.91	0.83	0.75	0.39	–								
537	0.93	0.93	0.90	0.90	0.69	0.38	–							
555	0.96	0.93	0.92	0.91	0.74	0.55	0.27	–						
584	0.98	0.98	0.98	0.98	0.93	0.86	0.78	0.67	–					
600	0.93	0.96	0.99	0.99	0.98	0.92	0.86	0.81	0.42	–				
610	0.91	0.93	0.98	1.00	0.98	0.98	0.95	0.96	0.63	0.26	–			
628	0.88	0.89	0.99	0.99	0.99	0.98	0.98	0.97	0.73	0.50	0.24	–		
651	0.87	0.87	0.95	0.98	0.98	0.98	0.98	0.98	0.80	0.59	0.38	0.15	–	
674	0.84	0.86	0.97	0.96	1.00	0.99	1.00	0.98	0.77	0.72	0.45	0.32	0.24	–

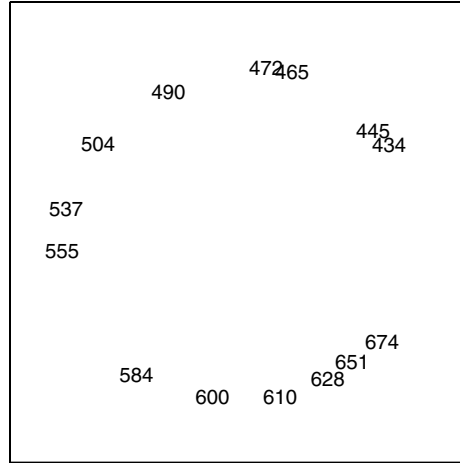


Figure 1 MDS solution in 2 dimensions of the color data in Table 1

Data for MDS

In the previous section, we introduced MDS as a method to describe relationships between objects on the basis of observed dissimilarities. However, instead of dissimilarities we often observe similarities between objects. Correlations, for example, can be interpreted as similarities. By converting the similarities into dissimilarities MDS can easily be applied to similarity data. There are several ways of transforming similarities into dissimilarities. For example, we may take one divided by the similarity or we can apply any monotone decreasing function that yields nonnegative values (dissimilarities cannot be negative). However, in Section ‘Transformations of the Data’, we shall see that by applying transformations in MDS, there is no need to transform similarities into dissimilarities. To indicate both similarity and dissimilarity data, we use the generic term proximities.

Data in MDS can be obtained in a variety of ways. We distinguish between the direct collection of proximities versus derived proximities. The color data of the previous section is an example of direct proximities. That is, the data arrives in the format of proximities. Often, this is not the case and our data does not consist of proximities between variables. However, by considering an appropriate measure, proximities can be derived from the original data. For example, consider the case where objects are rated on several variables. If the interest lies in representing the variables, we can calculate the correlation matrix

as measure of similarity between the variables. MDS can be applied to describe the relationship between the variables on the basis of the derived proximities. Alternatively, if interest lies in the objects, Euclidean distances can be computed between the objects using the variables as dimensions. In this case, we use high dimensional Euclidean distances as dissimilarities and we can use MDS to reconstruct these distances in a low dimensional space.

Co-occurrence data are another source for obtaining dissimilarities. For such data, a respondent groups the objects into partitions and an $n \times n$ incidence matrix is derived where a one indicates that a pair of objects is in the same group and a zero indicates that they are in different groups. By considering the frequencies of objects being in the same or different groups and by applying special measures (such as the so-called *Jaccard similarity measure*), we obtain proximities. For a detailed discussion of various (dis)similarity measures, we refer to [8] (see **Proximity Measures**).

Formalizing Multidimensional Scaling

To formalize MDS, we need some notation. Let n be the number of different objects and let the dissimilarity for objects i and j be given by δ_{ij} . The coordinates are gathered in an $n \times p$ matrix $\mathbf{X}$, where p is the dimensionality of the solution to be specified in advance by the user. Thus, row i from $\mathbf{X}$ gives the

coordinates for object i . Let $d_{ij}(\mathbf{X})$ be the Euclidean distance between rows i and j of $\mathbf{X}$ defined as

$$d_{ij}(\mathbf{X}) = \left(\sum_{s=1}^p (x_{is} - x_{js})^2 \right)^{1/2}, \quad (1)$$

that is, the length of the shortest line connecting points i and j . The objective of MDS is to find a matrix $\mathbf{X}$ such that $d_{ij}(\mathbf{X})$ matches δ_{ij} as closely as possible. This objective can be formulated in a variety of ways but here we use the definition of raw Stress $\sigma^2(\mathbf{X})$, that is,

$$\sigma^2(\mathbf{X}) = \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} (\delta_{ij} - d_{ij}(\mathbf{X}))^2 \quad (2)$$

by Kruskal [11, 12] who was the first one to propose a formal measure for doing MDS. This measure is also referred to as the least-squares MDS model. Note that due to the symmetry of the dissimilarities and the distances, the summation only involves the pairs ij where $i > j$. Here, w_{ij} is a user defined weight that must be nonnegative. For example, many MDS programs implicitly choose $w_{ij} = 0$ for dissimilarities that are missing.

The minimization of $\sigma^2(\mathbf{X})$ is a rather complex problem that cannot be solved in closed-form. Therefore, MDS programs use iterative numerical algorithms to find a matrix $\mathbf{X}$ for which $\sigma^2(\mathbf{X})$ is a minimum. One of the best algorithms available is the SMACOF algorithm [1, 3, 4, 5] based on iterative majorization. The SMACOF algorithm has been implemented in the SPSS procedure Proxscal [13]. In Section ‘The SMACOF Algorithm’, we give a brief illustration of the SMACOF algorithm.

Because Euclidean distances do not change under rotation, translation, and reflection, these operations may be freely applied to MDS solution without affecting the raw Stress. Many MDS programs use this indeterminacy to center the coordinates so that they sum to zero dimension wise. The freedom of rotation is often exploited to put the solution in so-called principal axis orientation. That is, the axes are rotated in such a way that the variance of $\mathbf{X}$ is maximal along the first dimension, the second dimension is uncorrelated to the first and has again maximal variance, and so on.

Here, we have discussed the Stress measure for MDS. However, there are several other measures for doing MDS. In Section ‘Alternative Measures for

Doing Multidimensional Scaling’, we briefly discuss other popular definitions of Stress.

Transformations of the Data

So far, we have assumed that the dissimilarities are known. However, this is often not the case. Consider for example the situation in which the objects have been ranked. That is, the dissimilarities between the objects are not known, but their order is known. In such a case, we would like to assign numerical values to the proximities in such a way that these values exhibit the same rank order as the data. These numerical values are usually called disparities, d-hats, or pseudo distances, and they are denoted by $\hat{d}$. The task of MDS now becomes to simultaneously obtain disparities and coordinates in such a way that the coordinates represent the disparities (and thus the original rank order of the data) as well as possible. This objective can be captured in minimizing a slight adaptation of raw Stress, that is,

$$\sigma^2(\hat{\mathbf{d}}, \mathbf{X}) = \sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} (\hat{d}_{ij} - d_{ij}(\mathbf{X}))^2, \quad (3)$$

over both the $\hat{\mathbf{d}}$ and $\mathbf{X}$, where $\hat{\mathbf{d}}$ is the vector containing $\hat{d}_{ij}$ for all pairs. The process of finding the disparities is called **optimal scaling** and was first introduced by Kruskal [11, 12]. Optimal scaling aims to find a transformation of the data that fits as well as possible the distances in the MDS solution. To avoid the trivial optimal scaling solution $\mathbf{X} = \mathbf{0}$ and $\hat{d}_{ij} = 0$ for all ij , we impose a length constraint on the disparities in such a way that the sum of squared d-hats equals a fixed constant. For example, $\sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} \hat{d}_{ij}^2 = n(n-1)/2$ [3].

Transformations of the data are often used in MDS. Figure 2 shows a few examples of transformation plots for the color example. Let us look at some special cases.

Suppose that we choose $\hat{d}_{ij} = \delta_{ij}$ for all ij . Then, minimizing (3) without the length constraint is exactly the same as minimizing (2). Minimizing (3) with the length constraint only changes $\hat{d}_{ij} = a\delta_{ij}$, where a is a scalar chosen in such a way that the length constraint is satisfied. This transformation is called a *ratio transformation* (Figure 2a). Note that, in this case, the relative differences of $a\delta_{ij}$ are the same as those for δ_{ij} . Hence, the relative differences

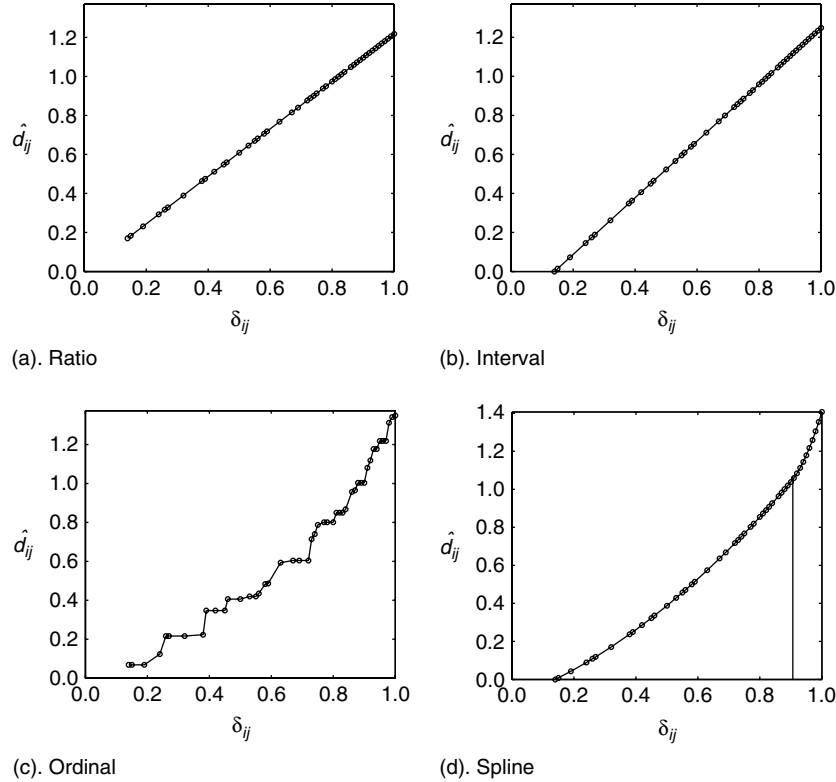


Figure 2 Four transformations often used in MDS

of the $d_{ij}(\mathbf{X})$ in (2) and (3) are also the same. Ratio MDS can be seen as the most restrictive transformation in MDS.

An obvious extension to the ratio transformation is obtained by allowing the $\hat{d}_{ij}$ to be a linear transformation of the δ_{ij} . That is, $\hat{d}_{ij} = a + b\delta_{ij}$, for some unknown values of a and b . Figure 2b depicts an interval transformation. This transformation may be chosen if there is reason to believe that $\delta_{ij} = 0$ does not have any particular interpretation. An interval transformation that is almost horizontal reveals little about the data as different dissimilarities are transformed to similar disparities. In such a case, the constant term will dominate the $\hat{d}_{ij}$'s. On the other hand, a good interval transformation is obtained if the line is not horizontal and the constant term is reasonably small with respect to the rest.

For ordinal MDS, the $\hat{d}_{ij}$ are only required to have the same rank order as δ_{ij} . That is, if for two pairs of objects ij and kl we have $\delta_{ij} \leq \delta_{kl}$ then the corresponding disparities must satisfy $\hat{d}_{ij} \leq \hat{d}_{kl}$.

An example of an ordinal transformation in MDS is given in Figure 2c. Typically, an ordinal transformation shows a step function. Similar to the case for interval transformations, it is not a good sign if the transformation plot shows a horizontal line. Moreover, if the transformation plot only exhibits a few steps, ordinal MDS does not use finer information available in the data. Ordinal MDS is particularly suited if the original data are rank orders. To compute an ordinal transformation a method called **monotone regression** can be used.

A monotone spline transformation offers more freedom than an interval transformation, but never more than an ordinal transformation (see **Scatterplot Smoothers**). The advantage of a spline transformation over an ordinal transformation is that it will yield a smooth transformation. Figure 2d shows an example of a spline transformation. A spline transformation is built on two ideas. First, the range of the δ_{ij} 's can be subdivided into connected intervals. Then, for each interval, the data are transformed

using a polynomial of a specified degree. For example, a second-degree polynomial imposes that $\hat{d}_{ij} = a\delta_{ij}^2 + b\delta_{ij} + c$. The special feature of a spline is that at the connections of the intervals, the so-called interior knots, the two polynomials connect smoothly. The spline transformation in Figure 2d was obtained by choosing one interior knot at .90 and by using second-degree polynomials. For MDS it is important that the transformation is monotone increasing. This requirement is automatically satisfied for monotone splines or I-Splines (see, [14, 1]). For choosing a transformation in MDS it suffices to know that a spline transformation is smooth and nonlinear. The amount of nonlinearity is governed by the number of interior knots specified. Unless the number of dissimilarities is very large, a few interior knots for a second-degree spline usually works well.

There are several reasons to use transformations in MDS. One reason concerns the fit of the data in low dimensionality. By choosing a transformation that is less restrictive than the ratio transformation a better fit may be obtained. Alternatively, there may exist theoretical reasons as to why a transformation of the dissimilarities is desired. Ordered from most to least restrictive transformation, we start with ratio, then interval, spline, and ordinal.

If the data are dissimilarities, then it is necessary that a transformation is monotone increasing (as in Figure 2) so that pairs with higher dissimilarities are indeed modeled by larger distances. Conversely, if the data are similarities, then the transformation should be monotone decreasing so that more similar pairs are

modeled by smaller distances. A ratio transformation is not possible for similarities. The reason is that the $\hat{d}_{ij}$'s must be nonnegative. This implies that the transformation must include an intercept.

In the MDS literature, one often encounters the terms metric and nonmetric MDS. Metric MDS refers to the ratio and interval transformations, whereas all other transformations such as ordinal and spline transformations are covered by the term nonmetric MDS. We believe, however, that it is better to refer directly to the type of transformation that is used.

There exist other a-priori transformations of the data that are not optimal in the sense described above. That is transformations that are not obtained by minimizing (3). The advantage of optimal transformations is that the exact form of the transformation is unknown and determined optimally together with the MDS configuration.

Diagnostics

In order to assess the quality of the MDS solution we can study the differences between the MDS solution and the data. One convenient way to do this is by inspecting the so-called **Shepard diagram**.

A Shepard diagram shows both the transformation and the error. Let p_{ij} denotes the proximity between objects i and j . Then, a Shepard diagram plots simultaneously the pairs $(p_{ij}, d_{ij}(\mathbf{X}))$ and $(p_{ij}, \hat{d}_{ij})$. In Figure 3, solid points denote the pairs $(p_{ij}, d_{ij}(\mathbf{X}))$ and open circles represent the pairs $(p_{ij}, \hat{d}_{ij})$. By

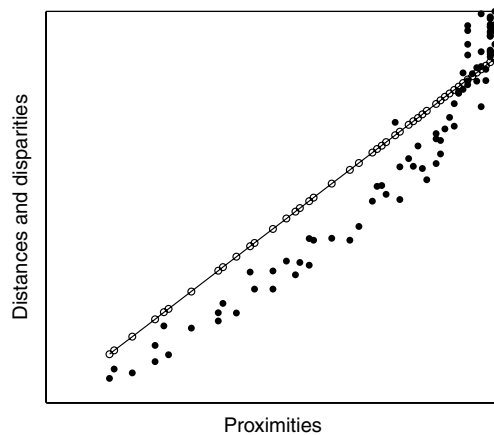


Figure 3 Shepard diagram for ordinal MDS of the color data, where the proximities are dissimilarities

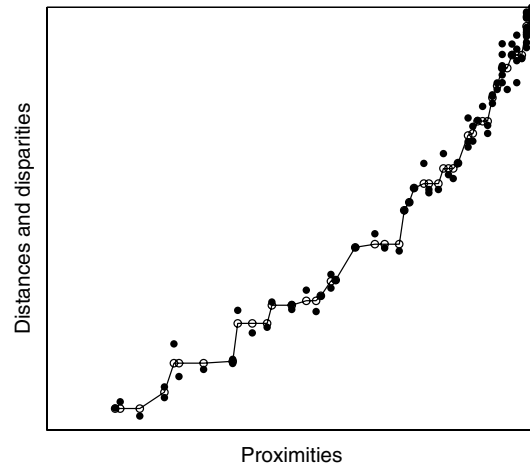


Figure 4 Shepard diagram for ordinal MDS of the color data, where the proximities are dissimilarities

connecting the open circles a line is obtained representing the relationship between the proximities and the disparities which is equivalent to the transformation plots in Figure 2. The vertical distances between the open and closed circles are equal to $\hat{d}_{ij} - d_{ij}(\mathbf{X})$, that is, they give the errors of representation for each pair of objects. Hence, the Shepard diagram can be used to inspect both the residuals of the MDS solution and the transformation. Outliers can be detected as well as possible systematic deviations. Figure 3 gives the Shepard diagram for the ratio MDS solution of Figure 1 using the color data. We see that all the errors corresponding to low proximities are positive whereas the errors for the higher proximities are all negative. This kind of heteroscedasticity suggests the use of a more liberal transformation. Figure 4 gives the Shepard diagram for an ordinal transformation. As the solid points are closer to the line connecting the open circles, we may indeed conclude that the heteroscedasticity has gone and that the fit has become better.

Choosing the Dimensionality

Several methods have been proposed to choose the dimensionality of the MDS solution. However, no definite strategy is present. **Unidimensional scaling**, that is, $p = 1$ (with ratio transformation) has to be treated with special care because the usual MDS algorithms will end up in local minima that can be far from global minima.

One approach to determine the dimensionality is to compute MDS solutions for a range of dimensions, say from 2 to 6 dimensions, and plot the Stress against the dimension. Similar to common practice in principal component analysis, we then use the elbow criterion to determine the dimensionality. That is, we choose the number of dimensions where a bend in the curve occurs.

Another approach (for ordinal MDS), proposed by [15], compares the Stress values against the Stress of generated data. However, perhaps the most important criterion for choosing the dimensionality is simply based on the interpretability of the map. Therefore, the vast majority of reported MDS solutions are done in two dimensions and occasionally in three dimensions. The interpretability criterion is a valid one especially when MDS is used for exploration of the data.

The SMACOF Algorithm

In Section ‘Formalizing Multidimensional Scaling’, we mentioned that a popular algorithm for minimizing Stress is the SMACOF Algorithm. Its major feature is that it guarantees lower Stress values in each iteration. Here we briefly sketch how this algorithm works.

Nowadays, the acronym SMACOF stands for Scaling by Majorizing a Complex Function. To understand how it works, consider Figure 5a. Suppose that we have dissimilarity data on 13 stock

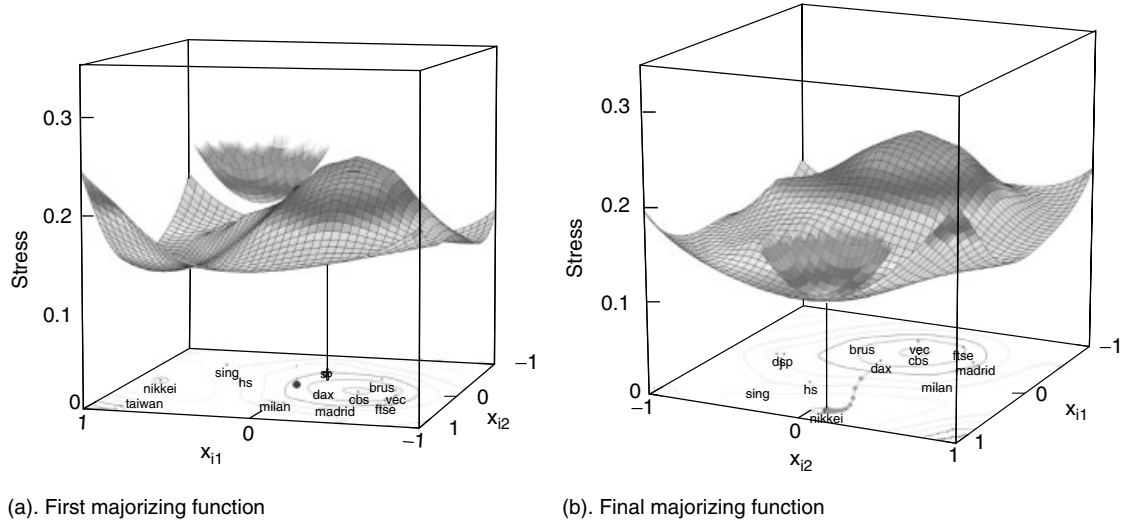


Figure 5 The Stress function and the majorizing function for the supporting (0,0) in Panel a., and the majorizing function at convergence in Panel b. Reproduced by permission of Vrieseborch Publishers

markets and that the two dimensional MDS solution is given by the points in the horizontal plane of Figure 5a and 5b. Now, suppose that the position of the 'nikkei' index was unknown. Then we can calculate the value of Stress as a function of the two coordinates for 'nikkei'. The surface in Figure 5a shows these values of Stress for every potential position of 'nikkei'. To minimize Stress we must find the coordinates that yield the lowest Stress. Hence, the final point must be located in the horizontal plane under the lowest value of the surface. To find this point, we use an iterative procedure that is based on majorization. First, as an initial point for 'nikkei' we choose the origin in Figure 5a. Then, a so-called majorization function is chosen in such a way that, for this initial point, its value is equal to the Stress value, and elsewhere it lies above the Stress surface. Here, the majorizing function is chosen to be quadratic and is visualized in Figure 5a as the bowl-shaped surface above the Stress function surface. Now, as the Stress surface is always below the majorization function, the value of Stress evaluated at the point corresponding to the minimum of the majorization function, will be lower than the initial Stress value. Hence, the initial point can be updated by calculating the minimum of the majorization function which is easy because the majorizing function is quadratic. Using the updated point we repeat this process until the coordinates remain practically

constant. Figure 5b shows the subsequent coordinates for 'nikkei' obtained by majorization as a line with connected dots marking the path to its final position.

Alternative Measures for Doing Multidimensional Scaling

In addition to the raw Stress measure introduced in Section 'Formalizing Multidimensional Scaling', there exist other measures for doing Stress. Here we give a short overview of some of the most popular alternatives. First, we discuss normalized raw Stress,

$$\sigma_n^2(\hat{\mathbf{d}}, \mathbf{X}) = \frac{\sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} (\hat{d}_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} \hat{d}_{ij}^2}, \quad (4)$$

which is simply raw Stress divided by the sum of squared dissimilarities. The advantage of this measure over raw Stress is that its value is independent of the scale of the dissimilarities and their number. Thus, multiplying the dissimilarities by a positive factor will not change (4) at a local minimum, whereas the coordinates will be the same up to the same factor.

The second measure is Kruskal's Stress-1 formula

$$\sigma_1(\hat{\mathbf{d}}, \mathbf{X}) = \left(\frac{\sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} (\hat{d}_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i=2}^n \sum_{j=1}^{i-1} w_{ij} d_{ij}^2(\mathbf{X})} \right)^{1/2}, \quad (5)$$

which is equal to the square root of raw Stress divided by the sum of squared distances. This measure is of importance because many MDS programs and publications report this value. It can be proved that at a local minimum of $\sigma_n^2(\hat{\mathbf{d}}, \mathbf{X})$, $\sigma_1(\hat{\mathbf{d}}, \mathbf{X})$ also has a local minimum with the same configuration up to a multiplicative constant. In addition, the square root of normalized raw Stress is equal to Stress-1 [1].

A third measure is Kruskal's Stress-2, which is similar to Stress-1 except that the denominator is based on the variance of the distances instead of the sum of squares. Stress-2 can be used to avoid the situation where all distances are almost equal.

A final measure that seems reasonably popular is called *S-Stress* (implemented in the program ALSCAL) and it measures the sum of squared error between *squared* distances and *squared* dissimilarities [16]. The disadvantage of this measure is that it tends to give solutions in which large dissimilarities are overemphasized and the small dissimilarities are not well represented.

Pitfalls

If missing dissimilarities are present, a special problem may occur for certain patterns of missing dissimilarities. For example, if it is possible to split the objects in two or more sets such that the between-set

weights w_{ij} are all zero, we are dealing with independent MDS problems, one for each set. If this situation is not recognized, you may inadvertently interpret the missing between set distances. With only a few missing values, this situation is unlikely to happen. However, when dealing with many missing values, one should verify that the problem does not occur.

Another important issue is to understand what MDS will do if there is no information in the data, that is, when all dissimilarities are equal. Such a case can be seen as maximally uninformative and therefore as a null model. Solutions of empirical data should deviate from this null model. This situation was studied in great detail by [2]. It turned out that for constant dissimilarities, MDS will find in one dimension points equally spread on a line (see Figure 6). In two dimensions, the points lie on concentric circles [6] and in three dimensions (or higher), the points lie equally spaced on the surface of a sphere. Because all dissimilarities are equal, any permutation of these points yield an equally good fit.

This type of degeneracy can be easily recognized by checking the Shepard diagram. For example, if all disparities (or dissimilarities in ratio MDS) fall into a small interval considerably different from zero, we are dealing with the case of (almost) constant dissimilarities. For such a case, we advise redoing the MDS analysis with a more restrictive transformation, for example, using monotone splines, an interval transformation or even ratio MDS.

A final pitfall for MDS are local minima. A local minimum for Stress implies that small changes in the configuration always have a worse Stress than the local minimum solution. However, larger changes in the configuration may yield a lower Stress. A configuration with the overall lowest Stress value is called a *global minimum*. In general, MDS algorithms that minimize Stress cannot guarantee the retrieval of

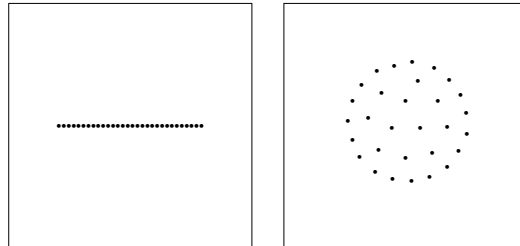


Figure 6 Solutions for constant dissimilarities with $n = 30$. The left plot shows the unidimensional solution and the right plot a 2D solution

a global minimum. However, if the dimensionality is exactly $n - 1$, it is known that ratio MDS only has one minimum that is consequently global. Moreover, when $p = n - 1$ is specified, MDS often yields a solution that fits in a dimensionality lower than $n - 1$. If so, then this MDS solution is also a global minimum. A different case is that of unidimensional scaling. For unidimensional scaling with a ratio transformation, it is well known that it has many local minima and can better be solved using combinatorial methods. For low dimensionality, like $p = 2$ or $p = 3$, experiments indicated that the number of different local minima ranges from a few to several thousands. For an overview of issues concerning local minima in ratio MDS, we refer to [9] and [10].

When transformations are used, there are fewer local minima and the probability of finding a global minimum increases. As a general strategy, we advice to use multiple random starts (say 100 random starts) and retain the solution with the lowest Stress. If most random starts end in the same candidate minimum, then there probably only exist few local minima. However, if the random starts end in many different local minima, the data exhibit a serious local minimum problem. In that case, it is advisable to increase the number of random starts and retain the best solution.

References

- [1] Borg, I. & Groenen, P.J.F. (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York.
- [2] Buja, A., Logan, B.F., Reeds, J.R. & Shepp, L.A. (1994). Inequalities and positive-definite functions arising from a problem in multidimensional scaling, *The Annals of Statistics* **22**, 406–438.
- [3] De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling, in *Recent Developments in Statistics*, J.R., Barra, F., Brodeau, G., Romier & B., van Cutsem, eds, North-Holland, Amsterdam, pp. 133–145.
- [4] De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling, *Journal of Classification* **5**, 163–180.
- [5] De Leeuw, J. & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration, in *Multivariate Analysis*, Vol. V, P.R., Krishnaiah, ed., North-Holland, Amsterdam, pp. 501–522.
- [6] De Leeuw, J. & Stoop, I. (1984). Upper bounds of Kruskal's Stress, *Psychometrika* **49**, 391–402.
- [7] Ekman, G. (1954). Dimensions of color vision, *Journal of Psychology* **38**, 467–474.
- [8] Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification* **3**, 5–48.
- [9] Groenen, P.J.F. & Heiser, W.J. (1996). The tunneling method for global optimization in multidimensional scaling, *Psychometrika* **61**, 529–550.
- [10] Groenen, P.J.F., Heiser, W.J. & Meulman, J.J. (1999). Global optimization in least-squares multidimensional scaling by distance smoothing, *Journal of Classification* **16**, 225–254.
- [11] Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.
- [12] Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: a numerical method, *Psychometrika* **29**, 115–129.
- [13] Meulman, J.J. Heiser, W.J. & SPSS. (1999). *SPSS Categories 10.0*, SPSS, Chicago.
- [14] Ramsay, J.O. (1988). Monotone regression splines in action, *Statistical Science* **3**(4), 425–461.
- [15] Spence, I. & Ogilvie, J.C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling, *Multivariate Behavioral Research* **8**, 511–517.
- [16] Takane, Y., Young, F.W. & De Leeuw, J. (1977). Non-metric individual differences multidimensional scaling: an alternating least-squares method with optimal scaling features, *Psychometrika* **42**, 7–67.

(See also **Multidimensional Unfolding; Scaling Asymmetric Matrices; Scaling of Preferential Choice**)

PATRICK J.F. GROENEN AND
MICHEL VAN DE VELDEN

Multidimensional Unfolding

JAN DE LEEUW

Volume 3, pp. 1289–1294

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multidimensional Unfolding

The unfolding model is a geometric model for preference and choice. It locates individuals and alternatives as points in a joint space, and it says that an individual will pick the alternative in the choice set closest to its ideal point. Unfolding originated in the work of Coombs [4] and his students. It is perhaps the dominant model in both **scaling of preference data** and **attitude scaling**.

The multidimensional unfolding technique computes solutions to the equations of unfolding model. It can be defined as **multidimensional scaling** of off-diagonal matrices. This means the data are dissimilarities between n row objects and m column objects, collected in an $n \times m$ matrix Δ . An important example is preference data, where δ_{ij} indicates, for instance, how much individual i dislikes object j . In unfolding, we have many of the same distinctions as in general multidimensional scaling: there is uni-dimensional and multidimensional unfolding, metric and nonmetric unfolding, and there are many possible choices of loss functions that can be minimized.

First we will look at (metric) unfolding as defining the system of equations $\delta_{ij} = d_{ij}(X, Y)$, where X is the $n \times p$ configuration matrix of row points, Y is the $m \times p$ configuration matrix of column points, and

$$d_{ij}(X, Y) = \sqrt{\sum_{s=1}^p (x_{is} - y_{js})^2}. \quad (1)$$

Clearly, an equivalent system of algebraic equations is $\delta_{ij}^2 = d_{ij}^2(X, Y)$, and this system expands to

$$\delta_{ij}^2 = \sum_{s=1}^p x_{is}^2 + \sum_{s=1}^p y_{js}^2 - 2 \sum_{s=1}^p x_{is} y_{js}. \quad (2)$$

We can rewrite this in matrix form as $\Delta^{(2)} = ae'_m + e_n b' - 2XY'$, where a and b contain the row and column sums of squares, and where e is used for a vector with all elements equal to one. If we define the centering operators $J_n = I_n - e_n e'_n / n$ and $J_m = I_m - e_m e'_m / m$, then we see that doubly centering the matrix of squared dissimilarities gives the basic result

$$H = -\frac{1}{2} J_n \Delta^{(2)} J_m = \tilde{X} \tilde{Y}', \quad (3)$$

where $\tilde{X} = J_n X$ and $\tilde{Y} = J_m Y$ are centered versions of X and Y . For our system of equations to be solvable, it is necessary that $\text{rank}(H) \leq p$. Solving the system, or finding an approximate solution by using the singular value decomposition, gives us already an idea about X and Y , except that we do not know the relative location and orientation of the two point-clouds.

More precisely, if $H = PQ'$ is full rank decomposition of H , then the solutions X and Y of our system of equations $\delta_{ij}^2 = d_{ij}^2(X, Y)$ can be written in the form

$$\begin{aligned} X &= (P + e_n \alpha') T, \\ Y &= (Q + e_m \beta') (T')^{-1}, \end{aligned} \quad (4)$$

which leaves us with only the $p(p+2)$ unknowns in α , β , and T still to be determined. By using the fact that the solution is invariant under translation and rotation, we can actually reduce this to $(1/2)p(p+3)$ parameters. One way to find these additional parameters is given in [10].

Instead of trying to find an exact solution, if one actually exists, by algebraic means, we can also define a multidimensional unfolding loss function and minimize it. In the most basic and classical form, we have the Stress loss function

$$\sigma(X, Y) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} (\delta_{ij} - d_{ij}(X, Y))^2. \quad (5)$$

This is identical to an ordinary multidimensional scaling problem in which the diagonal (row–row and column–column) weights are zero. Or, to put it differently, in unfolding, the dissimilarities between different row objects and different column objects are missing. Thus, any multidimensional scaling program that can handle weights and missing data can be used to minimize this loss function. Details are in [7] or [1, Part III]. One can also consider measuring loss using SStress, the sum of squared differences between the squared dissimilarities and squared distances. This has been considered in [6, 11].

We use an example from [9, p. 152]. The Department of Psychology at the University of Nijmegen has, or had, nine different areas of research and teaching. Each of the 39 psychologists working in the department ranked all nine areas in order of relevance for their work. The areas are given in Table 1.

2 Multidimensional Unfolding

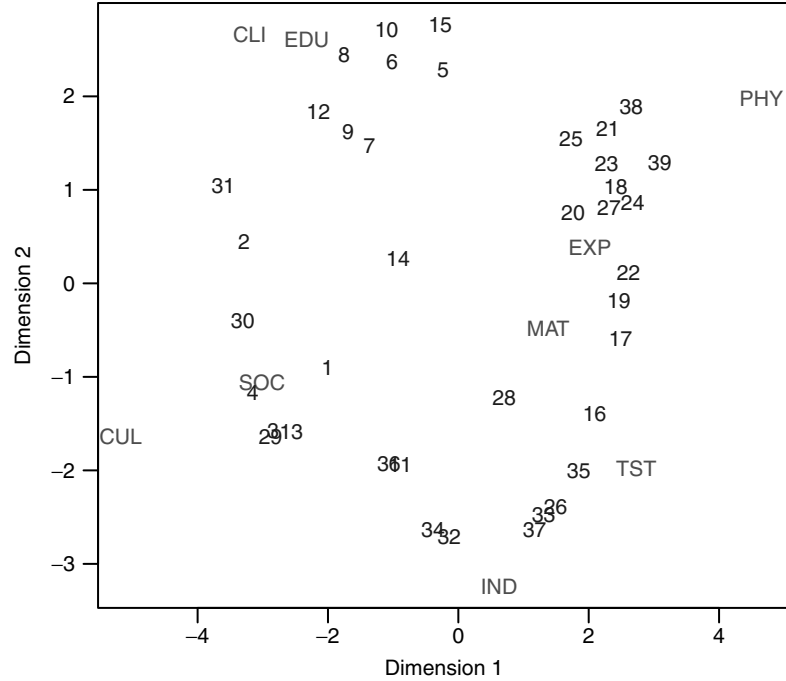


Figure 1 Metric Unfolding Roskam Data

We apply metric unfolding, in two dimensions, and find the solution in Figure 1.

In this analysis, we used the rank orders, more precisely the numbers 0 to 8. Thus, for good fit, first choices should coincide with ideal points. The grouping of the nine areas in the solution is quite natural and shows the contrast between the more scientific and the more humanistic and clinical areas.

Table 1 Nine psychology areas

Area	Plot code
Social Psychology	SOC
Educational and Developmental Psychology	EDU
Clinical Psychology	CLI
Mathematical Psychology and Psychological Statistics	MAT
Experimental Psychology	EXP
Cultural Psychology and Psychology of Religion	CUL
Industrial Psychology	IND
Test Construction and Validation	TST
Physiological and Animal Psychology	PHY

In this case, and in many other cases, the problems we are analyzing suggest that we really are interested in nonmetric unfolding. It is difficult to think of actual applications of metric unfolding, except perhaps in the life and physical sciences. This does not mean that metric unfolding is uninteresting. Most nonmetric unfolding algorithms solve metric unfolding subproblems, and one can often make a case for metric unfolding as a robust form to solve nonmetric unfolding problems.

The original techniques proposed by Coombs [4] were purely nonmetric and did not even lead to metric representations. In preference analysis, the prototypical area of application, we often only have ranking information. Each individual ranks a number of candidates, or food samples, or investment opportunities. The ranking information is row-conditional, which means we cannot compare the ranks given by individual i to the ranks given by individual k . The order is defined only within rows. Metric data are generally unconditional, because we can compare numbers both within and between rows. Because of the paucity of information (only rank order, only row-conditional, only off-diagonal), the usual Kruskal approach to

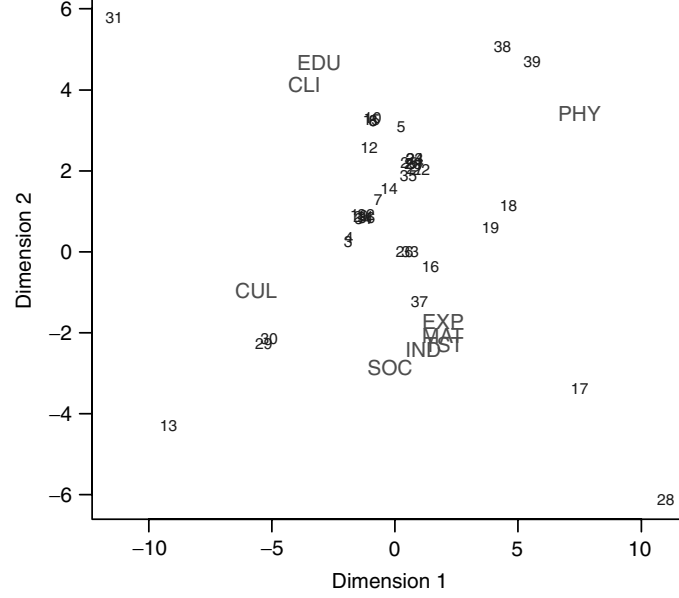


Figure 2 Nonmetric unfolding Roskam data

nonmetric unfolding often leads to degenerate solutions, even after clever renormalization and partitioning of the loss function [8]. In Figure 2, we give the solution minimizing

$$\sigma(X, Y, \Delta) = \sum_{i=1}^n \frac{\sum_{j=1}^m w_{ij}(\delta_{ij} - d_{ij}(X, Y))^2}{\sum_{j=1}^m w_{ij}(\delta_{ij} - \delta_{i*})^2} \quad (6)$$

over X and Y and over those Δ whose rows are monotone with the ranks given by the psychologists. Thus, there is a separate **monotone regression** computed for each of the 39 rows.

The solution is roughly the same as the metric one, but there is more clustering and clumping in the plot, and this makes the visual representation much less clear. It is quite possible that continuing to iterate to higher precision will lead to even more degeneracy. More recently, Busing et al. [2] have adapted the Kruskal approach to nonmetric unfolding by penalizing for the flatness of the monotone regression function.

One would expect even more problems when the data are not even rank orders but just binary choices. Suppose n individuals have to choose one alternative

from a set of m alternatives. The data can be coded as an *indicator matrix*, which is an $n \times m$ binary matrix with exactly one unit element in each row. The unfolding model says there are n points x_i and m points y_j in $\mathbb{R}^p$ such that if individual i picks alternative j , then $\|x_i - y_j\| \leq \|x_i - y_\ell\|$ for all $\ell = 1, \dots, m$. More concisely, we use the m points y_j to draw a *Voronoi diagram*. This is illustrated in Figure 3 for six points in the plane.

There is one Voronoi cell for each y_j , and the cell (which can be bounded or unbounded) contains exactly those points that are closer to y_j than to any of the other y_ℓ 's. The unfolding model says that individuals are in the Voronoi cells of the objects they pick. This clearly leaves room for a lot of indeterminacy in the actual placement of the points.

The situation becomes more favorable if we have more than one indicator matrix, that is, if each individual makes more than one choice. There is a Voronoi diagram for each choice and individuals must be in the Voronoi cells of the object they choose for each of the diagrams. Superimposing the diagrams creates smaller and smaller regions that each individual must be in, and the unfolding model requires the intersection of the Voronoi cells determined by the choices of any individual to be nonempty.

4 Multidimensional Unfolding

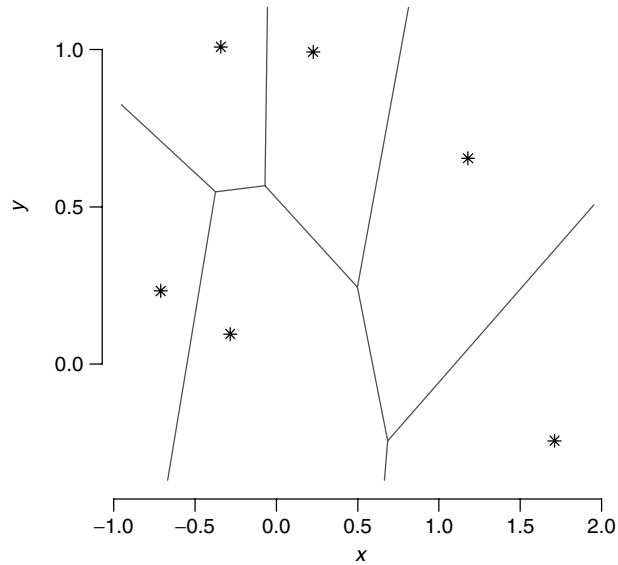


Figure 3 A Voronoi diagram

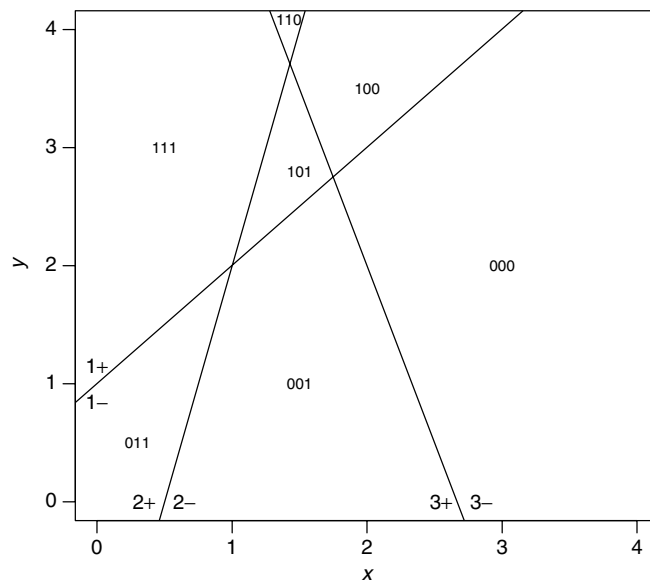


Figure 4 Unfolding binary data

It is perhaps simplest to apply this idea to binary choices. The Voronoi cells in this case are half spaces defined by hyperplanes dividing $\mathbb{R}^n$ in two parts. All individuals choosing the first of the two alternatives must be on one side of the hyperplane, all others must be on the other side. There is a hyperplane for each choice.

This is the nonmetric factor analysis model studied first by [5]. It is illustrated in Figure 4.

The prototype here is roll call data [3]. If 100 US senators vote on 20 issues, then the unfolding model says that (for a representation in the plane) there are 100 points and 20 lines, such that each issue-line separates the 'aye' and the 'nay' voters for

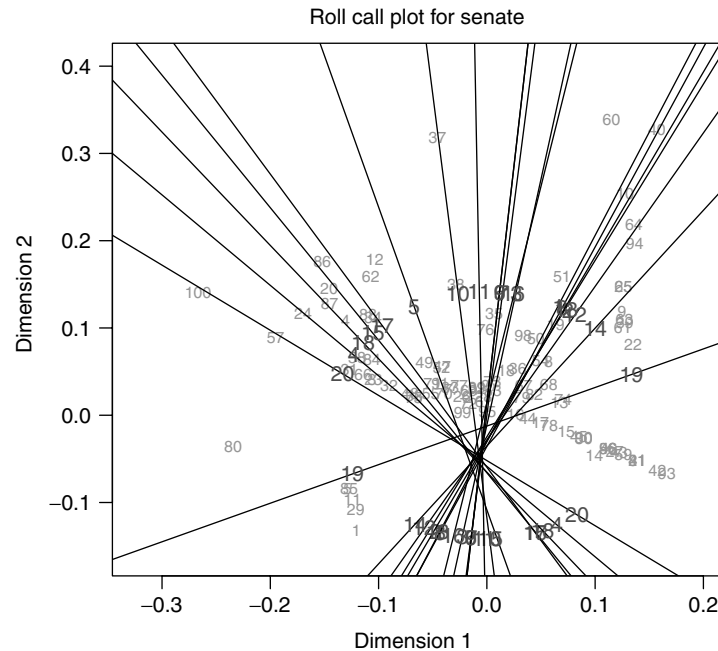


Figure 5 The 2000 US senate

that issue. Unfolding, in this case, can be done by **correspondence analysis**, or by maximum likelihood logit or probit techniques. We give an example, using 20 issues selected by Americans for Democratic Action, the 2000 US Senate, and the logit technique (Figure 5).

The issue lines are the perpendicular bisectors of the lines connecting the 'aye' and 'nay' points of the issues. We see for the figure how polarized American politics is, with almost all lines going through the center and separating Democrats from Republicans.

References

- [1] Borg, I. & Groenen, P.J.F. (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York.
- [2] Busing, F.M.T.A., Groenen, P.J.F. & Heiser, W.J. (2004). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation, *Psychometrika* (in press).
- [3] Clinton, J., Jackman, S. & Rivers, D. (2004). The statistical analysis of roll call data, *American Political Science Review* **98**, 355–370.
- [4] Coombs, C.H. (1964). *A Theory of Data*, Wiley.
- [5] Coombs, C.H. & Kao, R.C. (1955). *Nonmetric Factor Analysis*, *Engineering Research Bulletin* 38, Engineering Research Institute, University of Michigan, Ann Arbor.
- [6] Greenacre, M.J. & Browne, M.W. (1986). An efficient alternating least-squares algorithm to perform multidimensional unfolding, *Psychometrika* **51**, 241–250.
- [7] Heiser, W.J. (1981). *Unfolding analysis of proximity data*, Ph.D. thesis, University of Leiden.
- [8] Kruskal, J.B. & Carroll, J.D. (1969). Geometrical models and badness of fit functions, in *Multivariate Analysis*, Vol. II, P.R. Krishnaiah, ed., North Holland, pp. 639–671.
- [9] Roskam, E.E.C.H.I. (1968). *Metric analysis of ordinal data in psychology*, Ph.D. thesis, University of Leiden.
- [10] Schönemann, P.H. (1970). On metric multidimensional unfolding, *Psychometrika* **35**, 349–366.
- [11] Takane, Y., Young, F.W. & De Leeuw, J. (1977). Nonmetric individual differences in multidimensional scaling: an alternating least-squares method with optimal scaling features, *Psychometrika* **42**, 7–67.

JAN DE LEEUW

Multigraph Modeling

HARRY J. KHAMIS

Volume 3, pp. 1294–1296

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multigraph Modeling

13 ————— 23

Multigraph modeling is a graphical technique used to (1) identify whether a hierarchical **loglinear model** (HLM) is decomposable or not, (2) obtain an explicit factorization of joint probabilities in terms of marginal probabilities for decomposable HLMs, and (3) interpret the conditional independence structure (see **Probability: An Introduction**) of a given HLM of a **contingency table**. Consider a multiway contingency table and a corresponding HLM (see e.g., [1–3], or [11]). The HLM is uniquely characterized by its *generating class* (or *minimal sufficient configuration*), which establishes the correspondence between the model parameters and the associated minimal sufficient statistics. As an example, consider the model of conditional independence for a three-way contingency table (using Agresti’s [1] notation):

$$\log m_{ijk} = \mu + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ik}^{13} + \lambda_{jk}^{23}, \quad (1)$$

where m_{ijk} denotes the expected cell frequency for the i th row, j th column, and k th layer of the contingency table, and the parameters on the right side of the equation represent certain contrasts of $\log m_{ijk}$. The generating class for this model is denoted by [13][23]; it corresponds to the inclusion-maximal sets of indices in the model, $\{1, 3\}$ and $\{2, 3\}$, referred to as the *generators* of the model. This model represents conditional independence of factors 1 and 2 given factor 3, written as $[1 \otimes 2]3$.

A *graph* G is a mathematical object that consists of two sets: (1) a set of *vertices*, V , and (2) a set of *edges*, E , consisting of pairs of elements taken from V . The diagram of the graph is a picture in which a circle or dot or some other symbol represents a vertex and a line represents an edge. The *generator multigraph*, or simply *multigraph*, M of an HLM is a graph in which the vertex set consists of the set of generators of the HLM, and two vertices are joined by edges that are equal in number to the number of indices shared by them. The multigraph for the model in (1) is given in Figure 1. The vertices of this multigraph consist of the two generators of the model $\{1, 3\}$ and $\{2, 3\}$, and there is a single edge joining the two vertices because $\{1, 3\} \cap \{2, 3\} = \{3\}$. Note the one-to-one correspondence between the generating class and the multigraph representation.

Figure 1 Generator multigraph for the loglinear model [13] [23]

A *maximum spanning tree* T of a multigraph is a connected graph with no circuits (or closed loops) that includes each vertex of the multigraph such that the sum of all of the edges is maximum. Each maximum spanning tree consists of a family of sets of factor indices called the *branches* of the tree. For the multigraph in Figure 1, the maximum spanning tree is the edge (branch) joining the two vertices and is denoted by $T = \{3\}$.

An *edge cutset* of a multigraph is an inclusion-minimal set of multiedges whose removal disconnects the multigraph. For the model [13][23] with the multigraph given in Figure 1, there is a single-edge cutset that disconnects the two vertices, $\{3\}$, and it is the minimum number of edges that does so. General results concerning the multigraph are given as follows:

An HLM is *decomposable* if and only if [10]

$$d = \sum_{S \in V(T)} |S| - \sum_{S \in B(T)} |S|, \quad (2)$$

where d = number of factors in the contingency table, T is any maximum spanning tree of the multigraph, $V(T)$ and $B(T)$ are the set of vertices and set of branches of T respectively, and $S \in V(T)$ and $S \in B(T)$ represent the factor indices contained in $V(T)$ and $B(T)$ respectively.

For decomposable models, the joint distribution for the associated contingency table is [10]:

$$P[v_1, v_2, \dots, v_d] = \frac{\prod_{S \in V(T)} P[v : V \in S]}{\prod_{S \in B(T)} P[v : V \in S]}, \quad (3)$$

where $P[v_1, v_2, \dots, v_d]$ represents the probability associated with level v_1 of the first factor, level v_2 of the second factor, and so on, and level v_d of the d th factor. $P[v : V \in S]$ denotes the marginal probability indexed on those indices contained in S (and summing over all other indices). From this factorization, an explicit formula for the maximum likelihood estimator (see **Maximum Likelihood Estimation**) can be obtained (see, e.g., [2]).

2 Multigraph Modeling

In order to identify the conditional independence structure of a given HLM, the branches and edge cutsets of the multigraph are used. For a given multigraph M and set of factors S , construct the multigraph M/S by removing each factor of S from each generator (vertex in the multigraph) and removing each edge corresponding to that factor. For decomposable models, S is chosen to be a branch of any maximum spanning tree of M , and for nondecomposable models, S is chosen to be the factors corresponding to an edge cutset of M . Then, the conditional independence interpretation for the HLM is the following: The sets of factors in the disconnected components of M/S are mutually independent, conditional on S [10].

For the multigraph in Figure 1, $d = 3$, $T = \{3\}$, $V(T) = \{\{1, 3\}, \{2, 3\}\}$, and $B(T) = \{3\}$. By (2), $3 = (2 + 2) - 1$ so that the HLM [13][23] is decomposable, and by (3), the factorization of the joint distribution is (using simpler notation) $p_{ijk} = p_{i+k}p_{+jk}/p_{++k}$. Upon removing $S = \{3\}$ from the multigraph, the sets of factors in the disconnected components of M/S are $\{1\}$ and $\{2\}$. Hence, the interpretation of [13][23] is $[1 \otimes 2|3]$.

Edwards and Kreiner [6] analyzed a set of data in the form of a five-way contingency table from an investigation conducted at the Institute for Social Research, Copenhagen. A sample of 1592 employed men, 18 to 67 years old, were asked whether in the preceding year they had done any work they would have previously paid a craftsman to do. The variables included in the study are shown in Table 1.

One of the HLMs that fits the data well is $[AME][RME][AMT]$. The multigraph for this model is given in Figure 2. The maximum spanning tree for this multigraph is $T = \{\{M, E\}, \{A, M\}\}$, where $V(T) = \{\{A, M, E\}, \{R, M, E\}, \{A, M, T\}\}$ and $B(T) = \{\{M, E\}, \{A, M\}\}$. From (2), $5 = (3 + 3 + 3) - (2 + 2)$, indicating that $[AME][RME][AMT]$ is a decomposable model, and the factorization of the

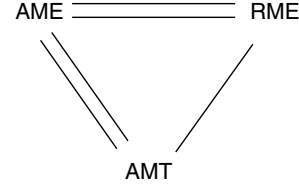


Figure 2 Generator multigraph for the loglinear model $[AME][RME][AMT]$

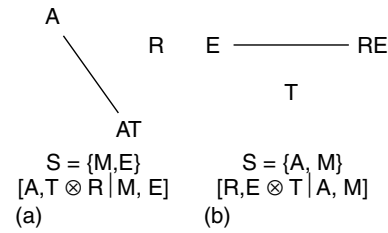


Figure 3 Multigraphs M/S and conditional independence interpretations for the loglinear model $[AME][RME][AMT]$ with branches: (a) $S = \{M, E\}$ and (b) $S = \{A, M\}$

joint probability can be obtained directly from (3). The multigraph M/S that results from choosing the branch $S = \{M, E\}$ is shown in Figure 3(a). The resulting interpretation is $[A, T \otimes R | M, E]$. Analogously, for $S = \{A, M\}$, we have $[R, E \otimes T | A, M]$ (Figure 3b).

The multigraph is typically smaller (i.e., has fewer vertices) than the first-order interaction graph introduced by Darroch et al. [4] (see also [5], [8], and [9]), especially for contingency tables with many factors and few generators. For the theoretical development of multigraph modeling, see [10]. For a further description of the application of multigraph modeling and additional examples, see [7].

References

- [1] Agresti, A. (2000). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.
- [2] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.
- [3] Christensen, R. (1990). *Log-linear Models*, Springer-Verlag, New York.
- [4] Darroch, J.N., Lauritzen, S.L. & Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables, *Annals of Statistics* **8**, 522–539.

Table 1

Variable	Symbol	Levels
Age	A	<30, 31–45, 46–67
Response	R	Yes, no
Mode of residence	M	Rent, own
Employment	E	Skilled, unskilled, other
Type of residence	T	Apartment, house

-
- [5] Edwards, D. (1995). *Introduction to Graphical Modelling*, Springer-Verlag, New York.
 - [6] Edwards, D. & Kreiner, S. (1983). The analysis of contingency tables by graphical models, *Biometrika* **70**, 553–565.
 - [7] Khamis, H.J. (1996). Application of the multigraph representation of hierarchical log-linear models, in *Categorical Variables in Developmental Research: Methods of Analysis*, Chap. 11, A. von Eye & C.C. Clogg, eds, Academic Press, New York.
 - [8] Khamis, H.J. & McKee, T.A. (1997). Chordal graph models of contingency tables, *Computers and Mathematics with Applications* **34**, 89–97.
 - [9] Lauritzen, S.L. (1996). *Graphical Models*, Oxford University Press, New York.
 - [10] McKee, T.A. & Khamis, H.J. (1996). Multigraph representations of hierarchical loglinear models, *Journal of Statistical Planning and Inference* **53**, 63–74.
 - [11] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Erlbaum, Hillsdale.

HARRY J. KHAMIS

Multilevel and SEM Approaches to Growth Curve Modeling

JOOP HOX AND REINOUD D. STOEL

Volume 3, pp. 1296–1305

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multilevel and SEM Approaches to Growth Curve Modeling

Introduction

A broad range of statistical methods exists for analyzing data from longitudinal designs (*see Longitudinal Data Analysis*). Each of these methods has specific features and the use of a particular method in a specific situation depends on such things as the type of research, the research question, and so on. The central concern of longitudinal research, however, revolves around the description of patterns of stability and change, and the explanation of how and why change does or does not take place [9].

A common design for longitudinal research in the social sciences is the panel or repeated measures design (*see Repeated Measures Analysis of Variance*), in which a sample of subjects is observed at more than one point in time. If all individuals provide measurements at the same set of occasions, we have a fixed occasions design. When occasions are varying, we have a set of measures taken at different points in time for different individuals. Such data occur, for instance, in growth studies, where individual measurements are collected for a sample of individuals at different occasions in their development (*see Growth Curve Modeling*). The data collection could be at fixed occasions, but the individuals have different ages. The distinction between fixed occasions designs and varying occasions designs is important, since they may lead to different analysis methods.

Several distinct statistical techniques are available for the analyses of panel data. In recent years, growth curve modeling has become popular [11, 12, 13, 24]. All subjects in a given population are assumed to have developmental curves of the same functional form (e.g., all linear), but the parameters describing their curves may differ. With linear developmental curves, for example, there may be individual differences in the initial level as well as in the growth rate or rate of change. Growth curve analysis is a statistical technique to estimate these parameters. Growth curve analysis is used to obtain a description of the mean growth in a population over a specific period of time (*see Growth Curve Modeling*). However, the

main emphasis lies in explaining variability between subjects in the parameters that describe their growth curves, that is, in interindividual differences in intra-individual change [25].

The model on which growth curve analysis is based, the growth curve model, can be approached from several perspectives. On the one hand, the model can be constructed as a standard two-level multilevel regression (MLR) model [4, 5, 20] (*see Linear Multilevel Models*). The repeated measures are positioned at the lowest level (level-1 or the occasion level), and are then treated as nested within the individuals (level-2 or the individual level), the same way as a standard cross-sectional multilevel model treats children as being nested within classes. The model can therefore be estimated using standard MLR software. On the other hand, the model can be constructed as a **structural equation model (SEM)**. Structural equation modeling uses latent variables to account for the relations between the observed variables, hence the name *latent growth curve (LGC)* model. The two approaches can be used to formulate equivalent models, providing identical estimates for a given data set [3].

The Longitudinal Multilevel or Latent Growth Curve Model

Both MLR and LGC incorporate the factor ‘time’ explicitly. Within the MLR framework time is modeled as an independent variable at the lowest level, the individual is defined at the second level, and explanatory variables can be included at all existing levels. The intercept and slope describe the mean growth. Interindividual differences in the parameters describing the growth curve are modeled as random effects for the intercept and slope of the time variable. The LGC approach adopts a latent variable view. Time is incorporated as specific constrained values for the factor loadings of the latent variable that represents the slope of the growth curve; all factor loadings of the latent variable that represents the intercept are constrained to the value of 1. The latent variable means for the intercept and slope factor describe the mean growth. Interindividual differences in the parameters describing the growth curve are modeled as the (co)variances of the intercept and slope factors. The mean and covariance structure of the latent variables in LGC analysis correspond to the fixed and

2 Multilevel and SEM Approaches to Growth Curve Modeling

random effects in MLR analysis, and this makes it possible to specify exactly the same model as a LGC or MLR model [23]. If this is done, exactly the same parameter estimates will emerge, as will be illustrated in the example.

The general growth curve model, for the repeatedly measured variable y_{ti} of individual i at occasion t , may be written as:

$$\begin{aligned} y_{ti} &= \lambda_{0t}\eta_{0i} + \lambda_{1t}\eta_{1i} + \gamma_{2t}x_{ti} + \varepsilon_{ti} \\ \eta_{0i} &= \nu_0 + \gamma_0z_i + \zeta_{0i} \\ \eta_{1i} &= \nu_1 + \gamma_1z_i + \zeta_{1i}, \end{aligned} \quad (1)$$

where λ_{1t} denotes the time of measurement and λ_{0t} a constant equal to the value of 1. Note that in a fixed occasions design λ_{1t} will typically be a consecutive series of integers (e.g., $[0, 1, 2, \dots, T]$) equal to all individuals, while in a varying occasions design λ_{1t} can take on different values across individuals. The individual intercept and slope of the growth curve are represented by η_{0i} and η_{1i} , respectively, with expectations ν_0 and ν_1 , and random departures or residuals, ζ_{0i} and ζ_{1i} , respectively. γ_{2t} represents the effect of the time-varying covariate x_{ti} ; γ_0 and γ_1 are the effects of the time-invariant covariate on the initial level and linear slope. Time-specific deviations are represented by the independent and identically standard normal distributed ε_{ti} , with variance σ_ε^2 . The variances of ζ_{0i} and ζ_{1i} , and their covariance are represented by:

$$\Sigma_\zeta = \begin{bmatrix} \sigma_0^2 & \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}. \quad (2)$$

Furthermore, it is assumed that $\text{cov}(\varepsilon_{it}, \varepsilon_{it'}) = 0$, $\text{cov}(\varepsilon_{it}, \eta_{i0}) = 0$, $\text{cov}(\varepsilon_{it}, \eta_{i1}) = 0$.

Within the longitudinal MLR model η_{0i} and η_{1i} are the random parameters, and λ_{1t} is an observed variable representing time. In the LGC model η_{0i} and η_{1i} are the latent variables and λ_{0t} and λ_{1t} are parameters, that is, factor loadings. Thus, the only difference between the models is the way time is incorporated in the model. In the MLR model time is introduced as a fixed explanatory variable, whereas in the LGC model it is introduced via the factor loadings. So, in the longitudinal MLR model an additional variable is added, and in the LGC model the factor loadings for the repeatedly measured variable are constrained in such a way that they represent time. The consequence of this is that with reference to the basic

growth curve model, MLR is essentially a univariate approach, with time points treated as observations of the same variable, whereas the LGC model is essentially a multivariate approach, with each time point treated as a separate variable [23]. Figure 1 presents a path diagram depicting a LGC model for four measurement occasions, for simplicity without covariates. Following SEM conventions, the first path for the latent slope factor, which is constrained to equal zero, is usually not present in the diagram.

The specific ways MLR and LGC model ‘time’ have certain consequences for the analysis. In the LGC approach, λ_{1t} cannot vary between subjects, which makes it best suited for a fixed occasions design. LGC modeling can be used for designs with varying occasions by modeling all existing occasions and viewing the varying occasions as a missing data problem, but when the number of existing cases is large this approach becomes unmanageable. In the MLR approach, λ_{1t} is simply a time-varying explanatory variable that can take on any variable, which makes MLR the best approach if there are a large number of varying occasions. There are also some differences between the LGC and MLR approach in the ways the model can be extended. In the LGC approach, it is straightforward to embed the LGC model in a larger path model, for instance, by combining several growth curves in one model, or by using the intercept and slope factors as predictors for outcome values measured at a later occasion. The MLR approach does not deal well with such extended models. On the other hand, in the MLR approach it is simple to add more levels, for instance to model a growth process of pupils nested in classes nested in schools. In the LGC approach, it is possible to embed a LGC model in a two-level structural equation model [14], but adding more levels is problematic.

Example

We will illustrate the application of both the MLR model and the LGC model using a hypothetical study in which data on the language acquisition of 300 children were collected during primary school at 4 consecutive occasions. Besides this, data were collected on the children’s intelligence, as well as, on each occasion a measure of their emotional well-being. The same data have been analyzed by Stoel, et al. [23], who also discuss extensions and applications of both models.

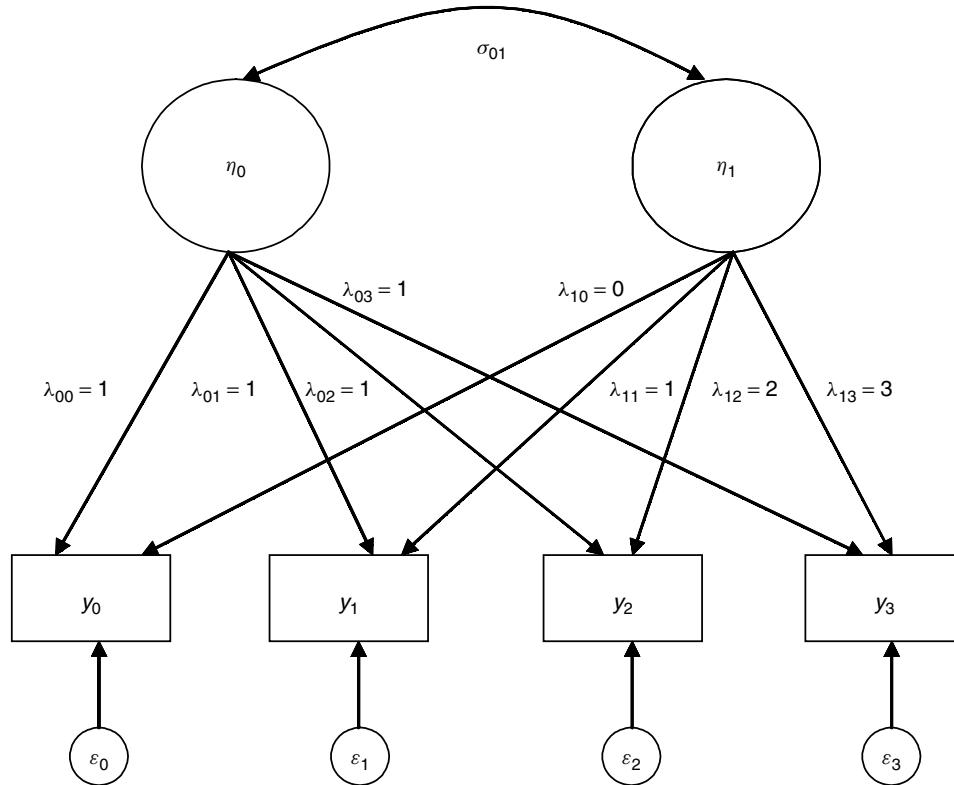


Figure 1 Path diagram of a four-wave latent growth curve model

The aim of the study is testing the hypothesis that there exists substantial growth in language acquisition, and that there is substantial variability between the children in their growth curves. Given interindividual differences in the growth curves, it is hypothesized that intelligence explains (part of) the interindividual variation in the growth

curves and that emotional well-being explains the time-specific deviations from the mean growth curve. The covariance matrix and mean vector are presented in Table 1.

Analyzing these data using both the MLR and LGC model with Maximum Likelihood estimation leads to the parameter estimates presented in Table 2.

Table 1 Covariance matrix and means vector

	y ₁	y ₂	y ₃	y ₄	x ₁	x ₂	x ₃	x ₄	z	means
y ₁	1.58									9.83
y ₂	1.28	4.15								11.72
y ₃	1.52	4.91	8.90							13.66
y ₄	1.77	6.83	11.26	17.50						15.65
x ₁	0.99	-0.08	-0.26	-0.33	2.17					0.00
x ₂	0.16	1.44	0.17	0.18	0.13	2.56				0.00
x ₃	0.07	-0.13	1.2	-0.27	-0.06	-0.09	2.35			0.00
x ₄	0.08	0.23	0.46	1.98	0.03	-0.04	0.10	2.24		0.00
z	0.34	1.28	2.01	3.01	-0.08	0.07	0.01	0.06	0.96	0.00

Note: y = language acquisition, x = emotional well-being, z = intelligence.

4 Multilevel and SEM Approaches to Growth Curve Modeling

Table 2 Maximum likelihood estimates of the parameters of (1), using multilevel regression and latent growth curve analysis

Parameter	MLR	LGC
<i>Fixed part</i>		
ν_0	9.89 (.06)	9.89 (.06)
ν_1	1.96 (.05)	1.96 (.05)
γ_0	0.40 (.06)	0.40 (.06)
γ_1	0.90 (.05)	0.90 (.05)
γ_2	0.55 (.01)	0.55 (.01)
<i>Random part</i>		
σ_ϵ^2	0.25 (.01)	0.25 (.01)
σ_0^2	0.78 (.08)	0.78 (.08)
σ_1^2	0.64 (.06)	0.64 (.06)
σ_{01}	0.00 (.05)	0.00 (.05)

Note: Standard errors are given in parentheses. The Chi-square test of model fit for the LGC model: $\chi^2(25) = 37.84(p = .95)$; RMSEA = 0.00. For the MLR model: $-2 \log \text{likelihood} = 3346.114$.

The first column of Table 2 presents the relevant parameters; the second and third columns show the parameter estimates of respectively the MLR, and LGC model.

As one can see in Table 2 the parameter estimates are the same and, consequently, both approaches would lead to the same substantive conclusions. According to the overall measure of fit provided by SEM, the model seems to fit the data quite well. Thus, the conclusions can be summarized as follows. After controlling for the effect of the covariates, a mean growth curve emerges with an initial level of 9.89 and a growth rate of 1.96. The significant variation between the subjects around these mean values implies that subjects start their growth process at different values and grow subsequently at different rates. The correlation between initial level and growth rate is zero. In other words, the initial level has no predictive value for the growth rate. Intelligence has a positive effect on both the initial level and growth rate, leading to the conclusion that children who are more intelligent show a higher score at the first measurement occasion and a greater increase in language acquisition than children with lower intelligence. Emotional well-being explains the time-specific deviations from the mean growth curve. That is, children with a higher emotional well-being at a specific time point show a higher score on language acquisition than is predicted by their growth curve.

Dichotomous and Ordinal Data

Both conventional structural equation modeling and multilevel regression analysis assume that the outcome variable(s) are continuous and have a (multivariate) normal distribution. In practice, many variables are measured as ordinal categorical variables, for example, the responses on a five- or seven-point Likert attitude question. Often, researchers treat such variables as if they were continuous and normal variables. If the number of response categories is fairly large and the response distribution is symmetric, treating these variables as continuous normal variable appears to work quite well. For instance, Bollen and Barb [2] show in a simulation study that if bivariate normal variables are categorized into at least five response categories, the differences between the correlation between the original variables and the correlation of the categorized variables is small (see **Categorizing Data**). Johnson and Creech [7] show that this also holds for parameter estimates and model fit. However, when the number of categories is smaller than five, the distortion becomes sizable. It is clear that such variables require special treatment.

A Categorical ordinal variable can be viewed as a crude observation of an underlying latent variable. The same model that is used for the continuous variables is used, but it is assumed to hold for the underlying latent response. The residuals are assumed to have a standard normal distribution, or a logistic distribution. The categories of the ordinal variable arise from applying thresholds to the latent continuous variable. Assume that we have an ordered categorical variable with three categories, for example, 'disagree', 'neutral', and 'agree.' The relation of this variable to the underlying normal latent variable is depicted in Figure 2.

The position on the latent variable determines which categorical response is observed. Specifically,

$$y_i = \begin{cases} 1, & \text{if } y_i^* \leq \tau_1 \\ 2, & \text{if } \tau_1 < y_i^* \leq \tau_2 \\ 3, & \text{if } \tau_2 < y_i^*, \end{cases} \quad (3)$$

where y_i is the observed categorical variable, y_i^* is the latent continuous variable, and τ_1 and τ_2 are the thresholds. Note that a dichotomous variable only has one threshold, which becomes the intercept in a regression equation.

To analyze ordered categorical data in a multilevel regression context, the common approach is

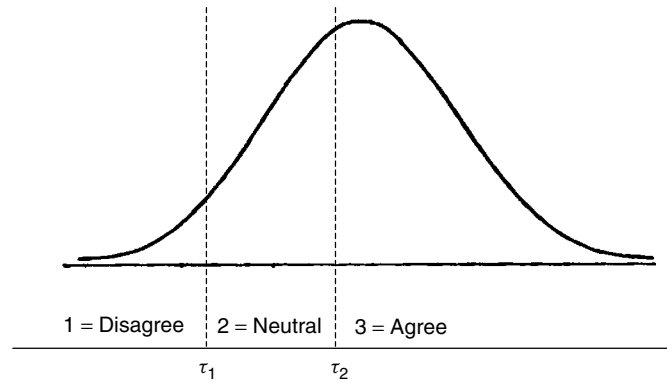


Figure 2 Illustration of two thresholds underlying a three-category variable

to assume a normal or a logistic distribution (*see Catalogue of Probability Density Functions*). There are several estimation methods available; most software relies on a Taylor series approximation to the likelihood, although some software is capable of numerical integration of the likelihood. For a discussion of estimation in nonnormal multilevel models computational details, we refer to the literature, for example [4, 20]. The common approach in structural equation modeling is to estimate **polychoric correlations**, that is, the correlations between the underlying latent responses, and apply conventional SEM methods. This approach also assumes a standard normal distribution for the residuals. For statistical issues and computational details, we refer to the literature, for example [1, 8] (*see Structural Equation Modeling: Categorical Variables*). The important point here for both types of modeling is that when the number of categories of an outcome variable is small, using an analysis approach that assumes continuous variables may lead to strongly biased results.

To give an indication of the extent of the bias, the outcome variables of the language acquisition example were dichotomized on their common median. This leads to a data set where the number of ‘zero’ and ‘one’ scores on the four Y variables taken together is 50% each, with the proportion of ‘one’ scores goes up from 0.04 through 0.44 at the first occasion to 0.72 to 0.81 at the last occasion. Table 3 presents the results of three multilevel analyses for a model with a linear random effect for time: (1) **Maximum likelihood estimation** on a standardized continuous outcome Z (mean zero and variance one across Y1 to Y4) assuming normality, (2) Maximum Likelihood estimation on the dichotomized outcome D assuming normality, and (3) approximate Maximum Likelihood estimation on the dichotomized outcome D assuming a logistic model and using a Laplace approximation (Laplace6 in HLM5, [22]).

The parameter estimates are different across all three methods. This is not surprising, because the Z -scores are standardized to a mean of 0 and a variance of 1, the dichotomous variables have an overall mean

Table 3 Estimates of the parameters of (1), using multilevel regression and different estimation methods

Parameter	ML on Z assuming normality	ML on D assuming normality	Approximate ML on D assuming logistic
<i>Fixed Part</i>			
ν_0	-0.82	0.11	-2.44
ν_1	0.54	0.26	1.80
<i>Random part</i>			
σ_ϵ^2	0.07	0.10	—
σ_0^2	0.07	0.01	0.24
σ_1^2	0.12	0.01	0.60
σ_{01}	0.03	0.01	0.38

of 0.5 and a variance of 0.25, and the underlying continuous variable y^* has a mean of 0 and a residual variance of approximately 3.29. However, statistical tests on the variance components lead to drastically different conclusions. An ordinary likelihood-ratio test on the variances is not appropriate, because the null-hypothesis is a value that lies on the boundary of the parameter space. Instead, we apply the chi-square test described by Raudenbush and Bryk [21], which is based on the residuals. The multilevel regression analysis on the continuous Z -scores leads to significant variance components ($p < 0.01$). The multilevel logistic regression analysis on the dichotomized D -scores leads to variance components that are also significant ($p < 0.01$). On the other hand, the multilevel regression analysis on the dichotomized D -scores using standard Maximum Likelihood estimation for continuous outcomes, leads to variance components that are *not* significant ($p > 0.80$). Thus, for our example data, using standard Maximum Likelihood estimation assuming a continuous outcome on the dichotomized variable leads to substantively different and in fact misleading conclusions.

Extensions

The model in (1) can be extended in a number of ways. We will describe some of these extensions in this section separately, but they can in fact be combined in one model.

Extending the Number of Levels

First, let us assume that we have collected data on several occasions from individuals within classes, and that there are (systematic) differences between classes in intercept and slope. The model in (1) can easily account for such a ‘three-level’ structure by adding the class-specific subscript j . The model then becomes:

$$\begin{aligned} y_{tij} &= \lambda_{0t}\eta_{0ij} + \lambda_{1t}\eta_{1ij} + \gamma_{2t}x_{tij} + \varepsilon_{tij} \\ \eta_{0ij} &= v_{0j} + \gamma_{0i}z_i + \zeta_{0ij} \\ \eta_{1ij} &= v_{1j} + \gamma_{1i}z_i + \zeta_{1ij} \\ v_{0j} &= v_0 + \zeta_{2j} \\ v_{1j} &= v_1 + \zeta_{3j}, \end{aligned} \quad (4)$$

and reflects the fact that the mean intercept and slope (v_{0j} and v_{1j} , respectively) may be different across classes. Note that (4) turns into (1) if ζ_{2j} and ζ_{3j} are constrained to zero. Incorporating class-level covariates and additional higher levels in the hierarchy are straightforward.

Extending the Measurement Model

Secondly, the model in (1) can be easily extended to include multiple indicators of a construct at each occasion explicitly. Essentially this extension merges the growth curve model with a measurement (latent factor) model at each occasion. For example if y_{ti} represented a mean score over R items, we may recognize that $y_{ti} = \sum_{r=1}^R y_{rti}/m$. Instead of modeling observed item parcels, the R individual items y_{rti} can be modeled directly. That is, to model them, explicitly, as indicators of a latent construct or factor at each measurement occasion. A growth model may then be constructed to explain the variance and covariance among the first-order latent factors. This approach has been termed *second-order growth modeling* in contrast to first-order growth modeling on the observed indicators. Different names for the same model are ‘curve-of-factors model’ and ‘multiple indicator latent growth model’ [12]. Note that modeling multiple indicators in a longitudinal setting requires a test on the structure of the measurements, that is a test of measurement invariance or factorial invariance [1, 8]. The model incorporating all y_{rti} explicitly then becomes:

$$\begin{aligned} y_{rti} &= \alpha_r + \lambda_r\eta_{ti} + \varepsilon_{ri} \\ \eta_{ti} &= \lambda_{0t}\eta_{0i} + \lambda_{1t}\eta_{1i} + \gamma_{2t}x_{ti} + \zeta_{ti} \\ \eta_{0i} &= v_0 + \gamma_{0i}z_i + \zeta_{0i} \\ \eta_{1i} &= v_1 + \gamma_{1i}z_i + \zeta_{1i}, \end{aligned} \quad (5)$$

where α_r and λ_r represent, respectively, the item-specific intercept and factor loading of item r , and ε_{ri} is a residual. η_{ti} is an individual and time-specific latent factor corresponding to y_{ti} of Model (1), ζ_{ti} is a random deviation corresponding to ε_{ti} of Model (1). The growth curve model is subsequently built on the latent factor scores η_{ti} with λ_{1t} representing the time of measurement and λ_{0t} a constant equal to the value of 1. This model thus allows for a separation of measurement error ε_{ri} and individual time-specific

deviation ζ_{ti} . In Model (1) these components are confounded in ε_{ti} .

Nonlinear Growth

The model discussed so far assumes linear growth. The factor time is incorporated explicitly in the model by constraining λ_{1t} in (1) explicitly to known values to represent the occasions at which the subjects are measured. This is, however, not a necessary restriction; it is possible to estimate a more general, that is nonlinear, model in which values of λ_{1t} are estimated (see **Nonlinear Mixed Effects Models; Nonlinear Models**). Thus, instead of constraining λ_{1t} to, for example $[0, 1, 2, 3 \dots T]$, some elements are left free to be estimated, providing information on the shape of the growth curve. For purposes of identification, at least two elements of λ_{1t} need to be fixed. The remaining values are then estimated to provide information on the shape of the curve; λ_{1t} then becomes $[0, 1, \lambda_{12}, \lambda_{13}, \dots, \lambda_{1T-1}]$. So, essentially, a linear model is estimated, while the nonlinear interpretation comes from relating the estimated λ_{1t} to the real time frame [13, 24]. The transformation of λ_{1t} to the real time frame gives the nonlinear interpretation.

Further Extensions

Further noteworthy extensions of the standard growth model in (1) which we will briefly sum up here are:

- The assumption of independent and identically distributed residuals can be relaxed. In other words, the model in (1) may incorporate a more complex type of residual structure. In fact, any type of residual structure can be implemented, provided the resulting model is identified.
- As stated earlier, the assumption that all individuals have been measured at the same measurement occasions as implied by Model (1) can be relaxed by giving λ_{1t} in (1) an individual subscript i . λ_{1ti} can subsequently be partly, or even completely different across individuals. However, using LGC modeling requires that we treat different λ_{1ti} 's across subjects as a balanced design with missing data (i.e., that not all subjects have been measured at all occasions), and assumptions about the missing data mechanism need to be made.

- Growth mixture modeling provides an interesting extension of conventional growth curve analysis. By incorporating a categorical latent variable it is possible to represent a mixture of subpopulations where population membership is not known but instead must be inferred from the data [15, 16, 18]. See Li et al. [10], for a didactic example of this methodology.

Estimation and Software

When applied to longitudinal data as described above, the MLR and LGC model are identical; they only differ in their representation. However, these models come from different traditions, and the software was originally developed to analyze different problems. This has consequences for the way the data are entered into the program, the choices the analyst must make, and the ease with which specific extensions of the model are handled by the software.

LGC modeling is a special case of the general approach known as structural equation modeling (SEM). Structural equation modeling is inherently a multivariate analyst method, and it is therefore straightforward to extend the basic model with other (latent or observed) variables. Standard SEM software abounds with options to test the fit of the model, compare groups, and constrain parameters within and across groups. This makes SEM a very flexible analysis tool, and LGC modeling using SEM shows this flexibility. Typically, the different measurement occasions are introduced as separate variables. Time-varying covariates are also introduced as separate variables that affect the outcome measures at the corresponding measurement occasions. Time invariant covariates are typically incorporated in the model by giving these an effect on the latent variables that represent the intercept or the slope. However, it is also possible to allow the time invariant covariates to have direct effects on the outcome variables at each measurement occasion. This leads to a different model. In LGC modeling using SEM, it is a straightforward extension to model effects of the latent intercept and slope variables on other variables, including analyzing two LGC trajectories in one model and investigating how their intercepts and slopes are related.

The flexibility of LGC analysis using SEM implies that the analyst is responsible for ensuring that the

model is set up properly. For instance, one extension of the basic LGC model discussed in the previous section is to use a number of indicators for the outcome measure and extend the model by including a measurement model. In this situation, the growth model is modeled on the consecutive latent variables. To ensure that the measurement model is invariant over time, the corresponding factor loadings for measurements across time must be constrained to be equal. In addition, since the LGC model involves changes in individual scores and the overall mean across time, means and intercepts are included in the model, and the corresponding intercepts must also be constrained equal over time.

Adding additional levels to the model is relatively difficult using the SEM approach. Muthén [14] has proposed a limited information Maximum Likelihood method to estimate parameters in two-level SEM. This approach works well [6], and can be implemented in standard SEM software [5]. Since the LGC model can be estimated using standard SEM, two-level SEM can include a LGC model at the individual (lowest) level, with groups at the second level. However, adding more levels is cumbersome in the SEM approach.

Multilevel Regression (MLR) is a univariate method, where adding an extra (lowest) level for the variables allows analysts to carry out multivariate analyses. So, growth curve analysis using MLR is accomplished by adding a level for the repeated measurement occasions. Most MLR software requires that the data matrix is organized by having a separate row for each measurement occasion within each individual, with the time invariant individual characteristics repeated for occasions within the same individual. Adding time-varying or time invariant covariates to the model is straightforward; they are just added as predictor variables. Allowing the time invariant covariates to have direct effects on the outcome variables at each measurement occasion is more complicated, because in the MLR approach this requires specifying interactions of these predictors with dummy variables that indicate the measurement occasions.

Adding additional levels is simple in MLR; after all, this is what the software was designed for. The maximum number of levels is defined by software restrictions; the current record is MLwiN [20], which is designed to handle up to 50 levels. This may seem excessive, but many special analyses are set up in

multilevel regression software by including an extra level. This is used, for instance, to model multivariate outcomes, cross-classified data, and specify measurement models. For such models, a nesting structure of up to five levels is not unusual, and not all multilevel regression software can accommodate this.

The MLR model is more limited than SEM. For instance, it is not possible to let the intercept and slopes act as predictors in a more elaborate path model. The limitations show especially when models are estimated that include latent variables. For instance, models with latent variables over time that are indicated by observed variables, easy to specify in SEM, can be set up in MLR using an extra variable level. At this (lowest) level, dummy variables are used to indicate variables that belong to the same construct at different measurement occasions. The regression coefficients for these dummies are allowed to vary at the occasion level, and they are interpreted as latent variables in a measurement model. However, this measurement model is more restricted than the measurement in the analogous SEM. In the MLR approach, the measurement model is a factor model with equal loadings for all variables, and one common error variance for the unique factors. In some situations, for instance, when the indicators are items measured using the same response scale, this restriction may be reasonable. It also ensures that the measurement model is invariant over time. The important issue is of course that this restriction cannot be relaxed in the MLR model, and it cannot be tested.

Most modern structural equation modeling software can be used to analyze LGC models. If the data are unbalanced, either by design or because of panel attrition, it is important that the software supports analyzing incomplete data using the raw Likelihood. If there are categorical response variables, it is important that the software supports their analysis. At the time of writing, only Muthén's software *Mplus* supports the combination of categorical incomplete data [17].

Longitudinal data can be handled by all multilevel software. Some software supports analyzing specific covariance structures over time, such as autoregressive models. When outcome variables may be categorical, there is considerable variation in the estimation methods employed. Most multilevel regression relies on Taylor series linearization, but increasingly numerical integration is used, which is regarded as more accurate.

A recent development in the field is that the distinction between MLR and LGC analysis is blurring. Advanced structural equation modeling software is now incorporating some multilevel features. *Mplus*, for example, goes a long way towards bridging the gap between the two approaches [15, 16]. On the other hand MLR software is incorporating features of LGC modeling. Two MLR software packages allow linear relations between the growth parameters: HLM [22] and GLLAMM [19]. HLM offers a variety of residual covariance structures for MLR models. The GLLAMM framework is especially powerful; it can be viewed as a multilevel regression approach that allows factor loadings, variable-specific unique variances, as well as structural equations among latent variables (both factors and random coefficients). In addition, it supports categorical and incomplete data. As the result of further developments in both statistical models and software, the two approaches to growth curve modeling may in time merge (*see Software for Statistical Analyses; Structural Equation Modeling: Software*).

Discussion

Many methods are available for the analysis of longitudinal data. There is no single preferred procedure, since different substantial questions dictate different data structures and statistical models. This entry focuses on growth curve analysis. Growth curve analysis, and its SEM variant latent growth curve analysis, has advantages for the study of change if it can be assumed that change is systematically related to the passage of time. Identifying individual differences in change, as well as understanding the process of change are considered critical issues in developmental research by many scholars. Growth curve analysis explicitly reflects on both intra-individual change and interindividual differences in such change.

In this entry, we described the general growth curve model and discussed differences between the multilevel regression approach and latent growth curve analysis using structural equation modeling. The basic growth curve model has a similar representation, and gives equivalent results in both approaches. Differences exist in the ways the model can be extended. In many instances, latent growth curve analysis is preferred because of its greater flexibility. Multilevel Regression is preferable if the

growth model must be embedded in a larger number of hierarchical data levels.

References

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [2] Bollen, K.A. & Barb, K. (1981). Pearson's R and coarsely categorized measures, *American Sociological Review* **46**, 232–239.
- [3] Curran, P. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavior Research* **38**(4), 529–569.
- [4] Goldstein, H. (2003). *Multilevel Statistical Models*, Edward Arnold, London.
- [5] Hox, J.J. (2002). *Multilevel Analysis: Techniques and Applications*, Lawrence Erlbaum, Mahwah.
- [6] Hox, J.J. & Maas, C.J.M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples, *Structural Equation Modeling* **8**, 157–174.
- [7] Johnson, D.R. & Creech, J.C. (1983). Ordinal measures in multiple indicator models: a simulation study of categorization error, *American Sociological Review* **48**, 398–407.
- [8] Kaplan, D. (2000). *Structural Equation Modeling*, Sage Publications, Thousand Oaks.
- [9] Kessler, R.C. & Greenberg, D.F. (1981). *Linear Panel Analysis*, Academic Press, New York.
- [10] Li, F., Duncan, T.E., Duncan, S.C. & Acock, A. (2001). Latent growth modeling of longitudinal data: a finite growth mixture modeling approach, *Structural Equation Modeling* **8**, 493–530.
- [11] McArdle, J.J. (1986). Latent variable growth within behavior genetic models, *Behavior Genetics* **16**, 163–200.
- [12] McArdle, J.J. (1988). Dynamic but structural equation modeling of repeated measures data, in *Handbook of Multivariate Experimental Psychology*, 2nd Edition, R.B. Cattell & J. Nesselroade, eds, Plenum Press, New York, pp. 561–614.
- [13] Meredith, W.M. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.
- [14] Muthén, B. (1994). Multilevel covariance structure analysis, *Sociological Methods & Research* **22**, 376–398.
- [15] Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling, in *New Methods for the Analysis of Change*, L.M. Collins & A.G. Sayer, eds, American Psychological Association, Washington, pp. 291–322.
- [16] Muthén, B.O. (2002). Beyond SEM: general latent variable modeling, *Behaviormetrika* **29**, 81–117.
- [17] Muthén, L.K. & Muthén, B.O. (2004). *Mplus Users Guide*, Muthén & Muthén, Los Angeles.

- [18] Muthén, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm, *Biometrics* **55**, 463–469.
- [19] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). Generalized multilevel structural equation modelling, *Psychometrika* **69**, 167–190.
- [20] Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. & Lewis, T. (2000). *A User's Guide to MlwiN*, Multilevel Models Project, University of London, London.
- [21] Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Newbury Park.
- [22] Raudenbush, S., Bryk, A., Cheong, Y.F. & Congdon, R. (2000). *HLM 5. Hierarchical Linear and Nonlinear Modeling*, Scientific Software International, Chicago.
- [23] Stoel, R.D., Van den Wittenboer, G. & Hox, J.J. (2003). Analyzing longitudinal data using multilevel regression and latent growth curve analysis, *Metodologia de las Ciencias del Comportamiento* **5**, 21–42.
- [24] Stoel, R.D., van den Wittenboer, G. & Hox, J.J. (2004). Methodological issues in the application of the latent growth curve model, in *Recent Developments in Structural Equation Modeling: Theory and Applications*, K. van Montfort, H. Oud & A. Satorra, eds, Kluwer Academic Publishers, Amsterdam, pp. 241–262.
- [25] Willett, J.B. & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time, *Psychological Bulletin* **116**, 363–381.

(See also **Structural Equation Modeling: Multilevel**)

JOOP HOX AND REINOUD D. STOEL

Multiple Baseline Designs

JOHN FERRON AND HEATHER SCOTT

Volume 3, pp. 1306–1309

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Baseline Designs

Single-case designs were developed to allow researchers to examine the effect of a treatment for a single case, where the case may be an individual participant or a group such as a class of students. Multiple-baseline designs [1] are an extension of the most basic single-case design, the AB, or interrupted time-series design. With an AB design, the researcher repeatedly measures the behavior of interest prior to intervention. These observations become the baseline phase (A). The researcher then introduces a treatment and continues to repeatedly measure the behavior, creating the treatment phase (B) of the design. If a change in behavior is observed, some may question whether the change was the result of the treatment or whether it resulted from maturation or some event that happened to coincide with treatment implementation. To allow researchers to rule out these alternative explanations for observed changes, more elaborate single-case designs were developed.

Multiple-baseline designs extend the AB design, such that a baseline phase (A) and treatment phase (B) are established for multiple participants, multiple behaviors, or multiple settings. The initiation of the treatment phases is staggered across time creating baselines of different lengths for the different participants, behaviors, or settings. The general form of the design with 12 observations and 4 baselines is presented in Figure 1.

To further illustrate the logic of this design, a graphical display of data from a multiple baseline across participant design is presented in Figure 2. When the first participant enters treatment, there is a notable change in behavior for this participant. The other participant, who is still in baseline, does not show appreciable changes in behavior when the treatment is initiated with the first participant. This makes history or maturational effects less plausible as explanations for why the first participant's behavior

changed. Put another way, when changes are due to history or maturation, we would anticipate change for each participant, but we would not expect those changes to stagger themselves across time in a manner that happened to coincide with the staggered interventions.

Applications

The multiple-baseline design is often employed in applied clinical and educational settings. However, its application may be extended to a variety of other disciplines and situations. The following are examples in which this design may be utilized:

- Educators might find the multiple baseline across individuals to be suitable for studying methods for reducing disruptive behaviors of students in the classroom.
- Psychologists might find the multiple baseline across behaviors to be suitable for investigating the effects of teaching children with autism to use socially appropriate gestures in combination with oral communication.
- Counselors might find the multiple baseline across participants to be suitable for studying the effects of training staff in self-supervision and uses of empathy in counseling situations.
- Therapists might find the multiple baseline across groups to be suitable for examining the effectiveness of treating anxiety disorders, phobias, and obsessive-compulsive disorders in adolescents.
- Retailers and grocery stores might find the multiple-baseline design across departments to be suitable for studying training effects on cleaning behaviors of employees.
- Teacher preparation programs might find the multiple baseline across participants to be suitable for examining the result of specific teaching practices.

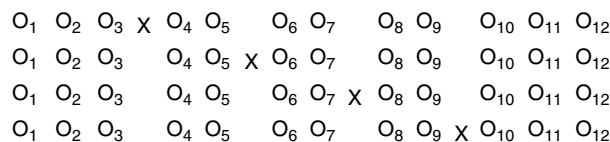


Figure 1 Diagram of a multiple-baseline design where *O*s represent observations and *X*s represent changes from baseline to treatment phases

2 Multiple Baseline Designs

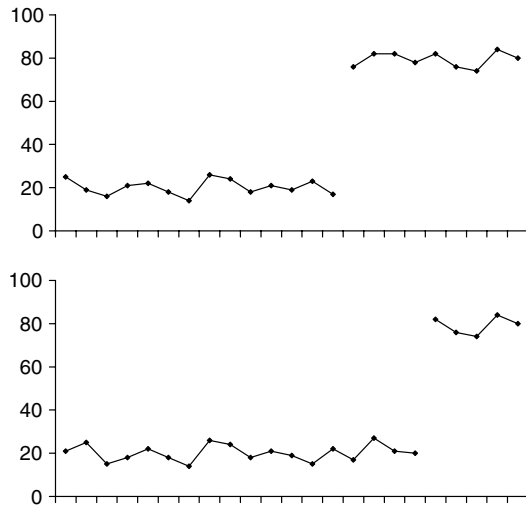


Figure 2 Graphical display of multiple-baseline design

- Medical practitioners or social workers might find the multiple baseline across behaviors to be suitable for studying the impact of teaching child care and safety skills to parents with intellectual disabilities.

Design Strengths and Weaknesses

The widespread use of multiple-baseline designs can be attributed to a series of practical and conceptual strengths. The multiple-baseline design allows researchers to focus extensively on a single participant or on a few participants. This may be advantageous in contexts where researchers wish to study cases that are relatively rare, when researchers are implementing particularly complex and time-consuming treatments, and when they are devoted to showing effects on specific individuals. Among single-case designs, the multiple-baseline design allows researchers to consider history and maturation effects without requiring them to withdraw the treatment. This can be seen as particularly valuable in clinical settings where it may not be ethical to withdraw an effective treatment.

Multiple-baseline designs provide relatively strong grounds for causal inferences when treatment effects can be seen that coincide with the unique intervention times for each participant, behavior, or setting. Things are less clear when there is a lack of independence between baselines so that treatment of one baseline

impacts another, or when treatment effects are not consistent across participants, behaviors, or settings. It should also be noted that the evidence for inferring a treatment effect for any one participant tends to be established less clearly than it could be in a single-case design involving a reversal. Finally, relative to group designs, the multiple-baseline design involves relatively few participants and thus there is little means for establishing generalizability.

Design Issues

More Complex Designs

The basic multiple-baseline design can be extended to have a more complex between series structure. For example, a researcher could conduct a multiple-baseline design across participants and behaviors by studying three behaviors of each of three participants. A researcher could also extend the multiple-baseline design by using a more complex within series structure. For example, a researcher conducting a multiple-baseline design across participants may for each participant include a baseline phase, followed by a treatment phase, followed by a second baseline phase. Whether one is considering the basic multiple-baseline design or a more complex extension, there are several additional design features that should be considered, including the number of baselines, the number of observations to be collected in each time series, and the decision about when to intervene within each series.

Number of Baselines

Multiple-baseline designs need to include at least two baselines, and the use of three or four baselines is more common. Assuming other things are equal, it is preferable to have greater numbers of baselines. Having more baselines provides a greater number of replications of the effect, and allows greater confidence that the observed effect in a particular time series was the result of the intervention.

Number of Observations

The number of observations within a time series varies greatly from one multiple-baseline study to the next. One study may have a phase with two

or three observations, while another may have more than 30 observations in each phase. When other things are equal, greater numbers of observations lead to stronger statements about treatment effects. The number of observations needed depends heavily on the amount of variation in the baseline data and the size of the anticipated effect. Suppose that a researcher wishes to intervene with four students that consistently spend no time on a task during a particular activity (i.e., 0% is observed for each baseline observation). Suppose further that the intervention is expected to increase the time on the task to above 80%. Under these conditions, a few observations within each phase is ample. However, if the baseline observations fluctuated between 0% and 80% with an average of 40% and the hope was to move the average to at least 60% for each student, many more observations would be needed.

Placement of Intervention Points

Systematic Assignment. The points at which the researcher will intervene can be established in several different ways. In some cases, the researcher simply chooses the points *a priori* on the basis of obtaining an even staggering across time. For example, the researcher may choose to intervene for the four baselines at times 4, 6, 8, and 10 respectively. This method seems to work well when baseline observations are essentially constant. Under these conditions, temporal stability is assumed and the researcher uses what has been referred to as the scientific solution to causal inference. When baseline observations are more variable, inferences become substantially more difficult and one may alter the method of assigning intervention points to facilitate drawing treatment effect inferences.

Response-guided Assignment. One option is to use a response-guided strategy where the data are viewed and intervention points are chosen on the basis of the emerging data. For example, a researcher gathering baseline data on each of the four individuals may notice variability in each participant's behavior. The researcher may be able to identify sources of this variability and make adjustments to the experiment to control the identified factors. After controlling these factors, the baseline data may stabilize, and the researcher may feel comfortable intervening with the first participant. The researcher would continue

watching the accumulating observations waiting until the data for the first participant demonstrated the anticipated effect. At this point, if the baseline data were still stable for the other three participants, the researcher would intervene with the second participant. Interventions for the third and fourth participants would be made in a similar manner, each time waiting until the data had shown a clear pattern before beginning the next treatment phase.

The response-guided strategy works relatively well when initial sources of variability to the baseline data can be identified and controlled. Variability may be resulting from unreliability in the measurement process that could be corrected with training, or it may be resulting from changes in the experimental conditions. As examples, variation may be associated with changes in the time of the observation, changes in the activities taking place during the observation period, changes in the people present during the observation period, or changes in the events preceding the observation period. By further standardizing conditions, variation in the baseline data can be reduced. If near constancy can be obtained in the baseline data, temporal stability can be assumed, and treatment effects can be readily seen.

When baseline variability cannot be controlled, one may turn to statistical methods for making treatment effect inferences. Interestingly, the response-guided strategy for establishing intervention points can lead to difficulties in establishing inferences statistically. Researchers wishing to make statistical inferences may turn to an approach that includes some form of random assignment.

Random Assignment. One approach is to randomly assign participants to designated intervention times (*see Randomization*). To illustrate, consider a researcher who plans to conduct a multiple-baseline study across participants where 12 observations will be gathered on each of the four participants. The researcher could decide that the interventions would occur on the 4th, 6th, 8th, and 10th point in time. The researcher could then randomly decide which participant would be treated at the 4th point in time, which would be treated at the 6th point in time, which on the 8th, and which on the 10th.

A second method of randomization would be to randomly select an intervention point for a participant under the restriction that there would be at least n observations in each phase. For example, the

researcher could randomly choose a time point for each participant under the constraint that the chosen point would fall between the 4th and 10th observation. It may turn out that a researcher chooses the interventions for the four participants to coincide with the 8th, 4th, 7th, and 5th observations, respectively. This leads to more possible assignments than the first method, but could possibly lead to the assignment of all interventions to the same point in time. This would be inconsistent with the temporal staggering, which is a defining feature in the multiple-baseline design, so one may wish to further restrict the randomization so this is not possible.

One way of doing this is to randomly choose intervention times under greater constraints and then randomly assign participants to intervention times. For example, a researcher could choose four intervention times by randomly choosing the 3rd or 4th, then randomly choosing the 5th or 6th, then randomly choosing the 7th or 8th, and then randomly choosing the 9th or 10th. The four participants could then be randomly assigned to these four intervention times. Finally, one could consider coupling one of these randomization methods with a response-guided strategy. For example, a researcher could work to control sources of baseline variability, and then make the random assignment after the data had stabilized as much as possible.

Analysis

Researchers often start their analysis by constructing a graphical display of the multiple-baseline data. In many applications, a visual analysis of the graphical display is the only analysis provided. There are, however, a variety of additional methods available to help make inferences about treatment effects in multiple-baseline studies.

If one makes intervention assignments randomly, then it is possible to use a randomization test to make

an inference about the presence of a treatment effect. Randomization tests [2] have been developed for each of the randomization strategies previously described, and these tests are statistically valid even when history or maturational effects have influenced the data. It is important to note, however, that these tests do not lead to inferences about the size of the treatment effect. When baseline data are variable, effect size inferences are often difficult and researchers should consider statistical modeling approaches.

A variety of statistical models are available for time series data [3], including models with a relatively simple assumed error structure, such as a standard regression model that assumes an independent error structure, to more complex time series models that have multiple parameters to capture dependencies among the errors, such as an autoregressive moving average model. Different error structure assumptions will typically lead to different estimated standard errors and potentially different inferential statements about the size of treatment effects. Consequently, care should be taken to establish the credibility of the assumed statistical model, which often requires a relatively large number of observations.

References

- [1] Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **1**, 91–97.
- [2] Koehler, M. & Levin, J. (1998). Regulated randomization: a potentially sharper analytical tool for the multiple-baseline design, *Psychological Methods* **3**, 206–217.
- [3] McKnight, S.D., McKean, J.W. & Huitema, B.E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms, *Psychological Methods* **5**, 87–101.

JOHN FERRON AND HEATHER SCOTT

Multiple Comparison Procedures

H.J. KESELMAN, BURT HOLLAND AND ROBERT A. CRIBBIE

Volume 3, pp. 1309–1325

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Comparison Procedures

Introduction

When more than two groups are compared for the effects of a treatment variable, researchers frequently adopt multiple comparison procedures (MCPs) in order to specify the exact nature of the treatment effects. Comparisons between two groups, pairwise comparisons, are frequently of interest to applied researchers, comparisons such as examining the effects of two different types of drugs, from a set of groups administered different drugs, on symptom dissipation. Complex comparisons, nonpairwise comparisons, the comparison of say one group (e.g., a group receiving a placebo instead of an active drug) to the average effect of two other groups (e.g., say two groups receiving some amount of the drug—50 and 100 mg) are also at times of interest to applied researchers. In both cases, the intent of the researcher is to examine focused questions about the nature of the treatment variable, in contrast to the answer provided by an examination of a global hypothesis through an omnibus test statistic (e.g., say, the use of the analysis of variance F test in order to determine if there is an effect of $J > 2$ ($j = 1, \dots, J$) different types of drugs on symptom dissipation).

A recent survey of statistical practices of behavioral science researchers indicates that the Tukey [60] and Scheffé [55] MCPs are very popular methods for conducting pairwise and complex comparisons, respectively (Keselman et al., [28]). However, behavioral science researchers now have available to them a plethora of procedures that they can choose from when their interest lies in conducting pairwise and complex comparisons among their treatment group means. In most cases, these newer procedures will provide either a test that is more insensitive, that is more robust, to the nonconformity of applied data to the derivational assumptions (i.e., homogeneity of population variances and normally distributed data) of the classical procedures (Tukey and Scheffé) or will provide greater sensitivity (statistical **power**) to detect effects when they are present. Therefore, the purpose of our paper is to present a brief introduction to some of the newer methods for conducting multiple comparisons. Our presentation will predominately

be devoted to an examination of pairwise methods since researchers perform these comparisons more often than complex comparisons. However, we will discuss some methods for conducting complex comparisons among treatment group means. Topics that we briefly review prior to our presentation of newer MCPs will be methods of Type I error control and simultaneous versus stepwise multiple testing.

It is important to note that the MCPs that are presented in our paper were also selected for discussion, because researchers can, in most cases, obtain numerical results with a statistical package, and in particular, through the SAS [52] system of computer programs. The SAS system presents a comprehensive up-to-date array of MCPs (see [65]). Accordingly, we acknowledge at the beginning of our presentation that some of the material we present follows closely Westfall et al.'s presentation. However, we also present procedures that are not available through the SAS system. In particular, we discuss a number of procedures that we believe are either new and interesting ways of examining pairwise comparisons (e.g., the model comparison approach of Dayton [6]) or have been shown to be insensitive to the usual assumptions associated with some of the procedures discussed by Westfall et al. (e.g., MCPs based on robust estimators).

Type I Error Control

Researchers who test a hypothesis concerning mean differences between two treatment groups are often faced with the task of specifying a significance level, or decision criterion, for determining whether the difference is significant. The level of significance specifies the maximum probability of rejecting the null hypothesis when it is true (i.e., committing a Type I error). As α decreases, researchers can be more confident that rejection of the null hypothesis signifies a true difference between population means, although the probability of not detecting a false null hypothesis (i.e., a Type II error) increases. Researchers faced with the difficult, yet important, task of quantifying the relative importance of Type I and Type II errors have traditionally selected some accepted level of significance, for example $\alpha = .05$.

However, determining how to control Type I errors is much less simple when multiple tests of significance (e.g., all possible pairwise comparisons between group means) will be computed. This is

2 Multiple Comparison Procedures

because when multiple tests of significance are computed, how one chooses to control Type I errors can affect whether one can conclude that effects are statistically significant or not.

The multiplicity problem in statistical inference refers to selecting the statistically significant findings from a large set of findings (tests) to support or refute one's research hypotheses. Selecting the statistically significant findings from a larger pool of results that also contain nonsignificant findings is problematic since when multiple tests of significance are computed, the probability that at least one will be significant by chance alone increases with the number of tests examined (*see Error Rates*).

Discussions on how to deal with multiplicity of testing have permeated many literatures for decades and continue to this day. In one camp are those who believe that the occurrence of any false positive must be guarded at all costs (see [13], [40], [48, 49, 50], [66]); that is, as promulgated by Thomas Ryan, pursuing a false lead can result in the waste of much time and expense, and is an error of inference that accordingly should be stringently controlled. Those in this camp deal with the multiplicity issue by setting α for the entire set of tests computed.

For example, in the pairwise multiple comparison problem, Tukey's [60] MCP uses a critical value wherein the probability of making at least one Type I error in the set of pairwise comparisons tests is equal to α . This type of control has been referred to in the literature as family-wise or experiment-wise (FWE) control. These respective terms come from setting a level of significance over all tests computed in an experiment, hence experiment-wise control, or setting the level of significance over a set (family) of conceptually related tests, hence FWE control. Multiple comparisonists seem to have settled on the family-wise label. As indicated, for the set of pairwise tests, Tukey's procedure sets a FWE for the family consisting of all pairwise comparisons.

Those in the opposing camp maintain that stringent Type I error control results in a loss of statistical power and consequently important treatment effects go undetected (see [47], [54], [72]). Members of this camp typically believe the error rate should be set per comparison (the probability of rejecting a given comparison) (hereafter referred to as the comparison-wise error-CWE rate) and usually recommend a five percent level of significance, allowing the overall error rate (i.e., FWE) to inflate with the number of tests

computed. In effect, those who adopt comparison-wise control ignore the multiplicity issue.

For example, a researcher comparing four groups ($J = 4$) may be interested in determining if there are significant pairwise mean differences between any of the groups. If the probability of committing a Type I error is set at α for each comparison (comparison-wise control = CWE), then the probability that at least one Type I error is committed over all $C = 4(4 - 1)/2 = 6$ pairwise comparisons can be much higher than α . On the other hand, if the probability of committing a Type I error is set at α for the entire family of pairwise comparisons, then the probability of committing a Type I error for each of the C comparisons can be much lower than α . Clearly then, the conclusions of an experiment can be greatly affected by the level of significance and unit of analysis over which Type I error control is imposed.

The FWE rate relates to a family (containing, in general, say k elements) of comparisons. A family of comparisons, as we indicated, refers to a set of conceptually related comparisons, for example, all possible pairwise comparisons, all possible complex comparisons, trend comparisons, and so on. As Miller [40] points out, specification of a family of comparisons, being self-defined by the researcher, can vary depending on the research paradigm. For example, in the context of a one-way design, numerous families can be defined: A family of all comparisons performed on the data, a family of all pairwise comparisons, a family of all complex comparisons. (Readers should keep in mind that if multiple families of comparisons are defined (e.g., one for pairwise comparisons and one for complex comparisons), then given that erroneous conclusions can be reached within each family, the overall Type I FWE rate will be a function of the multiple subfamily-wise rates.)

Specifying family size is a very important component of multiple testing. As Westfall et al. [65, p. 10] note, differences in conclusions reached from statistical analyses that control for multiplicity of testing (FWE) and those that do not (CWE) are directly related to family size. That is, the larger the family size, the less likely individual tests will be found to be statistically significant with family-wise control. Accordingly, to achieve as much sensitivity as possible to detect true differences and yet maintain control over multiplicity effects, Westfall et al. recommend that researchers 'choose smaller, more focused families rather than broad ones, and

(to avoid cheating) that such determination must be made *a priori* . . .’

Definitions of the CWE and FWE rates appear in many sources (e.g., see [34], [48], [40], [59], [60]).

Controlling the FWE rate has been recommended by many researchers (e.g., [16], [45], [48], [50], [60]) and is ‘the most commonly endorsed approach to accomplishing Type I error control’ [56, p. 577]. Keselman et al. [28] report that approximately 85 percent of researchers conducting pairwise comparisons adopt some form of FWE control.

Although many MCPs purport to control FWE, some provide ‘strong’ FWE control while others only provide ‘weak’ FWE control. Procedures are said to provide strong control if FWE is maintained across all null hypotheses; that is, under the complete null configuration ($\mu_1 = \mu_2 = \dots = \mu_J$) and all possible partial null configurations (An example of a partial null hypothesis is $\mu_1 = \mu_2 = \dots = \mu_{J-1} \neq \mu_J$). Weak control, on the other hand, only provides protection for the complete null hypothesis, that is, not for all partial null hypotheses as well.

The distinction between strong and weak FWE control is important because as Westfall et al. [65] note, the two types of FWE control, in fact, control different error rates. Weak control only controls the Type I error rate for falsely rejecting the complete null hypothesis and accordingly allows the rate to exceed, say 5%, for the composite null hypotheses. On the other hand, strong control sets the error rate at, say 5%, for all (component) hypotheses. For example, if $CWE = 1 - (1 - 0.05)^{1/k}$, the family-wise rate is controlled in a strong sense for testing k independent tests. Examples of MCPs that only weakly control FWE are the Newman [41] Keuls [33] and Duncan [8] procedures.

As indicated, several different error rates have been proposed in the multiple comparison literature. The majority of discussion in the literature has focused on the FWE and CWE rates (e.g., see [34], [48], [40], [59], [60]), although other error rates, such as the false discovery rate (FDR) also have been proposed (e.g., see Benjamini & Hochberg [2]).

False Discovery Rate Control. Work in the area of multiple hypothesis testing is far from static, and one of the newer interesting contributions to this area is an alternative conceptualization for defining errors in the multiple-testing problem; that is the FDR, presented by Benjamini and Hochberg [2]. FDR is defined

by these authors as the expected proportion of the number of erroneous rejections to the total number of rejections, that is, it is the expected proportion of false discoveries or false positives.

Benjamini and Hochberg [2] provide a number of illustrations where FDR control seems more reasonable than family-wise or comparison-wise control. Exploratory research, for example, would be one area of application for FDR control. That is, in new areas of inquiry where we are merely trying to see what parameters might be important for the phenomenon under investigation, a few errors of inference should be tolerable; thus, one can reasonably adopt the less stringent FDR method of control which does not completely ignore the multiple-testing problem, as does comparison-wise control, and yet, provides greater sensitivity than family-wise control. Only at later stages in the development of our conceptual formulations does one need more stringent family-wise control. Another area where FDR control might be preferred over family-wise control, suggested by Benjamini and Hochberg [2], would be when two treatments (say, treatments for dyslexia) are being compared in multiple subgroups (say, children of different ages). In studies of this sort, where an overall decision regarding the efficacy of the treatment is not of interest but, rather where separate recommendations would be made within each subgroup, researchers likely should be willing to tolerate a few errors of inference and accordingly would profit from adopting FDR rather than family-wise control.

Very recently, use of the FDR criterion has become widespread when making inferences in research involving the human genome, where family sizes in the thousands are common. See the review by Dudoit, Shaffer and Boldrick [7], and references contained therein.

Since multiple testing with FDR tends to detect more significant differences than testing with FWE, some researchers may be tempted to automatically prefer FDR control to FWE control. We caution that researchers who use FDR should be obligated to explain, in terms of the definitions of the two criteria, why it is more appropriate to control FDR than FWE in the context of their research.

Types of MCPs

MCPs can examine hypotheses either simultaneously or sequentially. A simultaneous MCP conducts all

4 Multiple Comparison Procedures

comparisons regardless of whether the omnibus test, or any other comparison, is significant (or not significant) using a constant critical value. Such procedures are frequently referred to as simultaneous test procedures (STPs) (see Einot & Gabriel [11]). A sequential (stepwise) MCP considers either the significance of the omnibus test or the significance of other comparisons (or both) in evaluating the significance of a particular comparison; multiple critical values are used to assess statistical significance. MCPs that require a significant omnibus test in order to conduct pairwise comparisons have been referred to as protected tests.

MCPs that consider the significance of other comparisons when evaluating the significance of a particular comparison can be either step-down or step-up procedures. Step-down procedures begin by testing the most extreme test statistic and nonsignificance of the most extreme test statistics implies nonsignificance for less extreme test statistics. Step-up procedures begin by testing the least extreme test statistic and significance of least extreme test statistics can imply significance for larger test statistics. In the equal sample sizes case, if a smaller pairwise difference is statistically significant, so is a larger pairwise difference, and conversely. However, in the unequal sample-size cases, one can have a smaller pairwise difference be significant and a larger pairwise difference nonsignificant, if the sample sizes for the means comprising the smaller difference are much larger than the sample sizes for the means comprising the larger difference.

One additional point regarding STP and stepwise procedures is important to note. STPs allow researchers to examine simultaneous intervals around the statistics of interest whereas stepwise procedures do not (see, however, [4]).

Preliminaries

A mathematical model that can be adopted when examining pairwise and/or complex comparisons of means in a one-way completely randomized design is:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad (1)$$

where Y_{ij} is the score of the i th subject ($i = 1, \dots, n$) in the j th ($j = 1, \dots, J$) group ($\sum_j n = N$), μ_j is the j th group mean, and ϵ_{ij} is the random error for the i th subject in the j th group. In the typical application

of the model, it is assumed that the ϵ_{ij} s are normally and independently distributed and that the treatment group variances (σ_j^2 s) are equal. Relevant sample estimates include

$$\begin{aligned} \hat{\mu}_j &= \bar{Y}_j = \sum_{i=1}^n \frac{Y_{ij}}{n} \text{ and } \hat{\sigma}^2 = \text{MSE} \\ &= \sum_{j=1}^J \sum_{i=1}^n \frac{(Y_{ij} - \bar{Y}_j)^2}{J(n-1)}. \end{aligned} \quad (2)$$

Pairwise Comparisons

A **confidence interval** for a pairwise difference $\mu_j - \mu_{j'}$

$$\bar{Y}_j - \bar{Y}_{j'} \pm c_\alpha \hat{\sigma} \sqrt{\frac{2}{n}}, \quad (3)$$

where c_α is selected such that $\text{FWE} = \alpha$. In the case of all possible pairwise comparisons, one needs a c_α for the set such that they simultaneously surround the true differences with a specified level of significance. That is, for all $j \neq j'$, c_α must satisfy

$$\begin{aligned} P \left(\bar{Y}_j - \bar{Y}_{j'} - c_\alpha \hat{\sigma} \sqrt{\frac{2}{n}} \leq \mu_j - \mu_{j'} \right. \\ \left. \leq \bar{Y}_j - \bar{Y}_{j'} + c_\alpha \hat{\sigma} \sqrt{\frac{2}{n}} \right) = 1 - \alpha. \end{aligned} \quad (4)$$

A hypothesis for the comparison ($H_c : \mu_j - \mu_{j'} = 0$) can be examined with the test statistic:

$$t_c = \frac{\bar{Y}_j - \bar{Y}_{j'}}{(2 \text{MSE} / n)^{1/2}}. \quad (5)$$

MCPs that assume normally distributed data and homogeneity of variances are given below.

Tukey. Tukey [60] proposed a STP for all pairwise comparisons. Tukey's MCP uses a critical value obtained from the Studentized range distribution. In particular, statistical significance, with FWE control, is assessed by comparing

$$|t_c| \geq \frac{q(J, J(n-1))}{\sqrt{2}}, \quad (6)$$

where q is a value from the Studentized range distribution (see [34]) based on J means and $J(n-1)$ error degrees of freedom. Tukey's procedure can

be implemented in SASs [52] **generalized linear model (GLM)** program.

Tukey–Kramer [35]. It is also important to note that Tukey’s method, as well as other MCPs, can be utilized when group sizes are unequal. A pairwise test statistic for the unequal sample-size case would be

$$t_{j,j'} = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{MSE(1/n_j + 1/n_{j'})}} (j \neq j'). \quad (7)$$

Accordingly, the significance of a pairwise difference is assessed by comparing

$$|t_{j,j'}| > \frac{q(J, \Sigma_j(n_j-1))}{\sqrt{2}}. \quad (8)$$

Hayter [17] proved that the Tukey–Kramer MCP only approximately controls the FWE – the rate is slightly conservative, that is, the true rate of Type I error will be less than the significance level. The GLM procedure in SAS will automatically compute the Kramer version of Tukey’s test when group sizes are unequal.

Fisher–Hayter’s Two-stage MCP. Fisher [12] proposed conducting multiple t Tests on the C pairwise comparisons following rejection of the omnibus ANOVA null hypothesis (see [29], [34]). The pairwise null hypotheses are assessed for statistical significance by referring t_c to $t_{(\alpha/2, \nu)}$, where $t_{(\alpha/2, \nu)}$ is the upper $100(1 - \alpha/2)$ percentile from Student’s distribution with parameter ν . If the ANOVA F is nonsignificant, comparisons among means are not conducted; that is, the pairwise hypotheses are retained as null.

It should be noted that Fisher’s [12] least significant difference (LSD) procedure only provides Type I error protection via the level of significance associated with the ANOVA null hypothesis, that is, the complete null hypothesis. For other configurations of means not specified under the ANOVA null hypothesis (e.g., $\mu_1 = \mu_2 = \dots = \mu_{J-1} < \mu_J$), the rate of family-wise Type I error can be much in excess of the level of significance (Hayter [18]; Hochberg & Tamhane [20]; Keselman et al. [29]; Ryan, [51]).

Hayter [18] proposed a modification to Fisher’s LSD that would provide strong control over FWE. Like the LSD procedure, no comparisons are tested

unless the omnibus test is significant. If the omnibus test is significant, then H_c is rejected if:

$$|t_c| > \frac{q(J-1, \nu)}{\sqrt{2}}. \quad (9)$$

Studentized range critical values can be obtained through SASs PROBMC (see [65], p. 46).

It should be noted that many authors recommend Fisher’s two-stage test for pairwise comparisons when $J = 3$ (see [27], [37]). These recommendations are based on Type I error control, power and ease of computation issues.

Hochberg’s Sequentially Acceptive Step-up Bonferroni Procedure. In this procedure [19], the P values corresponding to the m statistics (e.g., t_c) for testing the hypotheses $H_1, \dots, H_m$ are ordered from smallest to largest. Then, for any $i = m, m-1, \dots, 1$, if $p_i \leq \alpha/(m-i+1)$, the Hochberg procedure rejects all $H_{i'} (i' \leq i)$. According to this procedure, therefore, one begins by assessing the largest P value, p_m . If $p_m \leq \alpha$, all hypotheses are rejected. If $p_m > \alpha$, then H_m is accepted and one proceeds to compare $p_{(m-1)}$ to $\alpha/2$. If $p_{(m-1)} \leq \alpha/2$, then all $H_i (i = m-1, \dots, 1)$ are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha/3$, and so on.

Shaffer’s Sequentially Rejective Bonferroni Procedure that Begins with an Omnibus Test. Like the preceding procedure, the P values associated with the test statistics are rank ordered. In Shaffer’s procedure [57], however, one begins by comparing the smallest P value, p_1 , to α/m . If $p_1 > \alpha/m$, statistical testing stops and all pairwise contrast hypotheses ($H_i, 1 \leq i \leq m$) are retained; on the other hand, if $p_1 \leq \alpha/m$, H_1 is rejected and one proceeds to test the remaining hypotheses in a similar step-down fashion by comparing the associated P values to α/m^* , where m^* is equal to the maximum number of true null hypotheses, given the number of hypotheses rejected at previous steps. Appropriate denominators for each α -stage test for designs containing up to ten treatment levels can be found in Shaffer’s Table 2.

Shaffer [57] proposed a modification to her sequentially rejective Bonferroni procedure which involves beginning this procedure with an omnibus test. (Though MCPs that begin with an omnibus test frequently are presented with the F test, other

omnibus tests (e.g., a range statistic) can also be applied to these MCPs.) If the omnibus test is declared nonsignificant, statistical testing stops and all pairwise differences are declared nonsignificant. On the other hand, if one rejects the omnibus null hypothesis, one proceeds to test pairwise contrasts using the sequentially rejective Bonferroni procedure previously described, with the exception that p_m , the smallest P value, is compared to a significance level which reflects the information conveyed by the rejection of the omnibus null hypothesis. For example, for $m = 6$, rejection of the omnibus null hypothesis implies at least one inequality of means and, therefore, p_6 is compared to $\alpha/3$, rather than $\alpha/6$.

Benjamini and Hochberg's (BH) FDR Procedure.

In this procedure [2], the P values corresponding to the m pairwise statistics for testing the hypotheses $H_1, \dots, H_m$ are ordered from smallest to largest, that is, $p_1 \leq p_2 \leq \dots \leq p_m$, where $m = J(J-1)/2$. Let k be the largest value of i for which $p_i \leq i/m\alpha$, then reject all $H_i, i = 1, 2, \dots, k$. According to this procedure one begins by assessing the largest P value, p_m , proceeding to smaller P values as long as $p_i > i/m\alpha$. Testing stops when $p_k \leq k/m\alpha$.

Benjamini and Hochberg [3] also presented a modified (adaptive) (BH-A) version of their original procedure that utilizes the data to estimate the number of true H_c s. [The adaptive BH procedure has only been demonstrated, *not proven*, to control FDR, and only in the independent case.] With the original procedure, when the number of true null hypotheses (C_T) is less than the total number of hypotheses, the FDR rate is controlled at a level less than that specified (α).

To compute the Benjamini and Hochberg [2] procedure, the p_c -values are ordered (smallest to largest) $p_1, \dots, p_c$, and for any $c = C, C-1, \dots, 1$, if $p_c \leq \alpha(c/C)$, reject all $H_{c'} (c' \leq c)$. If all H_c s are retained, testing stops. If any H_c is rejected with the criterion of the BH procedure, then testing continues by estimating the slopes $S_c = (1 - p_c)/(C + 1 - c)$, where $c = 1, \dots, C$. Then, for any $c = C, C-1, \dots, 1$, if $p_c \leq \alpha(c/\hat{C}_T)$, reject all $H_{c'} (c' \leq c)$, where $\hat{C}_T = \min[(1/S^*) + 1, C]$, $[x]$ is the largest integer less than or equal to x and S^* is the minimum value of S_c such that $S_c < S_{c-1}$. If all $S_c > S_{c-1}$, S^* is set at C .

One disadvantage of the BH-A procedure, noted by both Benjamini and Hochberg [3] and Holland and Cheung [21], is that it is possible for an H_c to be rejected with $p_c > \alpha$. Therefore, it is suggested, by both authors, that H_c only be rejected if: (a) the hypothesis satisfies the rejection criterion of the BH-A; and (b) $p_c \leq \alpha$. To illustrate this procedure, assume a researcher has conducted a study with $J = 4$ and $\alpha = .05$. The ordered P values associated with the $C = 6$ pairwise comparisons are: $p_1 = .0014$, $p_2 = .0044$, $p_3 = .0097$, $p_4 = .0145$, $p_5 = .0490$, and $p_6 = .1239$. The first stage of the BH-A procedure would involve comparing $p_6 = .1239$ to $\alpha(c/C) = .05(6/6) = .05$. Since $.1239 > .05$, the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/C) = .05(5/6) = .0417$. Again, since $.0490 > .0417$, the procedure would continue by comparing $p_4 = .0145$ to $\alpha(c/C) = .05(4/6) = .0333$. Since $.0145 < .0333$, H_4 would be rejected. Because at least one H_c was rejected during the first stage, testing continues by estimating each of the slopes, $S_c = (1 - p_c)/(C - c + 1)$, for $c = 1, \dots, C$. The calculated slopes for this example are: $S_1 = .1664$, $S_2 = .1991$, $S_3 = .2475$, $S_4 = .3285$, $S_5 = .4755$ and $S_6 = .8761$. Given that all $S_c > S_{c-1}$, S^* is set at $C = 6$. The estimated number of true nulls is then determined by $\hat{C}_T = \min[(1/S^*) + 1, C] = \min[(1/6) + 1, 6] = \min[1.1667, 6] = 1$. Therefore, the BH-A procedure would compare $p_6 = .1239$ to $\alpha(c/\hat{C}_T) = .05(6/1) = .30$. Since $.1239 < .30$, but $.1239 > \alpha$, H_6 would not be rejected and the procedure would continue by comparing $p_5 = .0490$ to $\alpha(c/\hat{C}_T) = .05(5/1) = .25$. Since $.0490 < .25$ and $.0490 < \alpha$, H_5 would be rejected; in addition, all $H_{c'}$ would also be rejected (i.e., H_1, H_2, H_3 , and H_4).

Stepwise MCPs Based on the Closure Principle. As we indicated previously, researchers can adopt stepwise procedures when examining all possible pairwise comparisons, and typically they provide greater sensitivity to detect differences than do STPs, for example, Tukey's [60] method, while still maintaining strong FWE control. In this section, we present some methods related to closed-testing sequential MCPs that can be obtained through the SAS system of programs.

As Westfall et al. [65, p. 150] note, it was in the past two decades (from their 1999 publication) that a unified approach to stepwise testing has evolved.

The unifying concept has been the closure principle. MCPs based on this principle have been designated as closed-testing procedures. These methods are designated as closed-testing procedures because they address families of hypotheses that are closed under intersection ($\cap$). By definition, a closed family 'is one for which any subset intersection hypothesis involving members of the family of tests is also a member of the family'.

The closed-testing principle has led to a way of performing multiple tests of significance such that FWE is strongly controlled with results that are coherent. A coherent MCP is one that avoids inconsistencies in that it will not reject a hypothesis without rejecting all hypotheses implying it (see [20], pp. 44–45). Because closed-testing procedures were not always easy to derive, various authors derived other simplified stepwise procedures which are computationally simpler, though at the expense of providing smaller α values than what theoretically could be obtained with a closed-testing procedure. Naturally, as a consequence of having smaller α values (Type I errors are too tightly controlled), these simpler stepwise MCPs would not be as powerful as exact closed-testing methods. Nonetheless, these methods are still typically more powerful than STPs (e.g., Tukey) and therefore are recommended and furthermore, researchers can obtain numerical results through the SAS system.

REGWQ. One such method was introduced by Ryan [49], Einot and Gabriel [11] and Welsch [64] and is available through SAS. One can better understand the logic of the REGWQ procedure if we first introduce one of the most popular stepwise strategies for examining pairwise differences between means, the Newman-Keuls (NK) procedure.

In this procedure, the means are rank ordered from smallest to largest and the difference between the smallest and largest means is first subjected to a statistical test, typically with a range statistic (Q), at an α level of significance. If this difference is not significant, testing stops and all pairwise differences are regarded as null. If, on the other hand, this first range test is statistically significant, one 'steps-down' to examine the two $J - 1$ subsets of ordered means, that is, the smallest mean versus the next-to-largest mean and the largest mean versus the next-to-smallest mean, with each tested at an α level of significance. At each stage of testing, only subsets of ordered

means that are statistically significant are subjected to further testing. Although the NK procedure is very popular among applied researchers, it is becoming increasingly well known that when $J > 3$ it does not limit the FWE to α (see [20], p. 69).

Ryan [49] and Welsch [64], however, have shown how to adjust the subset levels of significance in order to provide strong FWE control. Specifically, in order to strongly control FWE a researcher must:

- Test all subset ($p = 2, \dots, J$) hypotheses at $\alpha_p = 1 - (1 - \alpha)^{\frac{p}{J}}$, for $p = 2, \dots, J - 2$ and at level $\alpha_p = \alpha$ for $p = J - 1, J$.
- Testing starts with an examination of the complete null hypothesis $\mu_1 = \mu_2 = \dots = \mu_J$, and if rejected one steps down to examine subsets of $J - 1$ means, $J - 2$ means, and so on.
- All subset hypotheses implied by a homogeneity hypothesis that has not been rejected are accepted as null without testing.

REGWQ can be implemented with the SAS GLM program. We remind the reader, however, that this procedure cannot be used to construct simultaneous confidence intervals.

MCPs that Allow Variances to be Heterogeneous

The previously presented procedures assume that the population variances are equal across treatment conditions. Given available knowledge about the non-robustness of MCPs with conventional test statistics (e.g., t , F), and evidence that population variances are commonly unequal (see [28], [67]), researchers who persist in applying MCPs with conventional test statistics increase the risk of Type I errors. As Olejnik and Lee [43, p. 14] conclude, 'most applied researchers are unaware of the problem [of using conventional test statistics with heterogeneous variances (see **Heteroscedasticity and Complex Variation**) and probably are unaware of the alternative solutions when variances differ'.

Although recommendations in the literature have focused on the Games–Howell [14], or Dunnett [10] procedures for designs with unequal σ_j^2 s (e.g., see [34], [59]), sequential procedures can provide more power than STPs while generally controlling the FWE (see [24], [36]).

8 Multiple Comparison Procedures

The SAS software can once again be used to obtain numerical results. In particular, Westfall et al. [65, pp. 206–207] provide SAS programs for logically constrained step-down pairwise tests when heteroscedasticity exists. The macro uses SASs mixed-model program (PROC MIXED), which allows for a nonconstant error structure across groups. As well, the program adopts the Satterthwaite [53] solution for error df. Westfall et al. remind the reader that the solution requires large data sets in order to provide approximately correct FWE control.

It is important to note that other non-SAS solutions are possible in the heteroscedastic case. For completeness, we note how these can be obtained. Specifically, sequential procedures based on the usual t_c statistic can be easily modified for unequal σ_j^2 s (and unequal n_j s) by substituting Welch's [62] statistic, $t_w(v_w)$ for $t_c(v)$, where

$$t_w = \frac{\bar{Y}_j - \bar{Y}_{j'}}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}}}, \quad (10)$$

and s_j^2 and $s_{j'}^2$ represent the sample variances for the j th and j' th group, respectively. This statistic is approximated as a t variate with critical value $t_{(1-\alpha/2, v_w)}$, the $100(1 - \alpha/2)$ quantile of Student's t distribution with df

$$v_w = \frac{\left\{ \frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}} \right\}^2}{\frac{(s_j^2/n_j)^2}{n_j - 1} + \frac{(s_{j'}^2/n_{j'})^2}{n_{j'} - 1}}. \quad (11)$$

For procedures, simultaneously comparing more than two means or when an omnibus test statistic is required (protected tests), robust alternatives to the usual **analysis of variance** (ANOVA) F statistic have been suggested. Possibly the best-known robust omnibus test is that due to Welch [63]. With the Welch procedure, the omnibus null hypothesis is rejected if $F_w > F_{(J-1, v_w)}$ where:

$$F_w = \frac{\sum_{j=1}^J w_j (\bar{Y}_j - \tilde{Y})^2 / (J - 1)}{1 + \frac{2(J - 2)}{(J^2 - 1)} \sum_{j=1}^J \frac{(1 - w_j / \Sigma w_j)^2}{n_j - 1}}, \quad (12)$$

$w_j = n_j / s_j^2$, and $\tilde{Y} = \Sigma w_j \bar{Y}_j / \Sigma w_j$. The statistic is approximately distributed as an F variate and

is referred to the critical value, $F_{(1-\alpha, J-1, v_w)}$, the $100(1 - \alpha)$ quantile of the F distribution with $J - 1$ and v_w df, where

$$v_w = \frac{J^2 - 1}{3 \sum_{j=1}^J \frac{(1 - w_j / \Sigma w_j)^2}{n_j - 1}}. \quad (13)$$

The Welch test has been found to be robust for largest to smallest variance ratios less than 10:1 (Wilcox, Charlin & Thompson [71]).

On the basis of the preceding, one can use the nonpooled Welch test and its accompanying df to obtain various stepwise MCPs. For example, Keselman et al. [27] verified that one can use this approach with Hochberg's [19] step-up Bonferroni MCP (see Westfall et al. [65], pp. 32–33) as well as with Benjamini and Hochberg's [2] FDR method to conduct all possible pairwise comparisons in the heteroscedastic case.

MCPs That Can be Used When Data are Nonnormal

An underlying assumption of all of the previously presented MCPs is that the populations from which the data are sampled are normally distributed; this assumption, however, may rarely be accurate (see [39], [44], [68]) (Tukey [61] suggests that most populations are skewed and/or contain outliers). Researchers falsely assuming normally distributed data risk obtaining biased Type I and/or Type II error rates for many patterns of nonnormality, especially when other assumptions are also not satisfied (e.g., variance homogeneity) (see [70]).

Bootstrap and Permutation Tests. The SAS system allows users to obtain both simultaneous and stepwise pairwise comparisons of means with methods that do not presume normally distributed data. In particular, users can use either **bootstrap** or **permutation** methods to compute all possible pairwise comparisons.

Bootstrapping allows users to create their own empirical distribution of the data and hence P values are accordingly based on the empirically obtained distribution, not a theoretically presumed distribution. For example, the empirical distribution, say $\hat{F}$, is obtained by sampling, *with replacement*, the pooled

sample residuals $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \bar{Y}_j$. That is, rather than assume that residuals are normally distributed, one uses empirically generated residuals to estimate the true shape of the distribution. From the pooled sample residuals one generates bootstrap data.

An example program for all possible pairwise comparisons is given by Westfall et al. [65, p. 229].

As well, pairwise comparisons of means (or ranks) can be obtained through permutation of the data with the program provided by Westfall et al. [65, pp. 233–234]. Permutation tests also do not require that the data be normally distributed. Instead of resampling with replacement from a pooled sample of residuals, permutation tests take the observed data $(Y_{11}, \dots, Y_{n1}, \dots, Y_{1J}, \dots, Y_{nJ})$ and randomly redistributes them to the treatment groups, and summary statistics (i.e., means or ranks) are then computed on the randomly redistributed data. The original outcomes (all possible pairwise differences from the original sample means) are then compared to the randomly generated values (e.g., all possible pairwise differences in the permutation samples).

When users adopt this approach to combat the effects of nonnormality they should take heed of the cautionary note provided by Westfall et al. [65, p. 234], namely, the procedure may not control the FWE when the data have heterogeneous variances, particularly when group sizes are unequal. Thus, we introduce another approach, pairwise comparisons based on robust estimators and a heteroscedastic statistic, an approach that has been demonstrated to generally control the FWE when data are non-normal and heterogeneous even when group sizes are unequal.

MCPs That Can be Used When Data are Neither Normal Nor When Variances are Heterogeneous

Trimmed Means Approach. A different type of testing procedure, based on trimmed (or censored) means (*see Trimmed Means*), has been discussed by Yuen and Dixon [73] and Wilcox [69, 70], and is purportedly robust to violations of normality.

That is, it is well known that the usual group means and variances, which are the bases for all of the previously described procedures, are greatly influenced by the presence of extreme observations

in distributions. In particular, the standard error of the usual mean can become seriously inflated when the underlying distribution has heavy tails. Accordingly, adopting a nonrobust measure ‘can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power’ [69, p. 66]. By substituting robust measures of location and scale for the usual mean and variance, it should be possible to obtain test statistics that are insensitive to the combined effects of variance heterogeneity and nonnormality.

While a wide range of robust estimators have been proposed in the literature (*see [15]*), the trimmed mean and Winsorized variance (*see Winsorized Robust Measures*) are intuitively appealing because of their computational simplicity and good theoretical properties [69]. The standard error of the trimmed mean is less affected by departures from normality than the usual standard error mean because extreme observations, that is, observations in the tails of a distribution, are censored or removed.

Trimmed means are computed by removing a percentage of observations from each of the tails of a distribution (set of observations). Let $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ represent the ordered observations associated with a group. Let $g = [\gamma n]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is notation for the largest integer not exceeding x . Wilcox [69] suggests that 20% trimming should be used. The effective sample size becomes $h = n - 2g$. Then the sample trimmed mean is

$$\bar{Y}_t = \frac{1}{h} \sum_{i=g+1}^{n-g} Y_{(i)}. \quad (14)$$

An estimate of the standard error of the trimmed mean is based on the Winsorized mean and Winsorized sum of squares (*see Winsorized Robust Measures*). The sample Winsorized mean is

$$\begin{aligned} \bar{Y}_w = \frac{1}{n} [& (g+1)Y_{(g+1)} + Y_{(g+2)} + \dots + Y_{(n-g-1)} \\ & + (g+1)Y_{(n-g)}], \end{aligned} \quad (15)$$

and the sample Winsorized sum of squared deviations is

$$SSD_w = (g+1)(Y_{(g+1)} - \bar{Y}_w)^2$$

$$\begin{aligned}
& + (Y_{(g+2)} - \bar{Y}_w)^2 + \cdots + (Y_{(n-g-1)} - \bar{Y}_w)^2 \\
& + (g+1)(Y_{(n-g)} - \bar{Y}_w)^2.
\end{aligned} \quad (16)$$

Accordingly, the sample Winsorized variance is $\hat{\sigma}_w^2 = SSD_w/(n-1)$ and the squared standard error of the mean is estimated as [58]

$$d = \frac{(n-1)\hat{\sigma}_w^2}{h(h-1)}. \quad (17)$$

To test a pairwise comparison null hypothesis, compute $\bar{Y}_t$ and d for the j th group, label the results $\bar{Y}_{tj}$ and d_j . The robust pairwise test (see Keselman, Lix & Kowalchuk [31]) becomes

$$t_w = \frac{\bar{Y}_{tj} - \bar{Y}_{tj'}}{\sqrt{d_j + d_{j'}}}, \quad (18)$$

with estimated df

$$v_w = \frac{(d_j + d_{j'})^2}{d_j^2/(h_j - 1) + d_{j'}^2/(h_{j'} - 1)}. \quad (19)$$

When trimmed means are being compared, the null hypothesis relates to the equality of population-trimmed means, instead of population means. Therefore, instead of testing $H_0: \mu_j = \mu_{j'}$, a researcher would test the null hypothesis, $H_0: \mu_{tj} = \mu_{tj'}$, where μ_t represents the population-trimmed mean (Many researchers subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters, when they are dealing with populations that are nonnormal in form.

Yuen and Dixon [73] and Wilcox [69] report that for long-tailed distributions, tests based on trimmed means and Winsorized variances can be much more powerful than tests based on the usual mean and variance. Accordingly, when researchers feel they are dealing with nonnormal data, they can replace the usual least squares estimators of central tendency and variability with robust estimators and apply these estimators in any of the previously recommended MCPs.

A Model Testing Procedure. The procedure to be described takes a completely different approach to specifying differences between the treatment group means. That is, unlike previous approaches which rely on a test statistic to reject or accept pairwise null hypotheses, the approach to be described uses an information criterion statistic to select a configuration

of population means which most likely corresponds with the observed data. Thus, as Dayton [6, p. 145] notes, ‘model-selection techniques are not statistical tests for which type I error control is an issue.’

When testing all pairwise comparisons, intransitive decisions are extremely common with conventional MCPs [6]. An intransitive decision refers to declaring a population mean (μ_j) not significantly different from two different population means ($\mu_j = \mu_{j'}$, $\mu_j = \mu_{j''}$), when the latter two means are declared significantly different ($\mu_{j'} \neq \mu_{j''}$). For example, a researcher conducting all pairwise comparisons ($J = 4$) may decide not to reject any hypotheses implied by $\mu_1 = \mu_2 = \mu_3$ or $\mu_3 = \mu_4$, but reject $\mu_1 = \mu_4$ and $\mu_2 = \mu_4$, based on results from a conventional MCP. Interpreting the results of this experiment can be ambiguous, especially concerning the outcome for μ_3 .

Dayton [6] proposed a model testing approach based on **Akaike’s Information Criterion (AIC)** [1]. Mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC is retained as the most probable population mean configuration, where:

$$AIC = SS_w + \sum_j n_j (\bar{Y}_j - \bar{Y}_{mj})^2 + 2q, \quad (20)$$

$\bar{Y}_{mj}$ is the estimated sample mean for the j th group (given the hypothesized population mean configuration for the m th model), SS_w is the ANOVA pooled within group sum of squares and q is the degree of freedom for the model. For example, for $J = 4$ (with ordered means) there would be $2^{J-1} = 8$ different models to be evaluated ($\{1234\}$, $\{1, 234\}$, $\{12, 34\}$, $\{123, 4\}$, $\{1, 2, 34\}$, $\{12, 3, 4\}$, $\{1, 23, 4\}$, $\{1, 2, 3, 4\}$). To illustrate, the model $\{12, 3, 4\}$ postulates a population mean configuration where groups one and two are derived from the same population, while groups three and four each represent independent populations. The model having the lowest AIC value would be retained as the most probable population model.

Dayton’s AIC model-testing approach has the virtue of avoiding intransitive decisions. It is more powerful in the sense of all-pairs power than Tukey’s MCP, which is not designed to avoid intransitive decisions. One finding reported by Dayton, as well as Huang and Dayton [25], is that the AIC has a slight bias for selecting more complicated models than the true model. For example, Dayton found

that for the mean pattern $\{12, 3, 4\}$, AIC selected the more complicated pattern $\{1, 2, 3, 4\}$ more than ten percent of the time, whereas AIC only rarely selected less complicated models (e.g., $\{12, 34\}$). This tendency can present a special problem for the complete null case $\{1234\}$, where AIC has a tendency to select more complicated models. Consequently, a recommendation by Huang and Dayton [25] is to use an omnibus test to screen for the null case, and then assuming rejection of the null, apply the Dayton procedure.

Dayton's [6] model testing approach can be modified to handle heterogeneous treatment group variances. Like the original procedure, mutually exclusive and transitive models are each evaluated using AIC, and the model having the minimum AIC is retained as the most probable population mean configuration. For heterogeneous variances:

$$\text{AIC} = -2 \left\{ \left(\frac{-N}{2} \right) (\ln(2\pi) + 1) - \frac{1}{2} (\sum n_j \ln(S)) \right\} + 2q, \quad (21)$$

where S is the biased variance for the j th group, substituting the estimated group mean (given the hypothesized mean configuration for the m th model) for the actual group mean in the calculation of the variance. As in the original Dayton procedure, an appropriate omnibus test can also be applied.

Complex Comparisons

To introduce some methods that can be adopted when investigating complex comparisons among treatment group means, we first expand on our introductory definitions. Specifically, we let

$$\psi = c_1\mu_1 + c_2\mu_2 + \cdots + c_J\mu_J, \quad (22)$$

(where the coefficients (c_j s) defining the contrast sum to one (i.e., $\sum_{j=1}^J c_j = 0$)) represent the population complex contrast that we are interested in subjecting to a test of significance. To test that $H_0: \psi = 0$, we replace the unknown population values with their **least squares estimates**, that is, the sample means, and subject the sample contrast estimate

$$\hat{\psi} = c_1\bar{Y}_1 + c_2\bar{Y}_2 + \cdots + c_J\bar{Y}_J \quad (23)$$

to a test with the statistic

$$t_{\hat{\psi}} = \frac{\hat{\psi}}{\sqrt{\text{MSE} \sum_{j=1}^J c_j^2/n_j}}. \quad (24)$$

As we indicated in the beginning of our paper, a very popular method for examining complex contrasts is Scheffé's [55] method. Scheffé's method, as we indicated, is an STP method that provides FWE control, that is, a procedure that uses one critical value to assess statistical significance of a set of complex comparisons. The simultaneous critical value is

$$\sqrt{(J-1)F_{1-\alpha, J-1, v}}, \quad (25)$$

where $F_{1-\alpha, J-1, v}$ is the $1-\alpha$ quantile from the sampling distribution of F based on $J-1$ and v numerator and denominator degrees of freedom (numerator and denominator df that are associated with the omnibus ANOVA F test). Accordingly, one rejects the hypothesis H_0 when

$$t_{\hat{\psi}} \geq \sqrt{(J-1)F_{1-\alpha, J-1, v}}. \quad (26)$$

Bonferroni. Another very popular STP method for evaluating a set of m complex comparisons is to compare the P values associated with the $t_{\hat{\psi}}$ statistics to the Dunn-Bonferroni critical value α/m or one may refer $|t_{\hat{\psi}}|$ to their simultaneous critical value (see Kirk [34], p. 829 for the table of critical values). As has been pointed out, researchers can compare, *a priori*, the Scheffé [55] and Dunn-Bonferroni critical values, choosing the smaller of the two, in order to obtain the more powerful of the two STPs. That is, the statistical procedure with the smaller critical value will provide more statistical power to detect true complex comparison differences between the means. In general though, if there are 20 or fewer comparisons, the Dunn-Bonferroni method will provide the smaller critical value and hence the more powerful approach with the reverse being the case, when there are more than 20 comparisons [65, p. 117].

One may also adopt stepwise MCPs when examining a set of m complex comparisons. These stepwise methods should result in greater sensitivity to detect effects than the Dunn-Bonferroni method. Many others have devised stepwise Bonferroni-type MCPs, for example, Holm [23], Holland and Copenhaver [22],

12 Multiple Comparison Procedures

Hochberg [19], Rom [46], and so on. All of these procedures provide FWE control; the minor differences between them can result in small differences in power to detect effects. Thus, we recommend the Hochberg [19] sequentially acceptable step-up Bonferroni procedure previously described because it is simple to understand and implement.

Finally, we note that the Scheffé [55], Dunn-Bonferroni [9] and Hochberg [19] procedures can be adopted to robust estimation and testing by adopting trimmed means and Winsorized variances, instead of the usual least squares estimators. That is, to circumvent the biasing effects of nonnormality and variance heterogeneity, researchers can adopt a heteroscedastic Welch-type statistic and its accompanying modified degrees of freedom, applying robust estimators. For example, the heteroscedastic test statistic for the Dunn-Bonferroni [9] and Hochberg [19] procedures would be

$$t_{\hat{\psi}} = \frac{\hat{\psi}}{\sqrt{\sum_{j=1}^J c_j^2 s_j^2 / n_j}}, \quad (27)$$

where the statistic is approximately distributed as a Student t variable with

$$df_w = \frac{\left(\sum_{j=1}^J c_j^2 s_j^2 / n_j \right)^2}{\sum_{j=1}^J [(c_j^2 s_j^2 / n_j)^2 / (n_j - 1)]}. \quad (28)$$

Accordingly, as we specified previously, one replaces the least squares means with trimmed means and least squares variances with variances based on Winsorized sums of squares. That is,

$$t_{\hat{\psi}_t} = \frac{\hat{\psi}_t}{\sqrt{\sum_{j=1}^J c_j^2 d_j}}, \quad (29)$$

where the sample comparison of trimmed means equals

$$\hat{\psi}_t = c_1 \bar{Y}_{t1} + c_2 \bar{Y}_{t2} + \cdots + c_J \bar{Y}_{tJ} \quad (30)$$

and error degrees of freedom is given by

$$df_{wt} = \frac{\left(\sum_{j=1}^J c_j^2 d_j \right)^2}{\sum_{j=1}^J [(c_j^2 d_j)^2 / (h_j - 1)]}. \quad (31)$$

A robust Scheffé [55] procedure would be similarly implemented, however, one would use the heteroscedastic Brown and Forsythe [5] statistic (see Kirk [34], p. 155). It is important to note that, although we are unaware of any empirical investigations that have examined tests of complex contrasts with robust estimators, based on empirical investigations related to pairwise comparisons, we believe these methods would provide approximate FWE control under conditions of nonnormality and variance heterogeneity and should possess more power to detect group differences than procedures based on least squares estimators.

Extensions

We have presented newer MCPs within the context of a one-way completely randomized design, highlighting procedures that should be robust to variance heterogeneity and nonnormality. We presented these methods because we concur with others that behavioral science data will not likely conform to these derivational requirements, assumptions which were adopted to derive the classical procedures, for example, [60] and [55]. For completeness, we note that robust procedures, that is, MCPs (omnibus tests as well) that employ a heteroscedastic test statistic and adopt robust estimators rather than the usual least squares estimators for the mean and variance, have also been enumerated within the context of factorial between-subjects and factorial between- by within-subjects repeated measures designs. Accordingly, applied researchers can note the generalization of the methods presented in our paper, to those that have been presented within these more general contexts by Keselman and Lix [30], Lix and Keselman [38], Keselman [26] and Keselman, Wilcox and Lix [32].

Numerical Example

We present a numerical example for the previously discussed MCPs so that the reader can check his/her facility to work with the SAS/Westfall et al. [65] programs and to demonstrate through example the differences between their operating characteristics. (Readers can obtain a copy of the SAS [52] and SPSS [42] syntax that we used to obtain numerical results from the third author.) In particular, the data ($n_1 = n_2 = \dots = n_J = 20$) presented in Table 1 were randomly generated by us though they could represent the outcomes of a problem solving task where the five groups were given different clues to solve the problem; the dependent measure was the time, in seconds, that it took to solve the task. The bottom two rows of the table contain the group means and standard deviations, respectively.

Table 2 contains FWE ($\alpha = .05$) significant (*) comparisons for the 10 pairwise comparisons for the five groups. The results reported in Table 2 generally conform, not surprisingly, to the properties of the MCPs that we discussed previously. In particular, of the ten comparisons, four were found to be statistically significant with the Tukey [60] procedure:

Table 1 Data values and summary statistics (means and standard deviations)

J_1	J_2	J_3	J_4	J_5
17	17	17	20	20
15	14	14	15	23
17	15	19	18	18
13	12	15	20	26
22	18	15	14	17
14	18	19	18	14
12	16	20	18	16
15	18	18	16	32
14	16	22	21	21
16	20	16	23	23
14	15	13	15	29
15	19	16	22	21
17	16	23	18	26
11	16	22	19	22
14	19	24	19	22
13	13	20	23	17
17	18	18	23	27
15	16	25	18	18
12	19	13	18	21
17	16	24	23	18
15.00	16.55	18.65	19.05	21.55
2.47	2.11	3.79	2.82	4.64

Table 2 FWE ($\alpha = .05$) significant (*) comparisons for the data from Table 1

$\mu_j - \mu_{j'}$	TK	HY	HC	SH	BH	BH-A	RG	BT	PM	TM
1 vs 2										
1 vs 3	*	*	*	*	*	*	*	*	*	*
1 vs 4	*	*	*	*	*	*	*	*	*	*
1 vs 5	*	*	*	*	*	*	*	*	*	*
2 vs 3						*				
2 vs 4					*	*				*
2 vs 5	*	*	*	*	*	*	*	*	*	*
3 vs 4										
3 vs 5		*	*	*	*	*	*	*	*	*
4 vs 5					*	*	*			

Note: TK-Tukey (1953); HY-Hayter (1986); HC-Hochberg (1988); SH-Shaffer (1979); BT-(Bootstrap)/PM (Permutation)-Westfall et al.; BH-Benjamini & Hochberg (1995); BH-A(Adaptive)-Benjamini & Hochberg (2002); RG-Ryan (1960)-Einot & Gabriel (1975)-Welsch (1977); TM-Trimmed means (and Winsorized variances) used with a nonpooled t Test and BH critical constants. Raw P values (1 vs 2, ..., 4 vs 5) for the SAS (1999) procedures are .1406, .0007, .0002, < .0001, .0469, .0185, < .0001, .7022, .0065 and .0185. The corresponding values for the trimmed means tests are .0352, .0102, .0001, .0003, .1428, .0076, .0044, .6507, .1271 and .1660.

$\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_1 - \mu_5$, and $\mu_2 - \mu_5$. In addition to these four comparisons, the Hayter [18], Hochberg [19], Shaffer [57], bootstrap [see 65] and permutation [see 65] MCPs found one additional comparison ($\mu_3 - \mu_5$) to be statistically significant. The REGWQ procedure also identified these comparisons as significant plus one additional one- $\mu_4 - \mu_5$. The BH [2] and BH-A [3] MCPs detected these contrasts as well, plus one and two others, respectively. In particular, BH and BH-A also identified $\mu_2 - \mu_4$ as significant, while BH-A additionally declared $\mu_2 - \mu_3$ significant. Thus, out of the ten comparisons, BH-A declared eight to be statistically significant. Clearly the procedures based on the more liberal FDR found more comparisons to be statistically significant than the FWE controlling MCPs. (Numerical results for BH-A were not obtained through SAS; they were obtained through hand-calculations.)

We also investigated the ten pairwise comparisons with the trimmed means and model-testing procedures; the results for the trimmed means analysis are also reported in Table 2. In particular, we computed the group trimmed means ($\bar{Y}_{t1} = 14.92$, $\bar{Y}_{t2} = 16.67$, $\bar{Y}_{t3} = 18.50$, $\bar{Y}_{t4} = 19.08$ and $\bar{Y}_{t5} = 21.08$) as well as the group Winsorized variances ($\hat{\sigma}_{W1}^2 = 2.58$, $\hat{\sigma}_{W2}^2 = 1.62$, $\hat{\sigma}_{W3}^2 = 8.16$, $\hat{\sigma}_{W4}^2 = 3.00$ and $\hat{\sigma}_{W5}^2 =$

= 10.26). These values can be obtained with the SAS/IML program discussed by Keselman et al. [32] or one can create a ‘special’ SPSS [42] data set to calculate nonpooled t statistics (t_W and v_W) and their corresponding P values (through the ONEWAY program) (These programs can be obtained from the first author). The results reported in Table 2 indicate that with this approach five comparisons were found to be statistically significant: $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_1 - \mu_5$, $\mu_2 - \mu_4$, and $\mu_2 - \mu_5$. Clearly, other MCPs had greater power to detect more pairwise differences. However, the reader should remember that robust estimation should result in more powerful tests when data are nonnormal as well as heterogeneous (see Wilcox [70]), which was not the case with our numerical example data. Furthermore, trimmed results were based on 12 subjects per group, not 20.

With regard to the model-testing approach, we examined the 2^{J-1} models of nonoverlapping subsets of ordered means and used the minimum AIC value to find the best model that ‘is expected to result in the smallest loss of precision relative to the true, but unknown, model (Dayton [6], p. 145).’ From the 16 models examined, the two models with the smallest AIC values were $\{1, 2, 34, 5\}$ ($AIC = 527.5$) and $\{12, 34, 5\}$ ($AIC = 527.8$). The ‘winning’ model combines one pair but clearly there is another model that is plausible given the data available. (Results were obtained through hand calculations. However, a GAUSS program is available from the Department of Measurement & Statistics, University of Maryland web site.) Though this ambiguity might

seem like a negative feature of the model-testing approach, Dayton [6] would maintain that being able to enumerate a set of conclusions (i.e., competing models) provides a broader more comprehensive perspective regarding group differences than does the traditional approach.

We also computed a set of complex contrasts (nine) among the five treatment group means to allow the reader to check his/her understanding of the computational operations associated with the MCPs that can be used when examining a set of complex contrasts. The population contrasts examined, their sample values, and the decisions regarding statistical significance ($FWE = .05$) are enumerated in Table 3. Again results conform to the operating characteristics of the MCPs. In particular, the Scheffé [55] procedure found the fewest number of significant contrasts, Hochberg’s [19] step-up Bonferroni procedure the most, and the number found to be significant according to the Dunn–Bonferroni [55] criterion was intermediate to Scheffé [55] and Hochberg [19]. When applying Hochberg’s [19] MCP with trimmed means and Winsorized variances six of the nine complex contrasts were found to be statistically significant.

References

- [1] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **AC-19**, 716–723.
- [2] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B* **57**, 289–300.

Table 3 FWE ($\alpha = .05$) significant (*) complex contrasts for the data from Table 1

Contrast (ψ)	$\hat{\psi}$	$\hat{\psi}_t$	Scheffé	Bonferroni	Hochberg	Hochberg/TM
$.5\mu_1 + .5\mu_2 - .33\mu_3 - .33\mu_4 - .33\mu_5$	−3.99	−3.78	*	*	*	*
$\mu_1 - .33\mu_3 - .33\mu_4 - .33\mu_5$	−4.77	−4.65	*	*	*	*
$\mu_2 - .33\mu_3 - .33\mu_4 - .33\mu_5$	−3.22	−2.90	*	*	*	*
$\mu_3 - .5\mu_4 - .5\mu_5$	−1.65	−1.58				
$-.5\mu_3 + \mu_4 - .5\mu_5$	−1.05	−0.71				
$.5\mu_3 + .5\mu_4 - \mu_5$	2.70	2.29		*	*	
$\mu_1 - .5\mu_2 - .5\mu_3$	−2.60	−2.67		*	*	*
$\mu_2 - .5\mu_3 - .5\mu_4$	−2.30	−2.13			*	
$.5\mu_1 + .5\mu_2 - \mu_3$	2.88	2.71	*	*	*	

Note: $\hat{\psi}$ is the value of the contrast for the original means; $\hat{\psi}_t$ is the value of the contrast for the trimmed means; Scheffé - Scheffé (1959); Hochberg-Hochberg (1988); Hochberg/TM-Hochberg FWE control with trimmed means (and Winsorized variances) and a nonpooled (Welch) t Test. Raw P values are $< .001$, $< .001$, $< .001$, .071, .248, .004, .005, .012, and .002. The corresponding values for the trimmed means tests are .0000, .0000, .0010, .2324, .5031, .1128, .0043, .0124, and .0339.

- [3] Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics* **25**, 60–83.
- [4] Bofinger, E., Hayter, A.J. & Liu, W. (1993). The construction of upper confidence bounds on the range of several location parameters, *Journal of the American Statistical Association* **88**, 906–911.
- [5] Brown, M.B. & Forsythe, A.B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances, *Biometrics* **30**, 719–724.
- [6] Dayton, C.M. (1998). Information criteria for the paired-comparisons problem, *The American Statistician* **52**, 144–151.
- [7] Dudoit, S., Shaffer, J.P. & Boldrick, J.C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science* **18**, 71–103.
- [8] Duncan, D.B. (1955). Multiple range and multiple F tests, *Biometrics* **11**, 1–42.
- [9] Dunn, O.J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association* **56**, 52–64.
- [10] Dunnett, C.W. (1980). Pairwise multiple comparisons in the unequal variance case, *Journal of the American Statistical Association* **75**, 796–800.
- [11] Einot, I. & Gabriel, K.R. (1975). A study of the powers of several methods of multiple comparisons, *Journal of the American Statistical Association* **70**, 574–583.
- [12] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [13] Games, P.A. (1971). Multiple comparisons of means, *American Educational Research Journal* **8**, 531–565.
- [14] Games, P.A. & Howell, J.F. (1976). Pairwise multiple comparison procedures with unequal n's and/or variances, *Journal of Educational Statistics* **1**, 113–125.
- [15] Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions, *Journal of the American Statistical Association* **71**, 409–416.
- [16] Hancock, G.R. & Klockars, A.J. (1996). The quest for α : Developments in multiple comparison procedures in the quarter century since Games (1971), *Review of Educational Research* **66**, 269–306.
- [17] Hayter, A.J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative, *Annals of Statistics* **12**, 61–75.
- [18] Hayter, A.J. (1986). The maximum familywise error rate of Fisher's least significant difference test, *Journal of the American Statistical Association* **81**, 1000–1004.
- [19] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.
- [20] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*, John Wiley & Sons, New York.
- [21] Holland, B. & Cheung, S.H. (2002). Family size robustness criteria for multiple comparison procedures, *Journal of the Royal Statistical Society, B* **64**, 63–77.
- [22] Holland, B.S. & Copenhaver, M.D. (1987). An improved sequentially rejective Bonferroni test procedure, *Biometrics* **43**, 417–423.
- [23] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.
- [24] Hsu, T. & Olejnik, S. (1994). Power of pairwise multiple comparisons in the unequal variance case, *Communications in Statistics: Simulation and Computation* **23**, 691–710.
- [25] Huang, C.J. & Dayton, C.M. (1995). Detecting patterns of bivariate mean vectors using model-selection criteria, *British Journal of Mathematical and Statistical Psychology* **48**, 129–147.
- [26] Keselman, H.J. (1998). Testing treatment effects in repeated measure designs: An update for psychophysiological researchers, *Psychophysiology* **35**, 70–478.
- [27] Keselman, H.J., Cribbie, R.A. & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control, *Psychological Methods* **4**, 58–69.
- [28] Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C. & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses, *Review of Educational Research* **68**, 350–386.
- [29] Keselman, H.J., Keselman, J.C. & Games, P.A. (1991). Maximum familywise Type I error rate: The least significant difference, Newman-Keuls, and other multiple comparison procedures, *Psychological Bulletin* **110**, 155–161.
- [30] Keselman, H.J. & Lix, L.M. (1995). Improved repeated measures stepwise multiple comparison procedures, *Journal of Educational and Behavioral Statistics* **20**, 83–99.
- [31] Keselman, H.J., Lix, L.M. & Kowalchuk, R.K. (1998). Multiple comparison procedures for trimmed means, *Psychological Methods* **3**, 123–141.
- [32] Keselman, H.J., Wilcox, R.R. & Lix, L.M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs, *Psychophysiology* **40**, 586–596.
- [33] Keuls, M. (1952). The use of the "Studentized range" in connection with an analysis of variance, *Euphytica* **1**, 112–122.
- [34] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Company, Toronto.
- [35] Kramer, C.Y. (1956). Extension of the multiple range test to group means with unequal numbers of replications, *Biometrics* **12**, 307–310.
- [36] Kromrey, J.D. & La Rocca, M.A. (1994). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances, *Journal of Experimental Education* **63**, 343–362.

- [37] Levin, J.R., Serlin, R.C. & Seaman, M.A. (1994). A controlled, powerful multiple-comparison strategy for several situations, *Psychological Bulletin* **115**, 153–159.
- [38] Lix, L.M. & Keselman, H.J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality, *Psychological Bulletin* **117**, 547–560.
- [39] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures, *Psychological Bulletin* **105**, 156–166.
- [40] Miller, R.G. (1981). *Simultaneous Statistical Inference*, 2nd Edition, Springer-Verlag, New York.
- [41] Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of standard deviation, *Biometrika* **31**, 20–30.
- [42] Norusis, M.J. (1997). *SPSS 9.0 Guide to Data Analysis*, Prentice Hall.
- [43] Olejnik, S. & Lee, J. (1990). Multiple comparison procedures when population variances differ, *Paper Presented at the Annual Meeting of the American Educational Research Association*, Boston.
- [44] Pearson, E.S. (1931). The analysis of variance in cases of nonnormal variation, *Biometrika* **23**, 114–133.
- [45] Petrionovich, L.F. & Hardyck, C.D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions, *Psychological Bulletin* **71**, 43–54.
- [46] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 663–665.
- [47] Rothman, K. (1990). No adjustments are needed for multiple comparisons, *Epidemiology* **1**, 43–46.
- [48] Ryan, T.A. (1959). Multiple comparisons in psychological research, *Psychological Bulletin* **56**, 26–47.
- [49] Ryan, T.A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics, *Psychological Bulletin* **57**, 318–328.
- [50] Ryan, T.A. (1962). The experiment as the unit for computing rates of error, *Psychological Bulletin* **59**, 305.
- [51] Ryan, T.A. (1980). Comment on “Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic”, *Psychological Bulletin* **88**, 354–355.
- [52] SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 7*, SAS Institute, Cary.
- [53] Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**, 110–114.
- [54] Saville, D.J. (1990). Multiple comparison procedures: The practical solution, *The American Statistician* **44**, 174–180.
- [55] Scheffé, H. (1959). *The Analysis of Variance*, Wiley.
- [56] Seaman, M.A., Levin, J.R. & Serlin, R.C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures, *Psychological Bulletin* **110**, 577–586.
- [57] Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures, *Journal of the American Statistical Association* **81**, 826–831.
- [58] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [59] Toothaker, L.E. (1991). *Multiple Comparisons for Researchers*, Sage Publications, Newbury Park.
- [60] Tukey, J.W. (1953). *The Problem of Multiple Comparisons*, Princeton University, Department of Statistics, Unpublished manuscript.
- [61] Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann, eds, *Contributions to Probability and Statistics*, Stanford University Press, Stanford.
- [62] Welch, B.L. (1938). The significance of the difference between two means when population variances are unequal, *Biometrika* **38**, 330–336.
- [63] Welch, B.L. (1951). On the comparison of several mean values: An alternative approach, *Biometrika* **38**, 330–336.
- [64] Welsch, R.E. (1977). Stepwise multiple comparison procedures, *Journal of the American Statistical Association* **72**, 566–575.
- [65] Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. & Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests*, SAS Institute, Cary.
- [66] Westfall, P.H. & Young, S.S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, Wiley, New York.
- [67] Wilcox, R.R. (1988). A new alternative to the ANOVA F and new results on James' second order method, *British Journal of Mathematical and Statistical Psychology* **41**, 109–117.
- [68] Wilcox, R.R. (1990). Comparing the means of two independent groups, *Biometrics Journal* **32**, 771–780.
- [69] Wilcox, R.R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies, *British Journal of Mathematical and Statistical Psychology* **48**, 99–114.
- [70] Wilcox, R.R. (1997). Three multiple comparison procedures for trimmed means, *Biometrical Journal* **37**, 643–656.
- [71] Wilcox, R.R., Charlin, V.L. & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics, *Communications in Statistics: Simulation and Computation* **15**, 933–943.
- [72] Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons, *Psychological Bulletin* **59**, 296–300.
- [73] Yuen, K.K. & Dixon, W.J. (1973). The approximate behavior of the two sample trimmed t, *Biometrika* **60**, 369–374.

H.J. KESELMAN, BURT HOLLAND AND
ROBERT A. CRIBBIE

Multiple Comparison Tests: Nonparametric and Resampling Approaches

H.J. KESELMAN AND RAND R. WILCOX

Volume 3, pp. 1325–1331

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Comparison Tests: Nonparametric and Resampling Approaches

Introduction

An underlying assumption of classical **multiple comparison procedures (MCPs)** is that the populations from which the data are sampled are normally distributed. Additional assumptions are that the population variances are equal (the homogeneity of variances assumption) and that the errors or observations are independent from one another (the independence of observations assumption). Although it may be convenient (both practically and statistically) for researchers to assume that their samples are obtained from normally distributed populations, this assumption may rarely be accurate [21, 31]. Researchers falsely assuming normally distributed data risk obtaining biased tests and relatively high Type II error rates for many patterns of nonnormality, especially when other assumptions are also not satisfied (e.g., variance homogeneity) (See [31]). Inaccurate **confidence intervals** occur as well.

The assumptions associated with the classical test statistics are typically associated with the following mathematical model that describes the sources that contribute to the magnitude of the dependent scores. Specifically, a mathematical model that can be adopted when examining pairwise and/or complex comparisons of means in a one-way completely randomized design is:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad (1)$$

where Y_{ij} is the score of the i th subject ($i = 1, \dots, n_j$) in the j th ($j = 1, \dots, J$) group, $\sum_j n_j = N$, μ_j is the j th group mean, and ϵ_{ij} is the random error for the i th subject in the j th group. As indicated, in the typical application of the model, it is assumed that the ϵ_{ij} s are normally and independently distributed and that the treatment group variances (σ_j^2 s) are equal. Relevant sample estimates include

$$\hat{\mu}_j = \bar{Y}_j = \sum_{i=1}^{n_j} \frac{Y_{ij}}{n_j} \text{ and } \hat{\sigma}^2 = \text{MSE}$$

$$= \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{\sum_{j=1}^J (n_j - 1)}. \quad (2)$$

Pairwise Comparisons

A confidence interval, assuming equal sample sizes n , for convenience, for a pairwise difference $\mu_j - \mu_{j'}$ has the form

$$\bar{Y}_j - \bar{Y}_{j'} \pm c_\alpha \hat{\sigma} \sqrt{\frac{2}{n_j}},$$

where c_α is selected such that the overall rate of Type I error (the probability of making at least one Type I error in the set of say m tests), that is, the familywise rate of Type I error (FWE) $= \alpha$. In the case of all possible pairwise comparisons, one needs a c_α such that the simultaneous probability coverage achieves a specified level. That is, for all $j \neq j'$, c_α must satisfy

$$P \left(\bar{Y}_j - \bar{Y}_{j'} - c_\alpha \hat{\sigma} \sqrt{\frac{2}{n_j}} \leq \mu_j - \mu_{j'} \leq \bar{Y}_j - \bar{Y}_{j'} + c_\alpha \hat{\sigma} \sqrt{\frac{2}{n_j}} \right) = 1 - \alpha. \quad (3)$$

Resampling Methods

Researchers can use both simultaneous and stepwise MCPs for pairwise comparisons of means with methods that do not assume normally distributed data. (Simultaneous MCPs use one critical value to assess statistical significance while stepwise procedures use a succession of critical value to assess statistical significance.) In particular, users can use either **permutation** or **bootstrap** methods to compute all possible pairwise comparisons, leading to hypothesis tests of such comparisons.

Pairwise comparisons of groups can be obtained through permutation of the data with the program provided by Westfall et al. [28, pp. 233–234]. Permutation tests do not require that the data be normally distributed. Instead of resampling with replacement from a pooled sample of residuals, permutation tests take the observed data ($Y_{11}, \dots, Y_{n11}, \dots, Y_{1J}, \dots, Y_{nJJ}$)

2 Multiple Comparison Tests

and randomly redistributes them to the treatment groups, and summary statistics (i.e., means or ranks) are then computed on the randomly redistributed data. The original outcomes (all possible pairwise differences from the original sample means) are then compared to the randomly generated values (e.g., all possible pairwise differences in the permutation samples). See [22] and [26].

Permutation tests can be used with virtually any measure of location, but regardless of which measure of location is used, they are designed to test the hypothesis that groups have identical distributions (e.g., [23]). If, for example, a permutation test based on means is used, it is not robust (*see Robust Testing Procedures*) if the goal is to make inferences about means (e.g., [1]).

When users adopt this approach to combat the effects of nonnormality, they should also heed the cautionary note provided by Westfall et al. [28, p. 234], namely, the procedure may not control the FWE when the data are heterogeneous, particularly when group sizes are unequal. Thus, we will introduce another approach, pairwise comparisons based on robust estimators and a heteroscedastic statistic, an approach that has been demonstrated to generally control the FWE when data are nonnormal and heterogeneous even when group sizes are unequal.

Prior to introducing bootstrapping with robust estimators it is important to note for completeness that researchers also can adopt nonparametric methods (*see Distribution-free Inference, an Overview*) to examine pairwise and/or complex contrasts among treatment group means (see e.g., [8] and [11]). However, one should remember that this approach is only equivalent to the classical approach of comparing treatment group means (comparing the same full and restricted models) when the distributions that are being compared are equivalent except for possible differences in location (i.e., a shift in location). That is, the classical and nonparametric approaches test the same hypothesis when the assumptions of the shift model hold; otherwise, the nonparametric approach is not testing merely for a shift in location parameters of the J groups (see [20], [4]).

Generally, conventional nonparametric tests are not aimed at making inferences about means or some measure of location. For example, the **Wilcoxon–Mann–Whitney test** is based on an estimate of $p(X < Y)$, the probability that an observation from the first group is less

than an observation from the second (e.g., see [3]). If restrictive assumptions are made about the distributions being compared, conventional nonparametric tests have implications about measures of location [6], but there are general conditions where a more accurate description is that they test the hypothesis of *identical distributions*. Interesting exceptions are given by Brunner, Domhof, and Langer [2] and Cliff [3]. For those researchers who believe nonparametric methods are appropriate (e.g., Kruskal–Wallis), we refer the reader to [8], [11], [20], or [25].

Robust Estimation

Bootstrapping methods provide an estimate of the distribution of the test statistic yielding P values that are not based on a theoretically presumed distribution (see e.g., [19]; [29]). An example SAS [22] program for all possible pairwise comparisons (of least squares means) is given by Westfall et al. [28, p. 229].

Westfall and Young's [29] results suggest that Type I error control could be improved further by combining a bootstrap method with one based on trimmed means. When researchers feel that they are dealing with populations that are nonnormal in form and thus subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters, then procedures, based on robust estimators, say trimmed means, should be adopted. Wilcox et al. [30] provide empirical support for the use of robust estimators and test statistics with bootstrap-determined critical values in one-way independent groups designs. This benefit has also been demonstrated in correlated groups designs (see [14]; [15]).

Accordingly, researchers can apply robust estimates of central tendency and variability to a heteroscedastic test statistic (*see Heteroscedasticity and Complex Variation*) (e.g., Welch's test [27]; also see [16]). When trimmed means are being compared the multiple comparison null hypothesis pertains to the equality of population trimmed means, that is, $H_0: \psi = \mu_{tj} - \mu_{tj'} = 0 (j \neq j')$. Although the null hypothesis stipulates that the population trimmed means are equal, we believe this is a reasonable hypothesis to examine since trimmed means, as opposed to the usual (least squares) means, provide better estimates of the typical individual in distributions that either contain **outliers** or are skewed. That

is, when distributions are skewed, trimmed means do not estimate μ but rather some value (i.e., μ_t) that is typically *closer* to the bulk of the observations. (Another way of conceptualizing the unknown parameter μ_t is that it is simply the population counterpart of $\hat{\mu}_t$ (see [12] and [9]). And lastly, as Zhou, Gao, and Hui [34] point out, distributions are typically skewed.

Thus, with robust estimation, the trimmed group means ($\hat{\mu}_{tj}$ s) replace the least squares group means ($\hat{\mu}_j$ s), the Winsorized group variances estimators (see **Winsorized Robust Measures**) ($\hat{\sigma}_{wj}^2$ s) replace the least squares variances ($\hat{\sigma}_j^2$ s), and h_j replaces n_j and accordingly one computes the robust version of a heteroscedastic test statistic (see [33], [32]). Definitions of trimmed means, Winsorized variances and the standard error of a trimmed mean can be found in [19] or [30, 31]. To test $H_0: \mu_{t1} - \mu_{t2} = 0 \equiv \mu_{t1} = \mu_{t2}$ (equality of population trimmed means), let $d_j = (n_j - 1)\hat{\sigma}_{wj}^2/h_j(h_j - 1)$, where $\hat{\sigma}_{wj}^2$ is the gamma-Winsorized variance and h_j is the effective sample size, that is, the size after trimming ($j = 1, 2$). Yuen's [33] test is

$$t_Y = \frac{\hat{\mu}_{t1} - \hat{\mu}_{t2}}{\sqrt{d_1 + d_2}}, \quad (4)$$

where $\hat{\mu}_{tj}$ is the γ -trimmed mean for the j th group and the estimated degrees of freedom are

$$v_Y = \frac{(d_1 + d_2)^2}{d_1^2/(h_1 - 1) + d_2^2/(h_2 - 1)}. \quad (5)$$

Bootstrapping

Following Westfall and Young [29] and as enumerated by Wilcox [30, 31], let $C_{ij} = Y_{ij} - \hat{\mu}_{tj}$; thus, the C_{ij} values are the empirical distribution of the j th group, centered so that the observed trimmed mean is zero. That is, the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each j , obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y_1^*, \dots, Y_{n_j}^*$. Let t_Y^* be the value of the test statistic based on the bootstrap sample. To control the FWE for a set of contrasts, the following approach can be used. Set $t_m^* = \max_j t_Y^*$, the maximum being taken over all $j \neq j'$. Repeat this process B times yielding

$t_{m1}^*, \dots, t_{mB}^*$. Let $t_{m(1)}^* \leq \dots \leq t_{m(B)}^*$ be the t_{mb}^* values written in ascending order, and let $q = (1 - \alpha)B$, rounded to the nearest integer. Then a test of a null hypothesis is obtained by comparing t_Y to $t_{m(q)}^*$ (i.e., whether $t_Y \geq t_{m(q)}^*$), where q is determined so that the FWE is approximately α . See [19, pp. 404–407], [29], or [32, 437–443].

Keselman, Wilcox, and Lix [16] present a SAS/IML [24] program which can be used to apply bootstrapping methods with robust estimators to obtain numerical results. The program can also be obtained from the first author's website at <http://www.umanitoba.ca/faculties/arts/psychology/>. This program is an extension of the program found in Lix and Keselman [18]. Tests of individual contrasts or families of contrasts may be performed (in addition omnibus main effects or interaction effects may be performed). The program can be applied in a variety of research designs. See [19].

Complex Comparisons

To introduce some methods that can be adopted when investigating complex comparisons among treatment group means we first provide some definitions. Specifically, we let

$$\psi = c_1\mu_1 + c_2\mu_2 + \dots + c_J\mu_J, \quad (6)$$

where the coefficients (c_j s) defining the contrast sum to one (i.e., $\sum_{j=1}^J c_j = 0$) represent the population contrast that we are interested in subjecting to a test of significance. To test $H_0: \psi = 0$, estimate ψ with

$$\hat{\psi} = c_1\bar{Y}_1 + c_2\bar{Y}_2 + \dots + c_J\bar{Y}_J. \quad (7)$$

The usual homoscedastic statistic is

$$t_{\hat{\psi}} = \frac{\hat{\psi}}{\sqrt{\text{MSE} \sum_{j=1}^J c_j^2/n_j}}. \quad (8)$$

A very popular simultaneous method for evaluating a set of m complex comparisons is to compare the P values associated with the $t_{\hat{\psi}}$ statistics to the Dunn–Bonferroni critical value α/m (see [17, p. 829] for the table of critical values).

One may also adopt stepwise MCPs when examining a set of m complex comparisons. These stepwise methods should result in greater sensitivity to

4 Multiple Comparison Tests

detect effects than the Dunn–Bonferroni [5] method. Many others have devised stepwise Bonferroni-type MCPs, for example, [10], [7], and so on. All of these procedures provide FWE control; the minor differences between them can result in small differences in power to detect effects. Thus, we recommend the Hochberg [7] sequentially acceptive step-up Bonferroni procedure because it is simple to understand and implement.

Hochberg’s [7] Sequentially Acceptive Step-up Bonferroni Procedure

In this procedure, the P values corresponding to the m statistics (e.g., $t_{\hat{\psi}_i}$) for testing the hypotheses $H_1, \dots, H_m$ are ordered from smallest to largest. Then, for any $i = m, m-1, \dots, 1$, if $p_i \leq \alpha/(m-i+1)$, the Hochberg procedure rejects all $H_{i'}$ ($i' \leq i$). According to this procedure, therefore, one begins by assessing the largest P value, p_m . If $p_m \leq \alpha$, all hypotheses are rejected. If $p_m > \alpha$, then H_m is accepted and one proceeds to compare $p_{(m-1)}$ to $\alpha/2$. If $p_{(m-1)} \leq \alpha/2$, then all H_i ($i = m-1, \dots, 1$) are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha/3$, and so on.

The Dunn–Bonferroni [5] and Hochberg [7] procedures can be adopted to robust estimation and testing by adopting trimmed means and Winsorized variances, instead of the usual least squares estimators. That is, to circumvent the biasing effects of non-normality and variance heterogeneity researchers can adopt a heteroscedastic Welch-type statistic and its accompanying modified degrees of freedom, applying robust estimators. For example, the heteroscedastic test statistic for the Dunn–Bonferroni [5] and Hochberg [7] procedures would be

$$t_{\hat{\psi}} = \frac{\hat{\psi}}{\sqrt{\sum_{j=1}^J c_j^2 s_j^2 / n_j}}, \quad (9)$$

where the statistic is approximately distributed as a Student t variable with

$$df_W = \frac{\left(\sum_{j=1}^J c_j^2 s_j^2 / n_j \right)^2}{\sum_{j=1}^J [(c_j^2 s_j^2 / n_j)^2 / (n_j - 1)]}. \quad (10)$$

Accordingly, as we specified previously, one replaces the least squares means with trimmed means and least squares variances with variances based on Winsorized sums of squares. That is,

$$t_{\hat{\psi}_i} = \frac{\hat{\psi}_i}{\sqrt{\sum_{j=1}^J c_j^2 d_j}}, \quad (11)$$

where the sample comparison of trimmed means equals

$$\hat{\psi}_i = c_1 \bar{Y}_{i1} + c_2 \bar{Y}_{i2} + \dots + c_J \bar{Y}_{iJ} \quad (12)$$

and error degrees of freedom is given by

$$df_{W_i} = \frac{\left(\sum_{j=1}^J c_j^2 d_j \right)^2}{\sum_{j=1}^J [(c_j^2 d_j)^2 / (h_j - 1)]}. \quad (13)$$

A bootstrap version of Hochberg’s [7] sequentially acceptive step-up Bonferroni procedure can be obtained in the following manner. Corresponding to the ordered P values are the $|t_{\hat{\psi}_i}|$ statistics. These pairwise statistics can be rank ordered according to their size and thus p_m (the largest P value) will correspond to the smallest $|t_{\hat{\psi}_i}|$ statistic. Thus, the smallest $t_{\hat{\psi}_i}$ (or largest P value) is bootstrapped. That is, as Westfall and Young [29, p. 47] maintain ‘The resampled P values... are computed using the same calculations which produced the original P values... from the original... data.’ Accordingly, let $|t_{\hat{\psi}_{i(1)}}^*| \leq \dots \leq |t_{\hat{\psi}_{i(B)}}^*|$ be the $|t_{\hat{\psi}_{i(b)}}^*|$ values for this smallest comparison written in ascending order, and let $m_r = [(1 - \alpha)B]$. Then statistical significance is determined by comparing

$$|t_{\hat{\psi}_i}| \geq t_{\hat{\psi}_{i(m_r)}}^*. \quad (14)$$

If this statistic fails to reach significance, then the next smallest $|t_{\hat{\psi}_i}|$ statistic is bootstrapped and compared to the $m_r = [(1 - \alpha/2)B]$ quantile. If necessary, the procedure continues with the next smallest $|t_{\hat{\psi}_i}|$. Other approaches for FWE control with resampling techniques are enumerated by Lunneborg [19, pp. 404–407] and Westfall and Young [29].

It is important to note that, although we are unaware of any empirical investigations that have examined tests of complex contrasts with robust estimators, based on empirical investigations related to pairwise comparisons, we believe these methods would provide approximate FWE control under conditions of nonnormality and variance heterogeneity and should possess more power to detect group differences than procedures based on least squares estimators.

Numerical Example

We use an example data set presented in Keselman et al. [16] to illustrate the procedures enumerated in this paper. Keselman et al. modified a data set given by Karayanidis, Andrews, Ward, and Michie [13] where the authors compared the performance of three age groups (Young, Middle, Old) on auditory selective attention processes. The dependent reaction time (in milliseconds) scores are reported in Table 1.

We used the SAS/IML program demonstrated in [16, pp. 589–590] to obtain numerical results for tests of pairwise comparisons and a SAS/IML program that we wrote to obtain numerical results for tests of complex contrasts. Estimates of group trimmed means and standard errors are presented in Table 2.

Table 1 Example data set

Young	Middle	Old
518.29	335.59	558.95
548.42	353.54	538.56
524.10	493.08	586.39
666.63	469.01	530.23
488.84	338.43	629.22
676.40	499.10	691.84
482.43	404.27	557.24
531.18	494.31	528.50
504.62	487.30	565.43
609.53	485.85	536.03
584.68	886.41	594.69
609.09	437.50	645.69
495.15		558.61
502.69		519.01
484.36		538.83
519.10		
572.10		
524.12		
495.24		

Table 2 Descriptive Statistics [Trimmed Means and (Standard Errors)]

Young	Middle	Old
532.98 (15.27)	453.11 (26.86)	559.41 (11.00)

Three pairwise comparisons were computed: (1) Young versus Middle, (2) Young versus Old, and (3) Middle versus Old. The values of t_Y and v_Y for the three comparisons are (1) 6.68 and 11.55, (2) 1.97 and 19.72, and (3) 13.41 and 9.31; for .05 FWE control the critical value is 12.56. Accordingly, based on the methodology enumerated in this paper and the critical value obtained with the program discussed by Keselman et al. [16] only the third comparison is judged to be statistically significant.

We as well computed three complex comparisons: (1) Young versus the average of Middle and Old, (2) Middle versus the average of Young and Old, and (3) Old versus the average Young and Middle. Using the bootstrapped version of Hochberg's [7] step-up Bonferroni MCP, none of the contrasts are statistically significant. That is, the ordered $t_{\hat{\psi}_i}$ were 1.27 (comparison 2), 3.27 (comparison 1) and 3.50 (comparison 3). The corresponding critical $t_{\hat{\psi}_{r(m_r)}}^*$ were 2.20, 4.14, and 4.15, respectively.

References

- [1] Boik, R.J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous, *British Journal Mathematical and Statistical Psychology* **40**, 26–42.
- [2] Brunner, E., Domhof, S. & Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*, Wiley, New York.
- [3] Cliff, N. (1996). *Ordinal Methods for Behavioral Data Analysis*, Lawrence Erlbaum, Mahwah.
- [4] Delaney, H.D. & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples, *Psychological Methods* **7**, 485–503.
- [5] Dunn, O.J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association* **56**, 52–64.
- [6] Hettmansperger, T.P. & McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*, Arnold, London.
- [7] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **75**, 800–802.
- [8] Hochberg, Y. & Tamhane, A.C. (1987). *Multiple Comparison Procedures*, Wiley, New York.

6 Multiple Comparison Tests

- [9] Hogg, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory, *Journal of the American Statistical Association* **69**, 909–927.
- [10] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.
- [11] Hsu, J.C. (1996). *Multiple Comparisons Theory and Methods*, Chapman & Hall, New York.
- [12] Huber, P.J. (1972). Robust statistics: a review, *Annals of Mathematical Statistics* **43**, 1041–1067.
- [13] Karayanidis, F., Andrews, S., Ward, P.B. & Michie, P.T. (1995). ERP indices of auditory selective attention in aging and Parkinson's disease, *Psychophysiology* **32**, 335–350.
- [14] Keselman, H.J., Algina, J., Wilcox, R.R. & Kowalchuk, R.K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: revisiting the robustness of the Welch-James test again, *Educational and Psychological Measurement* **60**, 925–938.
- [15] Keselman, H.J., Kowalchuk, R.K., Algina, J., Lix, L.M. & Wilcox, R.R. (2000). Testing treatment effects in repeated measures designs: trimmed means and bootstrapping, *British Journal of Mathematical and Statistical Psychology* **53**, 175–191.
- [16] Keselman, H.J., Wilcox, R.R. & Lix, L.M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs, *Psychophysiology* **40**, 586–596.
- [17] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks/Cole Publishing Company, Toronto.
- [18] Lix, L.M. & Keselman, H.J. (1995). Approximate degrees of freedom tests: a unified perspective on testing for mean equality, *Psychological Bulletin* **117**, 547–560.
- [19] Lunneborg, C.E. (2000). *Data Analysis by Resampling*, Duxbury, Pacific Grove.
- [20] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [21] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures, *Psychological Bulletin* **105**, 156–166.
- [22] Opdyke, J.D. (2003). Fast permutation tests that maximize power under conventional Monte Carlo sampling for pairwise and multiple comparisons, *Journal of Modern Applied Statistical Methods* **2**(1), 27–49.
- [23] Pesarin, F. (2001). *Multivariate Permutation Tests*, Wiley, New York.
- [24] SAS Institute. (1999). SAS/STAT user's guide, Version 7, SAS Institute, Cary.
- [25] Sprent, P. (1993). *Applied Nonparametric Statistical Methods*, 2nd Edition, Chapman & Hall, London.
- [26] Troendle, J.F. (1996). A permutational step-up method of testing multiple outcomes, *Biometrics* **52**, 846–859.
- [27] Welch, B.L. (1938). The significance of the difference between two means when population variances are unequal, *Biometrika* **38**, 330–336.
- [28] Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D. & Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests*, SAS Institute, Cary.
- [29] Westfall, P.H. & Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P value Adjustment*, Wiley, New York.
- [30] Wilcox, R.R. (1990). Comparing the means of two independent groups, *Biometrics Journal* **32**, 771–780.
- [31] Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego.
- [32] Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*, Academic Press, New York.
- [33] Yuen, K.K. (1974). The two-sample trimmed t for unequal population variances, *Biometrika* **61**, 165–170.
- [34] Zhou, X., Gao, S. & Hui, S.L. (1997). Methods for comparing the means of two independent log-normal samples, *Biometrics* **53**, 1129–1135.

Further Reading

- Keselman, H.J., Lix, L.M. & Kowalchuk, R.K. (1998). Multiple comparison procedures for trimmed means, *Psychological Methods* **3**, 123–141.
- Keselman, H.J., Othman, A.R., Wilcox, R.R. & Fradette, K. (2004). The new and improved two-sample t test, *Psychological Science* **15**, 47–51.

H.J. KESELMAN AND RAND R. WILCOX

Multiple Imputation

BRIAN S. EVERITT

Volume 3, pp. 1331–1332

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Imputation

A method by which **missing values** in a data set are replaced by more than one, usually between 3 and 10, simulated versions. Each of the simulated complete datasets is then analyzed by the method relevant to the investigation at hand, and the results combined to produce estimates, standard errors, and confidence intervals that incorporate missing data uncertainty. Introducing appropriate random errors into the imputation process makes it possible to get approximately unbiased estimates of all parameters, although the data must be missing at random (*see **Dropouts in Longitudinal Data; Dropouts in Longitudinal Studies: Methods of Analysis***) for this to

be the case. The multiple imputations themselves are created by a Bayesian approach (*see **Bayesian Statistics** and **Markov Chain Monte Carlo and Bayesian Statistics***), which requires specification of a parametric model for the complete data and, if necessary, a model for the mechanism by which data become missing. A comprehensive account of multiple imputation and details of associated software are given in Schafer [1].

Reference

- [1] Schafer, J. (1997). *The Analysis of Incomplete Multivariate Data*, CRC/Chapman & Hall, Boca Raton.

BRIAN S. EVERITT

Multiple Informants

KIMBERLY J. SAUDINO

Volume 3, pp. 1332–1333

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Informants

Most behavioral researchers agree that it is desirable to obtain information about subjects from multiple sources or informants. For example, the importance of using multiple informants for the assessment of behavior problems in children has long been emphasized in the phenotypic literature. Correlations between different informants' ratings of child behavior problems are typically low – in the range of 0.60 between informants with similar roles (e.g., parent–parent); 0.30 between informants who have different roles (e.g., parent–teacher); and 0.20 between self-reports and other informant ratings [1]. These low correlations are typically interpreted as indicating that different raters provide different information about behavior problems because they view the child in different contexts or situations; however, error and rater bias can also contribute to low agreement between informants. Whatever the reasons for disagreement among informants, the use of multiple informants allows researchers to gain a fuller understanding of the behavior under study.

In quantitative genetic analyses, relying on a single informant may not paint a complete picture of the etiology of the behavior of interest. Using the above example, parent and teacher ratings assess behaviors in very different contexts; consequently, the genetic and environmental factors that influence behaviors at home might differ from those that influence the same behaviors in the classroom. There may also be some question as to whether parents and teachers actually are assessing the *same* behaviors. In addition, informants' response tendencies, standards, or behavioral expectations may affect their ratings – **rater biases** that cannot be detected without information from multiple informants. Analysis of data from multiple informants can inform about the extent to which different informants' behavioral ratings are influenced by the same genetic and environmental factors, and can explain why there is agreement and/or disagreement amongst informants.

Three classes of quantitative genetic models have been applied to data from multiple informants: biometric models, psychometric models, and bias models [5]. Each makes explicit assumptions about the reasons for agreement and disagreement among informants. Biometric models such as the **Independent Pathway** model [3] posit that genes and environments

contribute to **covariance** between informants through separate genetic and environmental pathways. This model decomposes the genetic, **shared environmental**, and **nonshared environmental** variances of multiple informants' ratings (e.g., parent, teacher, and child) into genetic and environmental effects that are common to all informants, and genetic and environmental effects that are specific to each informant. Under this model, covariance between informants can arise due to different factors. That is, although all informants' ratings may be intercorrelated, the correlation between any two informants' ratings may be due to different factors (e.g., the correlation between parent and teacher could have different sources than the correlation between parent and child). Like the **Cholesky Decomposition** (also a biometric model), the Independent Pathway Model can be considered to be 'agnostic' in that it does not specify that the different informants are assessing the same phenotype, rather, it just allows that the phenotypes being assessed by each informant are correlated [2].

The Psychometric or **Common Pathway model** [3, 4] is more restrictive, positing that genes and environments influence covariation between raters through a *single* common pathway. This model suggests that correlations between informants arise because they are assessing a common phenotype. This common phenotype is then influenced by genetic and/or environmental influences. As is the case for the Independent Pathway model, this model also allows genetic and environmental effects specific to each informant. Under this model, genetic and/or environmental sources of covariance are the same across all informants. Informants' ratings agree because they tap the same latent phenotype (i.e., they are assessing the same behaviors). Informants' ratings differ because, to some extent, they also assess different phenotypes due to the fact that each informant contributes different but valid information about the target's behavior.

As with the Psychometric model, the **Rater Bias model** [2, 6] assumes that informants agree because they are assessing the same latent phenotype that is influenced by genetic and environmental factors; however, this model also assumes that disagreement between informants is due to rater bias and unreliability. That is, this model does not include informant-specific genetic or environmental influences – anything that is not reliable trait variance (i.e., the common phenotype) is bias or error, both of which are estimated in the model. Rater biases refer to the

2 Multiple Informants

informant's tendency to consistently overestimate or underestimate the behavior of targets [2]. Because bias is conceptualized as consistency *within* an informant *across* targets, the Rater Bias model requires that the same informant assess both members of a twin or sibling pair. Other multiple informant models do not require this.

All of the above models allow the estimation of **genetic and environmental correlations** (i.e., degree of genetic and/or environmental overlap) between informants, and the extent to which genetic and environmental factors contribute to the phenotypic correlations between informants (i.e., bivariate heritability and environmentality). Comparisons of the relative fits of the different models make it possible to get some understanding of differential informant effects. By comparing the Biometric and Psychometric models, it is possible to determine whether it is reasonable to assume that different informants are assessing the same phenotypes. That is, if a Biometric model provides the best fit to the data, then the possibility that informants are assessing different, albeit correlated, phenotypes must be considered. Similarly, comparisons between the Psychometric and Rater Bias models inform about the presence of valid informant differences versus rater biases.

References

- [1] Achenbach, T.M., McConaughy, S.H. & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity, *Psychological Bulletin* **101**, 213–232.
- [2] Hewitt, J.K., Silberg, J.L., Neale, M.C., Eaves, L.J. & Erikson, M. (1992). The analysis of parental ratings of children's behavior using LISREL, *Behavior Genetics* **22**, 292–317.
- [3] Kendler, K.S., Heath, A.C., Martin, N.G. & Eaves, L.J. (1987). Symptoms of anxiety and symptoms of depression: same genes, different environments? *Archives General Psychiatry* **44**, 451–457.
- [4] McArdle, J.J. & Goldsmith, H.H. (1990). Alternative common-factor models for multivariate biometric analyses, *Behavior Genetics* **20**, 569–608.
- [5] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, Dordrecht.
- [6] Neale, M.C. & Stevenson, J. (1989). Rater bias in the EASI temperament survey: a twin study, *Journal of Personality and Social Psychology* **56**, 446–455.

KIMBERLY J. SAUDINO

Multiple Linear Regression

STEPHEN G. WEST AND LEONA S. AIKEN

Volume 3, pp. 1333–1338

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Linear Regression

Multiple regression addresses questions about the relationship between a set of independent variables (*IVs*) and a dependent variable (*DV*). It can be used to describe the relationship, to predict future scores on the *DV*, or to test specific hypotheses based on scientific theory or prior research. Multiple regression most often focuses on linear relationships between the *IVs* and the *DV*, but can be extended to examine other forms of relationships.

Multiple regression begins by writing an equation in which the *DV* is a weighted linear combination of the independent variables. In general, the regression equation may be written as $Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p + e$. Y is the *DV*, each of the X s is an independent variable, each of the b s is the corresponding regression coefficient (weight), and e is the error in prediction (residual) for each case. The linear combination excluding the residual, $b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$, is also known as the predicted value or $\hat{Y}$, the score we would expect on the *DV* based on the scores on the set of *IVs*.

To illustrate, we use data from 56 live births taken from [4]. The *IVs* were the *Age* of the mother in years, the *Term* of the pregnancy in weeks, and the *Sex* of the infant (0 = girls, 1 = boys). The *DV* is

the *Weight* of the infant in grams. Figure 1(a) is a **scatterplot** of the *Term*, *Weight* pair for each case. Our initial analysis (Model 1) predicts infant *Weight* from one *IV*, *Term*. The regression equation is written as $Weight = b_0 + b_1Term + e$. The results are shown in Table 1. $b_0 = -2490$ is the intercept, the predicted value of *Weight* when *Term* = 0. $b_1 = 149$ is the slope, the number of grams of increase in *Weight* for each 1-week increase in *Term*. Each of the regression coefficients is tested against a population value of 0 using the formula, $t = b_i/s_{b_i}$, with $df = n - p - 1$, where s_{b_i} is the estimate of the standard error of b_i . Here, $n = 56$ is the sample size and $p = 1$ is the number of predictor variables. We cannot reject the null hypothesis that the infant's *Weight* at 0 weeks (the moment of conception) is 0 g, although a term of 0 weeks for a live birth is impossible. Hence, this conclusion should be treated very cautiously. The test of b_1 indicates that there is a positive weight gain for each 1-week increase in *Term*; the best estimate is 149 g per week. The 95% confidence interval for the corresponding population regression coefficient b_1 is $b_1 \pm t_{0.975} s_{b_1} = 149 \pm (2)(38.8) = 71.4$ to 226.6. $R^2 = 0.21$ is the squared correlation between Y and $\hat{Y}$ or, alternatively, $.21$ is $SS_{\text{predicted}}/SS_{\text{total}}$, the proportion of variation in *Weight* accounted for by *Term*.

Model 2 predicts *Weight* from *Term* and mother's *Age*. Figure 1(b) portrays the relationship between *Weight* and *Age*. To improve the interpretability of the intercept [9], we mean center *Term*, $Term_c = Term -$

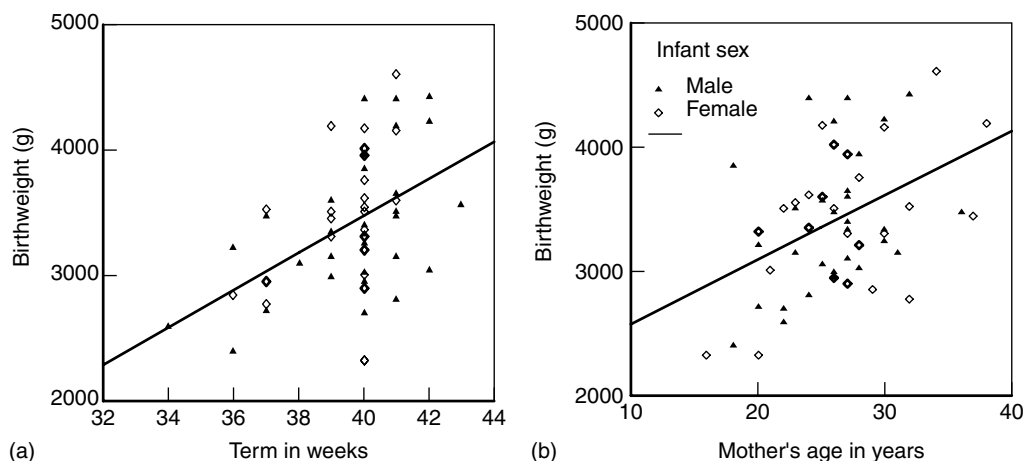


Figure 1 Scatterplots of raw data (a) Weight vs. Term (b) Weight vs. Age *Note:* x represents male; o represents female. Each point represents one observation. The best fitting straight line is superimposed in each scatterplot.

2 Multiple Linear Regression

Table 1 Model 1: regression of infant weight on term of pregnancy

Coefficient estimates					
Label	Estimate	Std. error	<i>t</i> -value	<i>P</i> value	
b_0 . Intercept	−2490	1537.0	−1.62	0.11	
b_1 . Term	149	38.8	3.84	0.0003	
R Squared:	0.2148				
Number of cases:					
Degrees of freedom:	54				
Summary Analysis of Variance Table					
Source	df	SS	MS	F	<i>P</i> value
Regression	1	3598397.	3598397.	14.77	0.0003
Residual	54	13155587.	243622.		
Total	55	16753984			

Table 2 Model 2: regression of infant weight on term of pregnancy and mother's age

Coefficient estimates					
Label	Estimate	Std. error	<i>t</i> -value	<i>P</i> value	
<i>b</i> ₀ . Intercept	3413	59.3	57.56	<0.0001	
<i>b</i> ₁ . Term-c	41	35.0	4.04	0.0002	
<i>b</i> ₂ . Age-c	48	13.0	3.72	0.0005	
R Squared:	0.3771				
Number of cases:	56				
Degrees of freedom:	53				
Summary Analysis of Variance Table					
Source	df	SS	MS	F	<i>P</i> value
Regression	2	6318298	3159149.	16.04	<0.0001
Residual	53	10435686	196900.		
Total	55	16753984			

$\text{Mean}(\text{Term})$, and Age , $\text{Age}_C = \text{Age} - \text{Mean}(\text{Age})$. In this sample, $\text{Mean}(\text{Term}) = 39.6$ weeks and $\text{Mean}(\text{Age}) = 26.3$ years. The results are presented in Table 2. The intercept b_0 now represents the predicted value of $\text{Weight} = 3413$ g when Term_C and Age_C both equal 0, at the mean Age of all mothers and mean Term length of all pregnancies in the sample. This value is equal to the mean birth weight in the sample. $b_1 = 141$ is the slope for Term , the gain in Weight for each 1-week increase in Term , holding the mother's Age constant. $b_2 = 48$ is the slope for Age , the gain in Weight for each 1-year increase in mother's Age , holding Term constant. Mean centering does not affect the value of the highest order regression

coefficients, here the slopes. The addition of Term to the equation results in an increase in R^2 from 0.21 to 0.38.

Model 3 illustrates the addition of a categorical IV , infant Sex , to the regression equation. For G categories, $G - 1$ code variables are needed to represent the categorical variable. Sex has two categories (female, male), so one code variable is needed. Several different schemes including dummy, effect, and contrast codes are available; the challenge is to select the scheme that provides the optimal interpretation for the research question [3]. Here we use a dummy code scheme in which 0 = female and 1 = male (see **Dummy Variables**). Table 3 presents

Table 3 Model 3: regression of infant weight on term, mother's age, and sex

Coefficient estimates								
Label	unstand. b_i	95% CI	stand. β_i	sr^2	pr^2	sb_i	t -value	P value
Intercept	3423	3246 to 3600	–	–	–	88.2	38.8	<0.0001
Age-c	48	21.6 to 74.8	.40	.159	.203	13.2	3.6	0.0006
Term-c	141	70.5 to 212.5	.44	.192	.235	35.4	4.0	0.0002
Sex	18	–261 to +225	–.02	.000	.000	121.0	–0.1	0.8822
R Squared:	0.3774							
Number of cases:	56							
Degrees of freedom:	52							

Summary Analysis of Variance Table					
Source	df	SS	MS	F	P value
Regression	3	6322743	2107581	10.51	<0.0001
Residual	52	10431241	200601		

Diagnostic Statistics	Largest Absolute Value
Leverage (h_{ii})	0.24
Distance (t_i)	2.49
Global Influence (DFFITS _{<i>i</i>})	0.66
Specific Influence	
DFBETAS _{<i>ij</i>} – Term	–0.46
DFBETAS _{<i>ij</i>} – Age	0.56
DFBETAS _{<i>ij</i>} – Sex	0.36

the results. Again, $Term_C$ and Age_C are mean centered.

The intercept b_0 now represents the predicted value of *Weight* when Age_C and $Term_C$ are 0 (their respective mean values) and $Sex = 0$ (female infant). $b_1 = 141$ represents the slope for $Term_C$, $b_2 = 48$ represents the slope for Age_C , and $b_3 = -18$ indicates that males are on average 18 g lighter holding Age and $Term$ constant. The t Tests indicate that the intercept, the slope for $Term$ and the slope for Age are greater than 0. However, we cannot reject the null hypothesis that the two sexes will have equal birth weights in the population, holding Age and $Term$ constant. The 95% confidence intervals indicate the plausible range for each of the regression coefficients in the population. Finally, the R^2 is still 0.38. The gain in prediction can be tested when any set of 1 or more predictors is added to the equation,

$$F = \frac{(SS_{\text{regression—full}} - SS_{\text{regression—reduced}})/df_{\text{set}}}{SS_{\text{residual—full}}/df_{\text{residual—full}}} \quad (1)$$

Comparing Model (3), the full model with all 3 predictors, with Model (1), the reduced model that only included $Term$ as a predictor, we find

$$F = \frac{(6322743 - 3598397)/2}{10431241/52} = 22.5;$$

$$df = 2, 54; \quad p < 0.001 \quad (2)$$

The necessary Sums of Squares (SS) are given in Tables 1 and 3 and the df_1 is the difference in the number of predictors between the full and reduced models, here $3 - 1 = 2$, and the df_2 is identical to that of the full model. This test indicates that adding Age and Sex to the model leads to a statistically significant increase in prediction.

Table 3 also presents the information about the relationship between each *IV* and the *DV* in three other metrics. The **standardized regression coefficient** indicates the number of standard deviations (*SD*) the *DV* changes for each 1 *SD* change in the *IV*, controlling for the other *IVs*. The squared semipartial (part) correlation indicates the unique proportion of the total variation in the *DV* accounted for by each

4 Multiple Linear Regression

IV. The squared partial correlation (*see Partial Correlation Coefficients*) indicates the unique proportion of the remaining variation in the DV accounted for by an IV after the contributions of each of the other IVS have been subtracted out. The squared partial correlation will be larger than the squared semipartial correlation unless the other IVS collectively do not account for any variation.

More complex forms of multiple regression models can also be specified. A quadratic relationship [1] can be expressed as $Y = b_0 + b_1X + b_2X^2 + e$. An interaction [1] in which Y depends on the combination of X_1 and X_2 can be expressed as $Y = b_0 + b_1X + b_2X_2 + b_3X_1X_2 + e$. In some cases such as when there is a very large range on the DV, Y may be transformed by replacing it by a function of Y , for example, $\log(Y)$. Similar transformations may also be performed on each IV [4]. Nonlinear models which follow a variety of mathematical functions can also be estimated [6, 8] (*see Nonlinear Models*). For example, the growth of learning over repeated trials often initially increases rapidly and then gets slower and slower until it approaches a maximum possible value (asymptote). This form can often be represented by an exponential function.

Multiple regression is adversely affected by violations of assumptions [3, 4, 6]. First, data may be clustered because individuals are measured in groups or repeated measurements are taken on the same individuals over time (nonindependence) (*see Linear*

Multilevel Models and Generalized Linear Mixed Models). Second, the variance of the residuals around the regression line may not be constant (heteroscedasticity). Third, the residuals may not have a normal distribution (nonnormality). Violations of these assumptions may require special procedures such as random coefficient models (3) for clustering and the use of alternative estimation procedures for non-constant variance to produce proper results, notably correct standard errors. Careful examination of plots of the residuals (see Figure 2) can diagnose problems in the regression model and lead to improvements in the model (e.g., addition of an omitted IV).

Another problem is **outliers**, extreme data points that are far from the mass of observations [3, 4, 6]. Outliers may represent observational errors or extremely unusual true cases (e.g., an 80-year old college student). Outliers can potentially profoundly affect the values of the regression coefficients, even reversing the sign in extreme cases. Outlier statistics are case statistics, with a set of different diagnostic statistics being computed for each case i in the data set. Leverage (h_{ii}) measures how extreme a case is in the set of IVS; it is a measure of the distance of case i from the centroid of the X s (the mean of the IVS). Distance measures the discrepancy between the observed Y and the predicted $\hat{Y}$ ($Y - \hat{Y}$) for each case, indicating how poorly the model fits the case in question. Because outliers can affect the slope of the

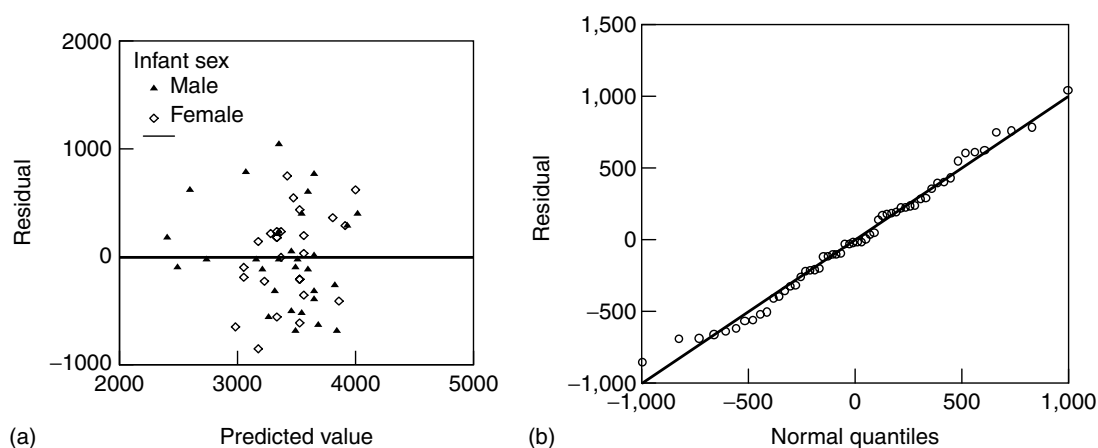


Figure 2 Model checking plots (a) Residuals vs. Predicted Values (b) Q-Q Plot of Residuals vs Normal Distribution
Note: In (a), residuals should be approximately evenly distributed around the 0-line if the residuals are homoscedastic. In (b), the residuals should approximately fit the superimposed straight line if they are normally distributed. The assumptions of homoscedasticity and normality of residuals appear to be met.

regression line, the externally Studentized residual t_i is used rather than the simple residual e as t_i provides a much clearer diagnosis. Conceptually, the regression equation is estimated with case i deleted from the data set. The values of the IVS for case i are substituted into this equation to calculate $\hat{Y}_{i(i)}$ for case i . t_i is the standardized difference between the observed value of Y_i and $\hat{Y}_{i(i)}$. Influence, which is the combination of high leverage and high distance, indicates the extent to which the results of the regression equation are affected by the outlier. $DFFITS_i$ is a standardized measure of global influence which describes the number of SD s by which $\hat{Y}$ would change if case i were deleted. $DFFITS_i = t_i \sqrt{(h_{ii}/1 - h_{ii})}$. (Cook's D_i is a closely related alternative measure of global influence in a different metric). Finally, $DFBETAS_{ij}$ is a standardized measure of influence which describes the number of SD s regression coefficient b_j would change if case i were deleted. Cases with extreme values on the outlier diagnostic statistics should be examined with suggested rule of thumb cutoff values for large sample sizes being $(2(p+1)/n)$ for h_{ii} , ± 3 or ± 4 for t_i , $\pm 2\sqrt{(p+1)/n}$ for $DFFITS_i$, and $\pm (2/\sqrt{n})$ for $DFBETAS_{ij}$. No extreme outliers were identified in our birth weight data set. Methods of addressing outliers include dropping them from the data set and limiting the conclusions that are reached and using robust regression procedures that down weight their influence (see **Influential Observations**).

In summary, multiple regression is a highly flexible method studying the relationship between a set of IVS and a DV. Different types of IVS may be studied and a variety of forms of the relationship between each IV and the DV may be represented. Multiple regression requires attention to its assumptions and to possible outliers to obtain optimal results and improve the model. A chapter length introduction [2], full length introductions

for behavioral scientists [3, 5, 7], and full length introductions for statisticians [6, 8, 10] to multiple regression are available. Numerous multiple regression programs exist, with SPSS Regression and SAS PROC Reg being most commonly used (see **Software for Statistical Analyses**). Cook and Weisberg [1] offer an excellent freeware program ARC <http://www.stat.umn.edu/arc/>.

References

- [1] Aiken, L.S. & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*, Sage Publications, Newbury Park.
- [2] Aiken, L.S. & West, S.G. (2003). Multiple linear regression, in *Handbook of Psychology, (Vol. 2): Research Methods in Psychology*, J.A. Schinka & W.F. Velicer, eds, Wiley, New York, pp. 483–532.
- [3] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.
- [4] Cook, R.D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, Wiley, New York.
- [5] Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, Thousand Oaks.
- [6] Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman W. (1996). *Applied Linear Regression Models*, 3rd Edition, Irwin, Chicago.
- [7] Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research*, 3rd Edition, Harcourt Brace, Forth Worth.
- [8] Ryan, T.P. (1996). *Modern Regression Methods*, Wiley, New York.
- [9] Wainer, H. (2000). The centercept: an estimable and meaningful regression parameter, *Psychological Science* **11**, 434–436.
- [10] Weisberg, S. (2005). *Applied Linear Regression*, 3rd Edition, Wiley, New York.

STEPHEN G. WEST AND LEONA S. AIKEN

Multiple Testing

RICHARD B. DARLINGTON

Volume 3, pp. 1338–1343

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multiple Testing

The multiple-testing problem arises when a person or team performs two or more significance tests, because the probability of finding at least one significant result exceeds the nominal probability α . Multiple-testing methods are designed to correct for this liberal bias. Typical cases of multiple tests are the multiple comparisons of cell means in **analysis of variance**, testing all the values in a matrix of correlations (see **Correlation and Covariance Matrices**), testing all the slopes in a **multiple linear regression**, testing each cell in a **contingency table** for the cell frequency's deviation from its expected value, and finding multiple tests of the same general question in a review of scientific literature. The multiple-comparisons problem has received far more attention than the others, and a separate article is devoted to that topic (see **A Priori v Post Hoc Testing**). We also cover that multiple-comparisons problem briefly here, but we focus more heavily on the more general problem.

If p_m denotes the most significant value among B significance levels, the problem is to find the probability that a value at least as small as p_m would have been found just by chance in the total set of B values. We shall denote this probability as PC , for 'corrected p_m '. We do not generally assume that the B trials are statistically independent; of the five examples in the opening paragraph, trials are independent only in the last one, in which the various significance levels are obtained from completely different experiments. When independence does exist, the exact formula for PC is well known: $(1 - p_m)$ is the probability of failing to find the p_m in question on any one trial, so $(1 - p_m)^B$ is the probability of failing on all B trials, so $1 - (1 - p_m)^B$ is the probability of finding at least that level on at least one trial. Thus $PC = 1 - (1 - p_m)^B$.

Here we review the most widely used methods for correcting for multiple tests and also consider more general questions that arise when dealing with multiple tests. We begin with a review of methods.

The Bonferroni Method

Ryan [5] and Dunn [3] were actually the first prominent advocates of the method that is now generally called the *Bonferroni* method, but nevertheless

here we shall call the method by its most widely used name. The Bonferroni method is noteworthy for its extreme flexibility; it can be applied in all of the cases mentioned in our opening paragraph, plus many more. It is also remarkably simple. If you have selected the most significant value p_m out of B significance levels that you computed, then you simply multiply these two values together, so $PC = B \times p_m$. The Bonferroni method has become increasingly practical in recent years because modern computer packages can usually give the exact values of very small p values; these are, of course, necessary for the Bonferroni method.

There is no requirement that the tests analyzed by the Bonferroni method be statistically similar. For instance, they may be any mixture of parametric and nonparametric tests. In theory, they could even be a mixture of one-tailed and two-tailed tests, but scientifically, the conclusions would usually be most meaningful if all tests were one or the other.

If tests are independent and the true PC is .05 or below, the Bonferroni method has only a slight conservative bias in comparison to the exact formula mentioned above. By trying the exact formula and the Bonferroni formula with various values of B , one can readily verify that when $B < 10^9$ and the exact formula yields $PC = .05$, the Bonferroni formula never yields PC over .0513. But when the exact PC is very high, the Bonferroni PC may be far higher still – even over 1.0. Thus, if the Bonferroni formula yields $PC > .05$ for independent tests, you should take the extra time to calculate the exact value.

Contrary to intuition, the Bonferroni method can be reasonably powerful even when B is very large. For instance, consider a Pearson correlation of .5 in a sample of 52 cases. We have $p < .000001$, which would yield PC below .05 even if B were 50 000.

Ryan [6] has shown that the Bonferroni method is never too liberal, even if the various tests in the analysis are nonindependent. The more positively correlated the tests in the analysis are, the more conservative the Bonferroni test is relative to some theoretically optimum test. One can think of the ordinary two-tailed test as a one-tailed test corrected by a Bonferroni factor of 2 to allow for the possibility that the investigator chose the direction *post hoc* to fit the data. There is no conservative bias in the two-tailed test because the two possible one-tailed tests are perfectly negatively correlated – they can never

2 Multiple Testing

both yield positive results. But there is no liberal bias even in this worst case.

The Bonferroni method easily handles a problem that arises in the area of multiple comparisons, which the classical methods for multiple comparisons cannot handle. If you have decided in advance that you are interested in only some subset of all the possible comparisons, and you test only those, then you can apply the Bonferroni method to correct for the exact number of tests you have performed.

There is also Bonferroni layering. If the most significant of k results is significant after a Bonferroni correction, then you can test the next most significant result using a Bonferroni factor of $(k - 1)$ because that is the most significant of the remaining $(k - 1)$ results. You can then continue, testing the third, fourth, and other most significant results using correction factors of $(k - 2)$, $(k - 3)$, and so on.

Multiple Comparisons

The opening paragraph gave five instances of the multiple-tests problem, and there are many others. Thus, the problem of multiple comparisons in analysis of variance is just one of many instances. But this one instance has attracted extraordinary attention from statisticians. Lengthy reviews of this work appear in works like [4] and [7], and the topic is covered in a separate article in this encyclopedia (*see A Priori v Post Hoc Testing*). Thus, we mention only a few highlights here.

In analysis of variance, a pairwise comparison is a test of the difference between two particular cell means. For simplicity, we shall emphasize one-way analyses of variance, though our discussion applies with little change to more complex designs. If there are k cells in a design, then the number of possible pairwise comparisons is $k(k - 1)/2$. Thus, one valid approach is to compute all possible pairwise values of t , test the most significant one using a Bonferroni correction factor of $k(k - 1)/2$, and then use Bonferroni layering to test the successively smaller values of t . This approach is very flexible: it does not require equal cell frequencies or the assumption of equal within-cell variances, and tests may be one-tailed, two-tailed, or even mixed. Virtually all alternative methods are less flexible but do gain a small amount of power relative to the method just mentioned.

Several of these methods use the *studentized range*, defined as $sr = (M_i - M_j)/\sqrt{(MSE/n_h)}$, where M_i and M_j are the two cell means in question, MSE is the mean squared error, and n_h is the **harmonic mean** of the two cell frequencies. This is essentially the t that one would calculate to test the difference between just those two cell means, with two differences: sr uses the pooled variance for the entire analysis rather than for just the two cells being tested and sr is multiplied by an extra factor of $\sqrt{2}$. The latter difference arises because $n_h = 2/(1/n_i + 1/n_j)$, which is twice the comparable value used in the ordinary t Test. The studentized range is then compared to critical values specific to the method in question.

Perhaps the best-known method using sr is the Tukey HSD (honestly significant difference) method. This provides critical values for the largest of all the $k(k - 1)/2$ values of sr for a given data set. The HSD test is slightly more powerful than the Bonferroni method. For instance, when $k = 10$, residual degrees of freedom = 60, and $\alpha = .05$ two-tailed, the critical t -values for the Tukey and Bonferroni methods are 3.29 and 3.43 respectively.

The Duncan method was designed to address this same problem; it also provides tables of critical values of sr . This method is widely available but is commonly considered logically flawed. It has the unusual property that the larger the k , the more likely the outcome is to be significant when the null is true. For instance, if $k = 100$ and if tests are at the .05 level and if all 100 true cell means are equal, the probability is nevertheless .9938 that at least one comparison will be found to be significant. The general expression for such values is $1 - (1 - \alpha)^{k-1}$. Duncan argued that this is reasonable, since when k is large, the largest cell difference is very likely to reflect a real difference. This view seems overly optimistic – if you were testing 100 different pills, each promising to make users lose 20 pounds of weight in the next month, would you want to use a statistical method almost guaranteed to conclude that at least one was effective?

The Fisher protected t Test is also called the LSD (least significant difference) or protected LSD test. Unlike the HSD or Duncan method, it requires no special tables. Rather the investigator first tests the overall F , and if it is significant, tests as many comparisons as desired with ordinary t Tests.

This test is controversial, and its advantages and disadvantages are discussed further below.

The Dunnett method is for a different problem in which the k cells include $k - 1$ treatment groups and one control group, and the only comparisons to be tested are between the control group and the treatment groups. It uses a studentized range. The Bonferroni method could be used here with a correction factor of $k - 1$. The Dunnett method is slightly more powerful than the Bonferroni method.

The Scheffé method is the only method discussed here that applies to comparisons involving more than two cell means, such as when you average the highest 3 of 10 cell means, average the lowest 3 cell means, and want to test the difference between the two averages. Weighted averages may also be used. Since the number of possible weighted averages is infinite, the Bonferroni method cannot be used because B would be infinite. The Scheffé method allows the investigator to use any *post hoc* strategy whatsoever to select the most significant comparisons and to test as many comparisons as desired. When corrected by the Scheffé formula, the tests will still be valid. The Scheffé formula is simple: compute t for the comparison of interest, then compute $F = t^2/(k - 1)$, and compare F to an F table using the same df one would use to test the overall F for that data set. The Scheffé method is less powerful than any of the other methods discussed here and is thus not recommended if any of the other methods are applicable.

Layering Methods for Multiple Comparisons

Suppose there are 10 cells in a one-way analysis of variance, and you number the cells in the order of their means, from 1 (lowest) to 10 (highest). Suppose that by the HSD method you have found a significant difference between cell means 1 and 10. You might then want to test the difference between means 1 and 9 and the difference between means 2 and 10. If the 1 to 9 difference is significant, you might want to test the difference between means 1 and 8 and the difference between means 2 and 9. Continuing in this way so long as results continue to be significant, you might ultimately want to test the difference between adjacent cell means. This entire process is called *layering* or *step-down analysis*, and several methods for it have been proposed.

The Newman–Keuls method simply uses the HSD tables for lower values of k , setting $k = m + 1$ where m is the difference between the ranks of the two means being compared. For instance, suppose you are testing the difference between means 2 and 7, so $m = 5$. These means are the highest and lowest of a set of 6 means, so use the HSD tables with $k = 6$. This method has been criticized as too liberal, but is defended in a later section.

The Ryan–Einot–Gabriel–Welsch method is a noncontroversial Bonferroni-type method for the layering problem in multiple comparisons. Compute the t and p for any comparison in the ordinary way, but then multiply p by a Bonferroni correction factor of $k \times m/2$.

Contingency Tables

Suppose you want to test each cell in a contingency table for a difference between that cell's observed and expected frequency. The sufficient statistics for that null hypothesis are the cell frequency, the corresponding column and row totals, and the grand total for the entire table. Thus you can use those four frequencies to construct a 2×2 table in which the four frequencies are the frequency in the focal cell, all other cases in the focal row, all other cases in the focal column, and all cases not in any of the first three. One can then apply any valid 2×2 test to that table. If you do this for each cell in the table, the Bonferroni method can be used to correct for multiple tests. If R and C are the numbers of rows and columns, then the total number of tests is $R \times C$, so that should be the Bonferroni correction. Bonferroni layering may be used.

It could be argued that once a cell has been identified as an outlying cell, its frequency should somehow be ignored in computing the expected frequencies for other cells. As explained in [1], that problem can be addressed by fitting models with 'structural zeros' for cells already identified as outliers.

What Tests Should One Correct for?

Particularly in the multiple-comparisons area, writers have generally assumed that the set of tests used in a single correction should be all the tests in a single 'experiment', or in a complex experiment, all

4 Multiple Testing

the tests in a single ‘effect’ – that is, all the tests in a single line of the analysis of variance table. The corrected significance levels calculated in this way are called ‘experimentwise’ levels in the former case or ‘familywise’ levels in the latter. But these answers do not apply to many of the examples in our opening paragraph, such as when one is reviewing many studies by different investigators. The broader answer would seem to be that one is trying to reach a particular conclusion, and one should correct for all the tests one scanned to try to support that conclusion. It should be plausible, to at least some people, that all the individual null hypotheses in the scanned set are true, so that any results that are significant individually appeared mainly because so many tests were scanned. Thus, for instance, we would never have to correct for all the tests ever performed in the history of science because there is no serious claim that all the null hypotheses in the history of science are true – the dozens of reliable appliances we use every day contradict that hypothesis.

A single result can drastically change the set that seems appropriate to correct for. For instance, suppose you search the literature on a topic and find 50 significance tests by various investigators. They do not all test the same null hypothesis, though all are in the same general area. One possibility is that all 50 null hypotheses are true, and that might be plausible. But suppose you can divide the 50 tests into 5 groups (A to E) on the basis of their scientific content. Suppose you find that single tests in groups B and D are significant even after any reasonable correction. Then you can dismiss the null hypotheses that there are no real effects in areas B and D, and you might evaluate other tests in those areas with no further correction for multiple tests. Of course, you have also dismissed the null hypothesis of no real effects in the total set of 50. Thus, in areas A, C, and E, you might apply corrections for multiple tests, correcting each only for the number of tests in those subareas. Thus, one or two highly significant results may, as in this example, greatly change your view of what ‘sets’ need to be tested. The choice of the proper sets should be made by subject-matter experts, not by statistical consultants.

This line of reasoning can be used to defend the Newman–Keuls method. In a one-way analysis of variance, suppose the set of k cell means is divided into subsets such that all true cell means within a subset are equal but all the subsets differ from each

other. In that situation, the Newman–Keuls method fails the familywise and experimentwise criteria of validity because the significant results found in the early stages of the layering process will lead the analyst to perform multiple tests in the final stages, and it is highly probable that some of those later tests will be significant (producing Type 1 errors) just by chance. But it can be argued that if all the cell means in one subset have been shown to differ from all those in another subset, then tests within the two subsets are now effectively in two different scientific areas, even though all the cell means were originally estimated in the same ‘experiment’. We all agree that tests in different scientific areas need not be corrected for each other, even though that policy does mean that on any given day some type 1 errors will be made somewhere in the world. By that line of reasoning, the Newman–Keuls method seems valid.

It is harder to see any similar justification for the Fisher LSD method. If one of 10 cell means is far above the other 9, the overall F may well be significant. But this provides no obvious reason for logically separating the lower 9 cell means from each other. If you have selected the largest difference among these 9 cell means, you would be selecting the largest of 36 comparisons. Yet, LSD would allow you to test that with no correction for *post hoc* selection. There is no clear justification for such a process. The Duncan method is far more liberal than LSD and also seems unreasonable.

An Apparent Contradiction

Investigators often observe significant results in an overall test of the null hypothesis – for instance, a significant overall F in a one-way analysis of variance – but then they find no significant individual results. This can happen even without corrections for multiple tests but is far more common when corrections are made. What conclusions can be drawn in such cases?

To answer this question, it is useful to distinguish between specific and vague conclusions. A vague conclusion includes the phrase ‘at least one’ or some similar phrase. For instance, in the significant overall F just mentioned, the proper conclusion is that ‘at least one’ of the differences between cell means must be nonzero. But no specific conclusions can be

drawn – conclusions specifying exactly which differences are nonzero, or even naming one difference known to be nonzero.

Results like this occur simply because it takes more evidence to reach specific conclusions than vague conclusions. This fact is not limited to statistics. For instance, a detective may conclude after only a few hours of investigation that a murder must have been committed by ‘at least one’ of five people but may take months to name one particular suspect, or may never be able to name one. Thus the situation we are discussing arises because the sample contains enough evidence to reach a vague conclusion but not enough evidence to reach a specific conclusion.

For another example, suppose five coins are each flipped 10 times to test the hypothesis that all are fair coins, against the one-tailed hypothesis that some or all are biased toward heads. Suppose the numbers of heads for the five coins are, respectively, 5, 6, 7, 8, and 9. Even the last of the 5 coins is not significantly different from the null at the .05 level after correction for *post hoc* selection; using the exact correction formula for independent tests yields $PC = .0526$ for that coin. But collectively, there are 35 heads in 50 flips, and that result is significant with $p = .0033$. We can reach the vague conclusion that at least one of the coins must be biased toward heads, but we cannot reach any specific conclusions about which one or ones.

More Techniques for Independent Tests, and the ‘File-Drawer’ Problem

Examples like the last one raise an interesting question: If an overall test is significant but no tests for specific conclusions can be reached, can we reach a conclusion that is more specific than the one already reached, even if not as specific as we might like? The answer is yes. For instance, in the five-coin example, consider the conclusion that at least one of the last two coins is biased toward heads. If we ignored the problem of *post hoc* selection we would notice that those two coins showed 17 heads in 20 trials, yielding $p = .0013$ one-tailed. But there are 10 ways to select two coins from five, so we must make a 10-fold Bonferroni correction, yielding $PC = .013$.

We thus have a significant result yielding a conclusion more specific than the conclusion that at least one of the five coins is biased toward heads. This kind of technique seems particularly useful in literature reviews, sometimes allowing conclusions that are positive and fairly specific even when no individual studies are significant after correction for *post hoc* selection. Methods for this problem are discussed at length in [2].

These methods also permit allowance for the file-drawer problem – the problem of possible unpublished negative results. In our example, suppose we were worried that as many as four other coins had been flipped but their results had not been given to us because they came out negative. Thus, our two coins with 17 heads between them in fact yielded the best results of nine coins, not merely the best of five. Thus our appropriate Bonferroni correction is $9!/(2! 7!) = 36$. Applying this correction we have $PC = 36 \times .0013 = .0464$. Thus our two-coin conclusion is significant beyond the .05 level even after allowing for the possibility that we are unaware of several negative results.

References

- [1] Bishop, Y.M., Feinberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge.
- [2] Darlington, R.B. & Hayes, A.F. (2000). Combining independent P values: extensions of the Stouffer and binomial methods, *Psychological Methods* **5**, 496–515.
- [3] Dunn, O.J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association* **56**, 52–64.
- [4] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [5] Ryan, T.A. (1959). Multiple comparisons in psychological research, *Psychological Bulletin* **56**, 26–47.
- [6] Ryan, T.A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics, *Psychological Bulletin* **57**, 318–328.
- [7] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Design*, 3rd Edition, McGraw-Hill, New York.

RICHARD B. DARLINGTON

Multitrait–Multimethod Analyses

SUSAN A. PARDY, LEANDRE R. FABRIGAR AND PENNY S. VISSER

Volume 3, pp. 1343–1347

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multitrait–Multimethod Analyses

The multitrait-multimethod (MTMM) design provides a framework for assessing the validity of psychological measures [5] (*see Measurement: Overview*). Although a variety of MTMM research designs exist, all share a critical feature: the factorial crossing of constructs and measurement procedures (i.e., a set of constructs is assessed with each of several different methods of measurement). MTMM designs have been used to confirm that measures intended to reflect different constructs do assess distinct constructs (i.e., discriminant validity), to confirm that measures intended to assess a given construct are in fact assessing the same construct (i.e., convergent validity), and to gauge the relative influence of method and trait effects.

Traditional Approach to MTMM Analyses

Although MTMM studies are ubiquitous, there has been surprisingly little consensus on how MTMM data should be analyzed. Traditionally, data have been analyzed by visually examining four properties of the resulting correlation matrix of measures in MTMM designs [5]. The first property is ‘monotrait-heteromethod’ correlations (i.e., correlations between different measures of the same construct). Large monotrait-heteromethod correlations indicate convergent validity. These correlations cannot be interpreted in isolation, however, because distinct measures of a construct may covary for other reasons (e.g., because measures tap unintended constructs or contain systematic measurement error). Thus, the ‘heterotrait-heteromethod’ correlations (i.e., correlations between different measures of different constructs) should also be examined. Each monotrait-heteromethod correlation should be larger than the heterotrait-heteromethod correlations that share the same row or column in the matrix. In other words, correlations between different measures of the same construct should be larger than correlations between construct/measure combinations that share neither construct nor measurement procedure. Third, the ‘heterotrait-monomethod’ correlations (i.e., correlations between measures of different

constructs with the same measurement procedures) should be inspected. Monotrait-heteromethod correlations should be larger than heterotrait-monomethod correlations. That is, observations of the same construct with different measurement procedures should be more strongly correlated than are observations of different constructs using the same measurement procedures. Finally, all of the heterotrait triangles should be examined. These triangles are the correlations between each of the constructs measured via the same procedures and the correlations between each of the constructs measured via different procedures. Similar patterns of associations among the constructs should be evident in each of the triangles.

Although the traditional analytical approach has proven useful, it has significant limitations. For example, it articulates no clear assumptions regarding the underlying nature of traits and methods (e.g., whether traits are correlated with methods) and it implies some assumptions that are unrealistic (e.g., random measurement error is ignored). Further, it provides no means for formally assessing the plausibility of underlying assumptions, and it is also cumbersome to implement with large matrices. Finally, implementation of the traditional approach is an inherently subjective and imprecise process. Some aspects of the matrix may be consistent with the guidelines, whereas others may not, and there are gradations to how well the requirements have been met. Because of these limitations, formal statistical procedures for analyzing MTMM data have been developed. The goal of these approaches is to allow for more precise hypothesis testing and more efficient summarizing of patterns of associations within MTMM correlation matrices.

Analysis of Variance (ANOVA) Approach

The **analysis of variance (ANOVA)** approach [6, 9, 14] is rooted in the notion that every MTMM observation is a joint function of three factors: trait, method, and person. ANOVA procedures can be used to parse the variance in MTMM data into that which is attributable to each of these factors and to their interactions. Variance attributable to person (i.e., a large person main effect) is often interpreted as evidence of convergent validity, but not as defined in the classic Campbell and Fiske [5] sense. Rather, it indicates that a general factor accounts for a

substantial portion of the variance in the MTMM matrix (which is useful as a comparison point for higher-order interactions). Variance attributable to the person by trait interaction reflects the degree to which people vary in their level of each trait, averaging across methods of measurement. A large person by trait interaction implies low heterotrait correlations (i.e., discriminant validity). Discriminant validity is high when the person by trait interaction accounts for substantially more variance than the person main effect. Variance attributable to the person by method interaction reflects the degree to which people vary in their responses to the measurement procedures, averaging across traits. Large person by method interactions are undesirable because they indicate that methods of measurement account for substantial variance in the MTMM matrix.

The ANOVA approach is useful in that it parsimoniously summarizes the pattern of correlations in an MTMM matrix, and in that it assesses the relative contribution of various sources of variance. However, the ANOVA also requires unrealistic assumptions (i.e., uncorrelated trait factors, uncorrelated method factors, equivalent reliability for measures, the person by trait by method interaction contains only error variance). Furthermore, the use of the person main effect as an index of convergent validity is problematic because it indicates that individual differences account for variation in observations, collapsing across traits and methods. This is conceptually different from the original notion of convergent validity, reflected in the correlations between different methods measuring a single construct. Finally, the ANOVA yields only omnibus information about trait and method effect. The approach precludes the exploration of correlations among traits, among methods, and between traits and methods.

Traditional MTMM-Confirmatory Factor Analysis (CFA) Model

The traditional MTMM-confirmatory factor analysis (see **Factor Analysis: Confirmatory**) model [7] uses covariance structure modeling to estimate a latent factor model in which each measure reflects a trait factor, a method factor, and a unique factor (i.e., random measurement and systematic measurement error specific to that measure). This model implies that measures assessing different constructs but employing a common method will be correlated because

of a shared method of measurement, and measures assessing the same construct but employing different methods will correlate because of a common underlying trait factor. The model also posits that trait factors are intercorrelated and that method factors are intercorrelated. High discriminant validity is reflected by weak correlations among trait factors. High convergent validity is reflected by large trait loadings and weak method factor loadings. Convergent validity is also improved when the variances in unique factors are small.

The traditional MTMM-CFA provides a parsimonious representation of MTMM data. Its assumptions are relatively explicit, and for the most part, intuitively follow from the original conceptualization of MTMM designs. Unfortunately, this approach routinely encounters computational difficulties in the form of improper solutions. The reasons for this problem are not entirely understood but might involve statistical or empirical identification problems. There are also certain ambiguities in the assumptions of this model that have generally gone unrecognized. For example, although the model assumes that traits and method effects are additive, interactions are possible [4, 15]. Moreover, the model actually allows for two potentially competing types of trait-method interactions. A hierarchical version of this model has been proposed [12]. This version has potential advantages over the first-order model in that it may be less susceptible to problems related to empirical identification, but it also suffers from some of the same limitations such as including competing types of trait-method interactions [15].

Restricted Versions of the Traditional CFA Model

In an effort to take advantage of the strengths of CFA while avoiding computational problems, some researchers [1, 10, 13] have proposed imposing constraints on the traditional MTMM-CFA model, thereby reducing the number of model parameters. For example, one may constrain the factor loadings from each method factor to the observed variables measured with that method to be equal to 1, and constrain the correlations between method factors to be zero.

By imposing constraints, the computational difficulties that arise from underidentification may be remedied. However, these constraints require

assumptions that are often unrealistic. In addition, the imposition of constraints reduces the amount of information provided by the model. For example, restrictions may preclude investigating correlations among method factors or which observed variables a particular trait most strongly influences.

Correlated Uniqueness Model

In this model [8, 11], only traits are represented as latent constructs. The influence of methods is represented in correlated unique variances for measures that share a common method. Thus, the unique variances in this model represent three sources of variability in measures: random error of measurement, specific factors, and method factors. Because both random error and specific factors are by definition uncorrelated, correlations among unique variances are attributable to a shared method of measurement. In this model, traits are permitted to correlate, but method factors are implied to be uncorrelated. Low correlations among traits indicate discriminant validity, and high convergent validity is represented by strong trait factor loadings. Method effects are gauged by the magnitude of the correlations among unique variances, with weak correlations implying greater convergent validity.

One advantage of this model is that it is less susceptible to computational difficulties. However, this model is based on similar assumptions to the traditional MTMM-CFA model, and therefore, the potential for the same logical inconsistencies exist. Also, because method factors are not directly represented, the model is not well suited for addressing specific questions regarding method effects.

Covariance Component Analysis

The Covariance Component Analysis (CCA) model [16] (see **Covariance Structure Models**) is based on the main effects portion of a random effects analysis of variance (see **Fixed and Random Effects**), but is typically tested using covariance structure modeling. Although there are a number of ways to parameterize this model, the Browne [3] parameterization is probably the most useful. In this parameterization, trait and method factors are treated as deviations from a general response factor (i.e., a ‘G’ factor). The scale of the latent variables is set by constraining the variance of the ‘G’ factor to 1.00.

Trait and method factors are represented by ‘Trait-deviation’ and ‘Method-deviation’ factors, and as such, the number of factors representing methods or traits is one less than the number of methods or traits included in the matrix. Trait deviations are permitted to correlate with each other but not with method deviations and vice versa. Neither trait nor method deviations correlate with the general response factor. Finally, a latent variable influencing each observed variable is specified. Random measurement error is incorporated into the model through error terms influencing each of the latent variables associated with each measured variable.

Unfortunately, even this parameterization does not provide immediately useful estimates, due to the fact that estimates for the traits and methods sum to zero. Thus, it is necessary to add ‘G’ separately to the trait and method deviations so that the correlation matrix among traits and the correlation matrix among methods can be computed. These matrices are then used to determine discriminant and convergent validity. Low correlations among traits imply discriminant validity. Convergent validity is supported by strong positive correlations among method factors, which indicate that the methods produce similar patterns of responses. The CCA also provides estimates of communalities for the measured variables.

The CCA model has advantages over many other models because it is a truly **additive model** and thus does not allow for the conceptual ambiguities that can occur in many CFA models. Furthermore, the matrices produced efficiently summarize properties of the data. Importantly, the CCA approach does not have the computational problems characteristic of the traditional MTMM-CFA model. A potential limitation is that it provides relatively little information about the performance of individual observed variables.

Composite Direct Product Model

Traditionally, MTMM researchers have assumed that trait and method factors combine in an additive fashion to influence measures. However, interactions among trait and method factors may exist, thus making additive models inappropriate for some data sets. The Composite Direct Product (CDP) model [2] was developed to deal with interactions between traits and

methods. In this model, each measure is assumed to be a function of a multiplicative combination of the trait and the method underlying that measure. Each measure is also assumed to be influenced by a unique factor. The model allows for correlations among trait factors and correlations among method factors. No correlations are permitted between method factors and trait factors. High discriminant validity is reflected in weak correlations among traits. As with the CCA, high correlations among methods suggest high convergent validity. Estimates of communalities provide information about the influence of random measurement error and specific factors.

One strength of the model is that its parameters can be related quite directly to the questions of interest in MTMM research. The CDP improves upon the traditional MTMM-CFA model and its variants in that it does not suffer from the conceptual ambiguities of these approaches. Moreover, the CDP is less susceptible to computational problems. Applications of the CDP have suggested that it usually produces satisfactory fit and sensible parameter estimates. Interestingly, when applied to data sets for which there is little evidence of trait-method interactions, the CDP model often provides similar fit and parameter estimates to the CCA model. However, the CDP model appears to sometimes perform better than the CCA model for data sets in which there is evidence of trait-method interactions. As with the CCA approach, the model provides relatively little information about the performance of individual measures.

Conclusions

When analyzing MTMM data, it seems most sensible to begin with a consideration of the four properties of the MTMM correlation matrix originally proposed by Campbell and Fiske [5]. If the properties of the original MTMM matrix are consistent with the results of more formal analyses, researchers can be more confident in the inferences they draw from those results. In addition, the assumptions of more formal analytical methods (e.g., do traits and methods combine in an additive or multiplicative manner?) can be evaluated by examining the original MTMM matrix [15]. Researchers should then consider what formal statistical analyses are most appropriate. Many approaches have serious limitations. The CCA and the CDP probably have the strongest conceptual

foundations and the fewest practical limitations. Of these two, the CDP has the advantages of being easier to implement and perhaps more robust to violations of its underlying assumptions about the nature of trait-method effects. However, until the contexts in which the CDP and CCA differ are better understood, it might be useful to supplement initial CDP analyses with a CCA analysis using Browne's [3] parameterization. If the results are similar, one can be confident that the conclusions hold under assumptions of either additive or multiplicative trait-method effects. If they are found to substantially differ, then it is necessary to carefully evaluate the plausibility of each model in light of conceptual and empirical considerations.

References

- [1] Alwin, D.F. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix, in *Sociological Methodology 1973–1974*, H.L. Costner, ed., Jossey-Bass, San Francisco pp. 79–105.
- [2] Browne, M.W. (1984). The decomposition of multitrait-multimethod matrices, *British Journal of Mathematical and Statistical Psychology* **37**, 1–21.
- [3] Browne, M.W. (1989). Relationships between an additive model and a multiplicative model for multitrait-multimethod matrices, in *Multivariate Data Analysis*, R. Coppi & S. Bolasco, eds, Elsevier Science, North-Holland, pp. 507–520.
- [4] Browne, M.W. (1993). Models for multitrait-multimethod matrices, in *Psychometric Methodology: Proceedings of the 7th European Meeting of the Psychometric Society in Trier*, R. Steyer, K.F. Wender & K.F. Widaman, eds, Gustav Fischer Verlag, New York, pp. 61–73.
- [5] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
- [6] Guilford, J.P. (1954). *Psychometric Methods*, McGraw-Hill, New York.
- [7] Joreskog, K.G. (1974). Analyzing psychological data by structural analysis of covariance matrices, in *Contemporary Developments in Mathematical Psychology*, Vol. 2, R.C. Atkinson, D.H. Krantz, R.D. Luce & P. Suppes, eds, W.H. Freeman, San Francisco, pp. 1–56.
- [8] Kenny, D.A. (1979). *Correlation and Causality*, Wiley, New York.
- [9] Kenny, D.A. (1995). The multitrait-multimethod matrix: design, analysis, and conceptual issues, in *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, P.E. Shrout & S.T. Fiske, eds, Erlbaum, Hillsdale, pp. 111–124.
- [10] Kenny, D.A. & Kashy, D.A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis, *Psychological Bulletin* **112**, 165–172.

-
- [11] Marsh, H.W. (1989). Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions, *Applied Psychological Measurement* **13**, 335–361.
- [12] Marsh, H.W. & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: application of second-order confirmatory factor analysis, *Journal of Applied Psychology* **73**, 107–117.
- [13] Millsap, R.E. (1992). Sufficient conditions for rotational uniqueness in the additive MTMM model, *British Journal of Mathematical and Statistical Psychology* **45**, 125–138.
- [14] Stanley, J.C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence, *Psychometrika* **26**, 205–219.
- [15] Visser, P.S., Fabrigar, L.R., Wegener, D.T. & Browne, M.W. (2004). *Analyzing Multitrait-Multimethod Data*, University of Chicago, Unpublished manuscript.
- [16] Wothke, W. (1995). Covariance components analysis of the multitrait-multimethod matrix, in *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske*, P.E. Shrout & S.T. Fiske, eds, Erlbaum, Hillsdale, pp. 125–144.
- (See also **Structural Equation Modeling: Overview**; **Structural Equation Modeling: Software**)
- SUSAN A. PARDY, LEANDRE R. FABRIGAR
AND PENNY S. VISSER

Multivariate Analysis: Bayesian

JIM PRESS

Volume 3, pp. 1348–1352

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Analysis: Bayesian

Multivariate Bayesian analysis is that branch of statistics that uses *Bayes' theorem* (see **Bayesian Belief Networks**) to make inferences about several, generally correlated, unobservable and unknown quantities. The unknown quantities may index probability distributions; they may be hypotheses or propositions; or they may be probabilities themselves (see **Bayesian Statistics**). Such procedures have widespread applications in the behavioral sciences.

Bayes' Theorem

In Bayesian analysis, an unknown quantity is assigned a probability distribution, to represent one's degree-of-belief about the unknown quantity. This degree-of-belief is then modified via Bayes' theorem, as new information becomes available through observational data, experience, or new theory. Multivariate Bayesian inference is based on Bayes' theorem for correlated random variables.

Symbolically, let Θ denote a collection (vector) of k unobservable jointly continuous random variables, and $\mathbf{X}$ a collection of N observable, p -dimensional jointly continuous random vectors. Let $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ denote probability density functions of their arguments. (Lowercase letters will be used to represent observed values of the random variables designated by uppercase letters. Sometimes, for convenience, we will use lowercase letters throughout. We use boldface to designate vectors and matrices.) Bayes theorem for continuous $\mathbf{X}$ and continuous Θ asserts that

$$h(\theta|\mathbf{x}) = \left(\frac{1}{Q} \right) f(\mathbf{x}|\theta)g(\theta), \quad (1)$$

where θ and $\mathbf{x}$ denote values of Θ and $\mathbf{X}$, respectively, and Q denotes a one-dimensional constant (possibly depending on $\mathbf{x}$, but not on θ), which is given by

$$Q = \int f(\mathbf{x}|\theta)g(\theta)d\theta. \quad (2)$$

The integration is taken over all possible values in k -dimensional space, and the notation $f(\mathbf{x}|\theta)$

should be understood to mean the density of the conditional distribution of $\mathbf{X}$, given $\Theta = \theta$. $f(\mathbf{x}|\theta)$ is the joint probability density function for $\mathbf{X}|\Theta = \theta$. When it is viewed as a function of θ , it is called the *likelihood function* (see **Maximum Likelihood Estimation**). When $\mathbf{X}$ and/or Θ are discrete, we replace integrals by sums, and probability density functions by probability mass functions. $g(\theta)$ is the *prior density* of Θ , since it is the probability density of Θ prior to having observed $\mathbf{X}$. Note that the prior density should not depend in any way upon the current data set, although it certainly could, and often does depend upon earlier-obtained data sets. If the prior were permitted to depend upon the current data set, using Bayes' theorem in this inappropriate way would violate the laws of probability. $h(\theta|\mathbf{x})$ is the *posterior density* of Θ , since it is the distribution of Θ 'subsequent' to having observed $\mathbf{X} = \mathbf{x}$.

Bayesian inference in multivariate distributions is based on the posterior distribution of the unobservable random variables, Θ , given the observable data (the unobservable random variable may be a vector or a matrix). A Bayesian estimator of θ is generally taken to be a measure of location of the marginal posterior distribution of Θ , such as its mean, median, or mode. To obtain the marginal posterior density of Θ given the data, it is often necessary to integrate the joint posterior density over spaces of other unobservable random variables that are jointly distributed with θ . For example, to determine the marginal posterior density of a mean vector θ , given data $\mathbf{x}$, from the joint posterior density of a mean vector θ and a covariance matrix Σ , given $\mathbf{x}$, we must integrate the joint posterior density of (θ, Σ) , given $\mathbf{x}$ over all elements of Σ that make it positive definite.

Bayesian Inferences

Bayesian confidence regions (called *credibility regions*) are obtainable for any preassigned level of credibility directly from the cumulative distribution function of the posterior distribution. We make a distinction here between 'credibility' and 'confidence' that is fundamental, and not just a simple choice of alternative words.

The *credibility region* is a probability region for the unknown, unobservable vector or matrix, conditional on the specific value of the observables that happened to have been observed in this instance,

2 Multivariate Analysis: Bayesian

regardless of what values of the observables might be observed in other instances (the region is based upon $P\{\Theta|\mathbf{X}\}$). For example, Ω denotes a 95% credibility region for $\Theta|\mathbf{X}$ if

$$P\{\Theta \in \Omega|\mathbf{X}\} = 95\%. \quad (3)$$

The *confidence region*, by contrast, is obtained from a probability statement about the observable variables, conditional on the unobservable ones. So it really represents a region based upon the distribution of $\mathbf{X}$ as to where the observables are likely to be, rather than where the unobservables are likely to be (the region is based upon $P\{\mathbf{X}|\theta\}$). (For more details, see **Confidence Intervals**).

When nonuniform, proper prior distributions are used, the resulting credibility and confidence regions will generally be quite different from one another.

Predictions about a data vector(s) or matrix not yet observed are carried out by averaging the likelihood for the future observation vector(s) or matrix over the best information we have about the indexing parameters of its distribution, namely, the posterior distribution of the indexing parameters, given the data already observed.

Hypothesis testing may be carried out by comparing the posterior probabilities of all competing hypotheses, given all data observed, and selecting the hypothesis with the largest posterior probability. These notions are identical with those in univariate Bayesian analysis. In multivariate Bayesian analysis, however, in order to make posterior inferences about a given hypothesis, conditional upon the observable data, it is generally necessary to integrate over all the components of Θ .

Multivariate Prior Distributions

The process of developing a prior distribution to express the beliefs of the analyst about the likely values of a collection of unobservables is called *multivariate subjective probability assessment*. None of the variables in a collection of unobservables, Θ , is ever known. The multivariate prior probability density function, $g(\theta)$ for continuous θ (or its counterpart, the prior probability mass function for discrete θ), is used to denote the degrees of belief the analyst holds about θ . The parameters that index the prior distribution are called *hyperparameters*.

Table 1 Prior distribution $g(\theta_1, \theta_2)$

		θ_2	
		0	1
θ_1	0	0.2	0.1
	1	0.3	0.4

For example, suppose a vector mean θ is bivariate ($k = 2$), so that there are two unobservable, one-dimensional random variables Θ_1 and Θ_2 . Suppose further (for simplicity) that Θ_1 and Θ_2 are discrete random variables, and let $g(\theta_1, \theta_2)$ denote the joint probability mass function for $\Theta = (\Theta_1, \Theta_2)'$.

Suppose Θ_1 and Θ_2 can each assume only two values, 0 and 1, and the analyst believes the probable values to be given by those in Table 1.

Thus, for example, the analyst believes that the chances that Θ_1 and Θ_2 are both 1 is 0.4, that is,

$$P\{\Theta_1 = 1, \Theta_2 = 1\} = g(1, 1) = 0.4. \quad (4)$$

Note that this bivariate prior distribution represents the beliefs of the analyst, and need not correspond to the beliefs of anyone else, neither any individual, nor any group. Other individuals may feel quite differently about Θ .

In some situations, the analyst does not feel at all knowledgeable about the likely values of unknown, unobservable variables. In such cases, he/she will probably resort to using a ‘vague’ (sometimes called ‘diffuse’ or ‘noninformative’) prior distribution. Let Θ denote a collection of k continuous, unknown variables, each defined on $(-\infty, \infty)$. $g(\theta)$ is a vague prior density for Θ if the elements of Θ are mutually independent, and if the probability mass of each variable is diffused evenly over all possible values. We write the (improper) prior density for Θ as

$$g(\theta) \propto \text{constant},$$

where $\propto$ denotes proportionality. Note that while this density characterization corresponds in form to the density of a uniform distribution, the fact that this uniform distribution must be defined over the entire real line means that the distribution is improper (it does not integrate to one) except in the case where the components of θ are defined on a finite interval. While prior distributions may sometimes be

improper, the corresponding posteriors must always be proper.

The vague prior for positive definite random variables is quite different. If Σ denotes a k -dimensional square and symmetric positive definite matrix of variances and covariances, a vague (improper) prior for Σ is given by (see [6]):

$$g(\Sigma) \propto |\Sigma|^{-(k+1)/2}$$

where $|\Sigma|$ denotes the determinant of the matrix Σ . Sometimes *natural conjugate* or other families of multivariate prior distributions are used (see, e.g., [10,11]).

Multivariate prior distributions are often difficult to assess because of the complexities of thinking in many dimensions simultaneously, and also because the number of parameters to assess is so much greater than in the univariate case. For some methods that have been proposed for assessing multivariate prior distributions, see, for example, [7, 10] or [11, Chapter 5].

Numerical Methods

Numerical methods are of fundamental importance in Bayesian multivariate analysis. They are used for evaluating and approximating the normalizing constants of multivariate posterior densities, for finding marginal densities, for calculating moments and other functions of marginal densities, for sampling from multivariate posterior densities, and for many other needs associated with multivariate Bayesian statistical inference. The principal methods currently in use are computer-intensive: Markov Chain Monte Carlo (MCMC) methods (see **Markov Chain Monte Carlo and Bayesian Statistics**) and the *Metropolis-Hastings Algorithm* ([1–4, 8, 9], [11, Chapter 6], and [16, 17]).

An Example

The data for this illustrative example first appeared in Kendall [8]. There are 48 applicants for a certain job and they have been scored on 15 variables regarding their acceptability. The variables are as shown in Table 2.

The question is: Is there an underlying subset of factors (of dimension less than 15) that

Table 2 Data reproduced from Kendall, M. *Multivariate Analysis*, 2nd Edition, Griffin Publisher, Charles, 34

1	Form of letter application
2	Appearance
3	Academic ability
4	Likability
5	Self-confidence
6	Lucidity
7	Honesty
8	Salesmanship
9	Experience
10	Drive
11	Ambition
12	Grasp
13	Potential
14	Keenness to join
15	Suitability

explain the variation observed in the 15-dimensional scores? If so, the applicants could then be compared more easily.

Adopt the Bayesian **factor analysis** model

$$\mathbf{x}_j = \underset{(p \times m)}{\mathbf{\Lambda}} \underset{(m \times 1)}{\mathbf{f}_j} + \underset{(p \times 1)}{\boldsymbol{\epsilon}_j} \quad j = 1, \dots, N, \quad m < p. \quad (5)$$

Here, $\mathbf{\Lambda}$ denotes the (unobservable) factor-loading matrix; $\mathbf{f}_j$ denotes the (unobservable) factor score vector for subject j , with $\mathbf{F}' \equiv (\mathbf{f}_1, \dots, \mathbf{f}_N)$; the $\boldsymbol{\epsilon}_j$'s are assumed to be (unobservable) disturbance vectors, mutually uncorrelated and normally distributed as $N(\mathbf{0}, \Psi)$, where Ψ is permitted to be a general, symmetric, positive definite matrix, although $E(\Psi)$ is assumed to be a diagonal matrix. The dimension of the problem is p ; in this example, $p = 15$; m denotes the number of underlying factors (in this example, $m = 4$, but it is generally unknown); and N denotes the number of subjects; in this example, $N = 48$. There are three unknown matrix parameters: $(\mathbf{\Lambda}, \mathbf{F}, \Psi)$. Assume *a priori* that

$$g(\mathbf{\Lambda}, \mathbf{F}, \Psi) = g_1(\mathbf{\Lambda}|\Psi)g_2(\Psi)g_3(\mathbf{F}), \quad (6)$$

where the g 's denote prior probability densities of their respective arguments. Thus, $\mathbf{F}$ is assumed to be *a priori* independent of $(\mathbf{\Lambda}, \Psi)$. Adopt, respectively, the normal and inverted Wishart prior densities:

$$g_1(\mathbf{\Lambda}|\Psi) \propto |\Psi|^{-0.5m}$$

$$\times \exp\{(-0.5)\text{tr}[\Psi^{-1}(\Lambda - \Lambda_0)\mathbf{H}(\Lambda - \Lambda_0)']\}$$

$$g_2(\Psi) \propto |\Psi|^{-0.5\nu} \exp\{-0.5\text{tr}(\Psi^{-1}\mathbf{B})\},$$

where ‘tr’ denotes the matrix trace operation. There are many prior densities for $\mathbf{F}$ that might be used (see [10]). For simplicity, adopt the vague prior density

$$g_3(\mathbf{F}) \propto \text{constant}.$$

The quantities $\mathbf{H}$, Λ_0 , ν , $\mathbf{B}$ are hyperparameters that must be assessed. A model with four factors is postulated on the basis of having carried out a principal components analysis and having found that four factors accounted for 81.5% of the variance. This is therefore the first guess for m .

Adopt the simple hyperparameter structure: $\mathbf{H} = n_0\mathbf{I}_m$, for some preassigned scalar n_0 , and $\mathbf{I}_m$ denotes the identity matrix of order m . In this example, take $\mathbf{H} = 10\mathbf{I}_4$, $\mathbf{B} = 0.2\mathbf{I}_{15}$, and $\nu = 33$. On the basis of the underlying beliefs, the following transposed location hyperparameter, Λ'_0 , for the factor-loading matrix, Λ was constructed:

$$\Lambda'_0 = \begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 & 0.7 & 0.7 & 0.0 & 0.7 & 0.0 & 0.7 & 0.7 & 0.7 & 0.7 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.7 & 0.0 & 0.0 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \end{pmatrix} \quad (7)$$

The four underlying factors are found as linear combinations of the original 15 variables. (See [11, Chapter 15], [12, 13], for additional details.) Factor scores and credibility intervals for the factor scores are given in [11, Chapter 15].

It is shown in [12] how the number of underlying factors may be selected by maximizing the posterior probability over m , the number of factors. An empirical Bayes estimation approach is used in [14] to assess the hyperparameters in the model. Other proposals for assessing the hyperparameters may be found in [5] and [15].

References

- [1] Chib, S. (2003). Markov chain Monte Carlo methods, Chapter 6 of *Subjective and Objective Bayesian Methods: Principles, Models and Applications*, S. James Press, New York: John Wiley & Sons.
- [2] Gelfand, A.E. & Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**, 398–409.
- [3] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- [4] Hastings, W.K. (1970). Mont Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.
- [5] Hyashi, K. (1997). *The Press-Shigemasu factor analysis model with estimated hyperparameters*, Ph.D. thesis, Department of Psychology, University of North Carolina, Chapel Hill.
- [6] Jeffreys, H. (1961). *Theory of Probability*, 3rd Edition, Clarendon, Oxford.
- [7] Kass, R. & Wasserman, L. (1996). The selection of prior distributions by formal rules, *Journal of the American Statistical Association* **91**, 1343–1370.
- [8] Kendall, M. *Multivariate Analysis*, 2nd Edition, Griffin Publisher, Charles, 34.
- [9] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.
- [10] Press, S.J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd Edition, (Revised), Krieger Pub. Co, Melbourne.
- [11] Press, S.J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, John Wiley & Sons, New York.
- [12] Press, S.J. & Shigemasu, K. (1989). Bayesian inference in factor analysis, in *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, Gleser, L.J. Perlman, M.D. Press S.J. & Simpson, A.R. eds, Springer-Verlag, New York, 271–287.
- [13] Press, S.J. & Shigemasu, K. (1999). A note on choosing the number of factors, with K. Shigemasu, *Communications in Statistics: Theory and Methods* **28**(8).
- [14] Press, S.J., Shin, K.-Il. & Lee, S.E. (2004). Empirical Bayes Assessment of the hyperparameters in Bayesian factor analysis, submitted to *Communications in Statistics, Theory and Methods*.
- [15] Shera, D.M. & Ibrahim, J.G. (1998). *Prior elicitation and computation for Bayesian factor analysis*, submitted manuscript, Department of Biostatistics,

- School of Public Health, Harvard University, Boston.
- [16] Spiegelhalter, D., Thomas, A., Best, N. & Gilks, W. (1994). *BUGS: Bayesian inference using Gibbs sampling*, available from MRC Biostatistics Unit, Cambridge.
- [17] Tanner, M.A. & Wong, W. (1987). The calculation of posterior distributions (with discussion), *Journal of the American Statistical Association*, **82**, 528–550.

JIM PRESS

Multivariate Analysis: Overview

WOJTEK J. KRZANOWSKI

Volume 3, pp. 1352–1359

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Analysis: Overview

Many social or behavioral studies lead to the collection of data in which more than one variable (feature, attribute) is measured on each sample member (individual, subject). This can happen either by design, where the variables have been chosen because they represent essential descriptors of the system under study, or because of expediency or economy, where the study has been either difficult or expensive to organize, so as much information as possible is collected once it is in progress. If p variables ($X_1, X_2, \dots, X_p$, say) have been measured on each of n sample individuals, then the resulting values can be displayed in a *data matrix* $\mathbf{X}$. Standard convention is that rows relate to sample individuals and columns to variables, so that $\mathbf{X}$ has n rows and p columns, and the element x_{ij} in the i th row and j th column is the value of the j th variable exhibited by the i th individual. For example, the data matrix in Table 1 gives the values of seven measurements in inches (chest, waist, hand, head, height, forearm, wrist) taken on each of 17 first-year university students.

In most cases, the rows of a data matrix will not be related in any way, as the values observed on

any one individual in a sample will not be influenced by the values of any other individual. This is obvious in the case of the students' measurements. However, since the whole set of variables is measured on *each* individual, the columns will be related to a greater or lesser extent. In the example above, tall students will tend to have large values for all the variables, while short ones will tend to have small values across the board. There is, therefore, correlation between any pair of variables, and a typical multivariate data matrix will have a possibly complex set of interdependencies linking the columns. This means that it is dangerous to analyze the variables individually if general conclusions are desired about the overall system. Rather, techniques that analyze the whole ensemble of variables are needed, and this is the province of multivariate analysis.

Multivariate methods have had a slightly curious genesis and development. The earliest work, dating from the end of the nineteenth century, was rooted in practical problems arising from social and educational research, specifically in the disentangling of factors underlying correlated Intelligence Quotient (IQ) tests. This focus gradually widened and led to many developments in the first half of the twentieth century, but because computational aids were very limited, these developments were primarily mathematical. Indeed, multivariate research at that time might almost be viewed as a branch of linear algebra, and practical applications were very limited as a consequence of the formidable computational requirements attaching to most techniques. However, the advent and rapid development of electronic computers in the second half of the twentieth century removed these barriers. The first effect was an expansion in the practical use of established techniques, which was then followed by renewed interest in development of new techniques. Inevitably, this development became more and more dependent on sophisticated computational support, and, as the century drew to a close, the research emphasis had shifted perceptibly from mathematics to computer science. Currently, new results in multivariate analysis are reported as readily in sociological, psychological, biological, or computer science journals as they are in the more traditional statistical, mathematical, or engineering ones.

The general heading 'multivariate analysis' covers a great variety of different techniques that can be grouped together for ease of description in various

Table 1 Data matrix representing seven measures on each of 17 students

Row	Chest	Waist	Hand	Head	Height	Arm	Wrist
1	36.0	29.00	8.00	22.00	70.50	10.0	6.50
2	36.0	29.00	8.50	23.00	70.00	10.5	6.50
3	38.5	31.00	8.50	24.00	72.50	13.0	7.00
4	36.0	31.00	8.50	23.00	71.00	11.0	7.00
5	39.0	33.00	9.00	24.00	75.00	11.0	7.50
6	37.0	33.00	9.00	22.00	71.00	10.0	7.00
7	38.0	30.50	8.50	23.00	70.00	10.0	7.00
8	39.0	32.00	9.00	23.00	70.00	11.0	7.00
9	39.0	32.00	8.00	24.00	73.00	11.0	7.00
10	36.0	30.00	7.00	22.00	70.00	14.0	6.50
11	38.0	32.00	8.00	23.00	71.00	11.0	7.00
12	41.0	36.00	8.50	23.00	75.00	12.0	7.00
13	38.0	33.00	9.00	23.00	73.50	12.0	7.00
14	36.0	31.00	9.00	23.00	70.50	11.5	7.00
15	43.0	32.00	9.00	22.50	74.00	12.0	8.00
16	36.0	30.00	8.50	22.00	68.00	10.0	6.50
17	38.0	34.00	8.00	23.00	70.00	13.0	7.00

ways. Different authors have different preferences, but, for the purpose of this necessarily brief survey, I will consider them under four headings: Visualization and Description; Extrapolation and Inference; Discrimination and Classification; Modeling and Explanation. I will then end the article by briefly surveying current trends and developments.

Visualization and Description

A data matrix contains a lot of information, often of quite complex nature. Large-scale investigations or surveys may contain thousands of individuals and hundreds of variables, in which case assimilation of the raw data is a hopeless task. Even in a relatively small dataset, such as that of the university students above, the interrelationships among the variables make direct assimilation of the numbers difficult. As a first step towards analysis, a judicious pictorial representation may help to uncover patterns, identify unusual observations, suggest relationships between variables, and indicate whether assumptions made by formal techniques are reasonable or not. So, effort has been expended in devising ways of representing multivariate data with these purposes in mind, and several broad classes of descriptive techniques can be identified.

The simplest idea derives from the *pictogram* and is a method for directly representing a multivariate sample in two dimensions (so that it can be drawn on a sheet of paper or shown on a computer screen). The aim here is to represent each individual in the sample by a single object (such as a box, a star, a tree, or a face) using the values of the variables for that individual to determine the size or features of the object (see **Star and Profile Plots**). So, for example, each variable might be associated with a branch of a tree, and the value of the variable determines its length, or each variable is associated with a facial feature, and the value determines either its size or its shape. The idea is simple and appealing, and the methods can be applied whether the data are quantitative, qualitative, or a mixture of both, but the limitations are obvious: difficulties in relating individuals and variables or directly comparing the latter across individuals, inability to cope with large samples, and no scope for developing any interpretations or analyses from the pictures. Andrews' curves [2], where each individual is represented by a Fourier curve over a range

of values, introduces more quantitative possibilities, but limitations still exist.

A more fruitful line of approach when the data are quantitative is to imagine the $n \times p$ data matrix as being represented by n points in p dimensions. Each variable corresponds to a dimension and each individual to a point, the coordinates of each point on p orthogonal axes being given by the values of the p variables for the corresponding individual. Of course, if p is greater than 2, then such a configuration cannot be seen directly and must somehow be approximated. The **scatterplot** has long been a powerful method of data display for bivariate data, and computer graphics now offer the user a large number of possibilities. The basic building block is the **scatterplot matrix**, in which the scatterplots for all $p(p-1)$ ordered pairs of variables are obtained and arranged in matrix fashion (so that each marginal scale applies to $p-1$ plots). The separate plots can be linked by *tagging* (where individual points are tagged so that they can be identified in each plot), *brushing* (where sets of points can either be highlighted in, or deleted from, each plot), and *painting* (where color is used to distinguish grouping or overplotting in the diagrams). The development of such depth-cueing devices as *kinematic* and *stereoscopic* displays has led to **three-dimensional scatterplot** depiction, and modern software packages now routinely include such facilities as dynamic rotations of three-dimensional scatterplots, tourplots, and rocking rotations (i.e., ones with limited angular displacements). Moreover, window-based systems allow the graphics to be displayed in multiple windows that are linked by their observations and/or their variables. For details, see [25] or [26].

Despite their ingenuity and technical sophistication, the above displays are still limited to pairs or triples of variables and stop short at pure visualization. To encompass all variables simultaneously, and to allow the possibility of substantive interpretation, we must move on to *projection* techniques. The idea here is to find a low-dimensional subspace of the p -dimensional space into which the n points can be projected; for a sufficiently low dimensionality (2 or 3), the points can then be plotted and any patterns in the data easily discerned. The problem is to find a projection that minimizes the distortion caused by approximation in a lower dimensionality. However, what is meant by 'minimizing distortion' will depend on what features of the data are considered to

be important; many different projection methods now exist, each highlighting different aspects of the data.

The oldest such technique, dating back to the turn of the twentieth century, is **principal component analysis** (PCA) [17]. This has many different characterizations, but the two most common ones are as the subspace into which the points are orthogonally projected with least sum of squared (perpendicular) displacements, or, equivalently, as the subspace which maximizes the total variance among the coordinates of the points. The latter characterization identifies the subspace as the one with maximum inter-point scatter, so is the one in which the overall configuration of points is best viewed to detect interesting patterns. The technique has various side benefits: it has an analytical solution from which configurations in any chosen dimensionality can be obtained, and the chosen subspace is defined by a set of orthogonal linear combinations or contrasts among the original variables. The latter property enables directions of scatter to be substantively interpreted, or *reified*, which adds power to any analysis.

If the data have arisen from a number of *a priori* groups, and one objective in data visualization is to highlight group differences, then *canonical variate analysis* (CVA) [6] is the appropriate technique to use (see **Canonical Correlation Analysis; Discriminant Analysis**). This has similarities to PCA in that the requisite subspaces are again defined by linear combinations or contrasts of the original variables, and the solution is also obtained analytically for any dimensionality, but the projections in this case are not orthogonal, and, hence, interpretation is a little harder. If other facets of the data are to be highlighted (e.g., subspaces in which either **skewness** or **kurtosis** is maximized, or subspaces in which the data are as nonnormal as possible), then computational searches rather than analytical solutions must be undertaken. All such problems fall into the category known as **projection pursuit** [18]. Moreover, dimensionality reduction can be optimally linked to some other statistical techniques in this way, as with *projection pursuit regression*.

The above projection methods require quantitative data, in order for the underlying model of n points in p dimensions to be realizable. So what can be done if qualitative data (such as presence/absence of attributes, or categories like different colors) are present? Here, we can link the concepts of *dissimilarity* and *distance* to enable us to *construct* a

low-dimensional representation of the data for visualization and interpretation (see **Proximity Measures**). There are many ways of measuring the *dissimilarity* between two individuals [15]. If we choose a suitable measure appropriate for our data, then we can compute a *dissimilarity matrix* that gives the dissimilarity between every pair of individuals in the sample. **Multidimensional scaling** is a suite of techniques that represents sample individuals by points in low-dimensional space in such a way that the distance between any two points approximates as closely as possible the dissimilarity between the corresponding pair of individuals. If the approximation is in terms of exact values of distance/dissimilarity, then *metric* multidimensional scaling provides an analytical solution along similar lines to PCA, while if the approximation is between the rank orders of the dissimilarities and distances, then *nonmetric* multidimensional scaling provides a computational solution along the lines of projection pursuit. Many other variants of multidimensional scaling now also exist [5].

The above are the most commonly used multivariate visualization techniques in practical applications. They are essentially *linear* techniques, in that they deal with linear combinations of variables and involve linear algebra for their solution. Various *non-linear* techniques have recently been developed [13], but are only slowly finding their way into general use.

Extrapolation and Inference

Frequently, a multivariate data set comes either as a sample from some single population of interest or as a collection of samples from two or more populations of interest. The main focus then may not be on the sample data *per se*, but rather on extrapolating from the sample(s) to the population(s) of interest, that is, on drawing *inferences* about the population(s).

For example, the data on university students given above is actually a subset of a larger collection, in which the same measurements were taken on samples of both male and female students from each of a number of different institutions. Interest may then center on determining whether there are any differences overall, either between genders or institutions, in terms of the features measured for the samples.

In order to make inferences about populations, it is necessary to postulate probability models for these

4 Multivariate Analysis: Overview

populations, and then to couch the inferences in terms of statements about the parameters of these models. In classical (frequentist) inference, the traditional ways of doing this are through either estimation (whether point or interval) of the parameters or testing hypotheses about them. In Bayesian inference, all statements are channeled through the posterior distribution of the model parameters (*see Bayesian Statistics*).

Taking classical methods first, the underlying population model for quantitative multivariate data has long been the multivariate normal distribution (*see Catalogue of Probability Density Functions*), although some attention has also been paid in recent years to the class of *elliptically contoured distributions* [11]. The multivariate normal distribution is characterized completely by its mean vector and dispersion matrix, and most traditional methods of estimation (such as **maximum likelihood**, least squares, or invariance) lead to the sample mean vector and some constant times the sample covariance matrix as the estimates of the corresponding population parameters. Moreover, distribution theory easily leads to confidence regions, at least for the population mean vector. So, for estimation, multivariate results appear to be ‘obvious’ generalizations of familiar univariate quantities.

Hypothesis testing is not so straightforward, however. Early results were derived by researchers who sought ‘intuitive’ generalizations of univariate test statistics, such as Hotelling’s T^2 statistics for one- and two-sample tests of means as generalizations of the familiar one- and two-sample t statistics in univariate theory. It soon became evident that this avenue afforded limited prospects, however, so a more structured approach was necessary. Several general methods of test construction thus evolved, and the most common of these were the likelihood ratio, the union-intersection, and the invariant test procedures. An unfortunate side effect of this multiplicity is that each approach in general leads to a different test statistic in any single situation, the only exceptions being those that lead to the T^2 statistics mentioned above. Typically, the likelihood ratio approach produces test statistics involving products of all **eigenvalues** of particular matrices, the union-intersection approach produces test statistics involving just the largest eigenvalue of such matrices, and other approaches frequently produce test statistics involving the sum of the eigenvalues; in

statistical software, the product statistic is generally termed *Wilks’ lambda*, the largest eigenvalue is usually referred to as *Roy’s largest root*, while the sum statistic is either *Pillai’s trace* or *Hotelling–Lawley’s trace* (for more details *see Multivariate Analysis of Variance*). Tables of critical values exist [24] for conducting significance tests with these statistics, but it should be noted that the different statistics test for different departures from the null hypothesis, and so, in any given situation, may provide contradictory inferences.

One- or two-sample tests can be derived for most situations by using one of the above approaches, and most standard textbooks on multivariate analysis quote the relevant statistics and associated distribution theory. When there are more than two groups, or if there is a more complex design structure on the data (such as a factorial cross-classification, for example), then the *multivariate* analysis of variance (MANOVA) provides the natural analog of *univariate analysis of variance* (ANOVA). The calculations for any particular MANOVA design are exactly the same as for the corresponding ANOVA design, except that they are carried out for each separate variable and then for each pair of variables, and the results are collected together into matrices. All resulting ANOVA sums of squares are thereby translated into MANOVA sums of squares and product matrices, and in place of a ratio of mean squares for any particular ANOVA effect, we have a product of one matrix with the inverse of another. Test statistics are then based on the eigenvalues of these matrices, using one of the four variants described above. Moreover, if any effect proves to be significant, then *canonical variate analysis* (also referred to as *descriptive discriminant analysis*; *see Discriminant Analysis*) can be used to investigate the reasons for the significance.

Most classical multivariate tests require the data to come from either the multivariate normal or an elliptically contoured distribution for the theory to be valid and for the resulting tests to have good properties. This is sometimes a considerable restriction (*see Multivariate Normality Tests*). However, the computing revolution that took place in the last quarter of the twentieth century means that very heavy computing can now be undertaken for even the most routine analysis. This has opened up the possibility of conducting inference computationally rather than mathematically, and such computer-intensive procedures as

permutation testing, **bootstrapping**, and **jackknifing** [7] now form a routine approach to inference.

Bayesian inference requires the specification of a prior distribution for the unknown parameters, and this prior distribution is converted into a posterior distribution by multiplying it by the likelihood of the data. Inference is then conducted with reference to this posterior distribution, perhaps after integrating out those parameters in which there is no interest. For many years, **multivariate Bayesian inference** was hampered by the complicated multidimensional integrals that had to be effected during the derivation of the posterior distribution, and older textbooks restricted attention either to a small set of conjugate prior distributions or to the case of prior uncertainty. Since the development of **Markov Chain Monte Carlo** (MCMC) methods in the early 1990s, however, the area has been opened up considerably, and, now, virtually any prior distribution can be handled numerically using the software package BUGS.

Discrimination and Classification

Whenever there are *a priori* groups in a data set, for example the genders or the institutions for the measurements on university students, the above methods give a mechanism for deciding whether the populations from which the groups came are different or not. If the populations are different, then further questions become relevant: How can we best characterize the differences? How can we best allocate a future unidentified individual to his or her correct group? The first of these questions concerns *discrimination* between the groups, while the second concerns *classification* of (future) individuals to the specified groups. Such questions have assumed great currency, particularly in such areas as health, where it is important to be able to distinguish between, and to correctly allocate individuals to, groups such as disease classes. This topic is a genuinely multivariate one, as intuitively one feels that increasing the number of variables on each individual should increase the ability to discriminate between groups (see **Discriminant Analysis**). Unfortunately, however, the word ‘classification’ has also been used in a rather different context, namely, in the sense of partitioning a collection of individuals into distinct classes. This usage has come down from biological/taxonomic applications, and in the statistical literature, has been distinguished

from allocation by calling it *clustering*. Another way of distinguishing the two usages, popularized in the **pattern recognition** and computer science literature, is to refer to the allocation usage as *supervised classification*, and to the partitioning usage as *unsupervised classification*. This is the nomenclature that we will adopt in the present overview.

At the heart of supervised classification is a function or set of functions of the measured variables. For simplicity, we will refer just to a single function $f(\mathbf{x})$. If distinguishing the groups is the prime purpose, then $f(\mathbf{x})$ is usually termed a *discriminant function*, while if allocation of individuals to groups is the intended usage, it is termed a *classification function* or *classifier*. The first appearance of such a quantity in the literature was the now famous *linear discriminant function* between two groups, derived by **R.A. Fisher** in 1935 on purely intuitive grounds. Subsequently, Welch demonstrated in 1938 that it formed an optimal classifier into one of two multivariate normal populations that shared the same dispersion matrix, while Anderson provided the sample version in 1951.

In the years since then, a great variety of discriminant functions and classifiers have been suggested. Relaxing the assumption of common dispersion matrices in normal populations leads to a *quadratic discriminant function*; eschewing distributional assumptions altogether requires a *non-parametric discriminant function*; estimating probabilities of class membership for classification purposes can be done via *logistic discriminant functions* (see **Logistic Regression**); producing a classifier by means of a sequence of decisions taken on individual variables results in a *tree-based classifier* (see **Classification and Regression Trees**); while recent years have seen the development of a variety of computer-intensive classifiers using **neural networks**, *regularized discriminant functions*, and *radial basis functions*. Bayesian approaches to parametric discriminant analysis were hampered in the early days for the same reasons as already described above, but MCMC procedures are now beginning to be applied in the classification field also. With so many potential classifiers to choose from, ways of assessing their performance is an important consideration. There are various possible measures of performance [16], but the most common one in practical use is the *error rate* associated with a classifier (see **Misclassification Rates**). Another focus of interest

in supervised classification is the question of *variable selection*; in many problems, only a restricted number of variables provide genuine discriminatory information between groups, while the other variables essentially only contribute noise. Classification performance, as well as economy of future measurement, can, therefore, be enhanced by finding this ‘effective’ subset of variables, and much research has, therefore, been conducted into variable selection procedures. In the Bayesian approach, variable selection requires the use of reversible jump MCMC. Full technical details of all these aspects can be found in [8, 16, 22].

Unsupervised classification has an even longer history than the supervised variety, dating back essentially to the numerical taxonomy and biological classification of the nineteenth century. It has developed as a series of ad-hoc advances rather than as a systematic field, and can now be characterized in terms of several groups of loosely connected methods. The oldest set of methods are the *hierarchical clustering* methods in which the sample of n individuals is either progressively grouped, starting with separate groups for each individual and finishing with a single group containing all individuals by fusing two existing groups at each stage, or progressively splitting from a single group containing all individuals down to n groups, each containing a single individual, by dividing an existing group into two at each stage. Both approaches result in a ‘family tree’ structure, showing how individuals either successively fuse together or split apart, and this structure fitted well into the biological classification schemes that were its genesis (see **Hierarchical Clustering**). However, over the years, many different criteria have been proposed for fusing or splitting groups, so now the class of hierarchical methods is a very large one. By contrast with hierarchical methods are the *optimization* methods of clustering, in which the sample of n individuals is partitioned into a prespecified number g of groups in such a way as to optimize some numerical criterion that quantifies the two desirable features of *internal cohesion* within groups and *separability* between groups (see **k-means Analysis**). As can be imagined, many different criteria have been proposed to quantify these features, so there are also many such clustering methods. Moreover, each of these methods often requires formidable computing resources if n is large and needs to be repeated for different values of g if an optimum number of groups is also desired.

Other possible approaches to clustering involve probabilistic models for groups (*mode-splitting* or *mixture separation*; see **Model Based Cluster Analysis**) or *latent variable models* (see below). Further technical details may be found in [10, 14].

Modeling and Explanation

It has already been stressed that association between the variables is one of the chief characteristics of a multivariate data set. If the variables are quantitative, then the correlation coefficient is a familiar measure of the strength of association between two variables, while if the variables are categorical, then the data set can be summarized in a **contingency table** and the strength of association between the variables can be assessed through a chi square test using the chi square measure X^2 . If there are many variables in a data set, then there are potentially many associations between pairs of variables to consider. In the students’ measurements, for example, there are seven variables and, hence, 21 correlations to assess. With 20 variables, there will be 190 correlations, and the number rapidly escalates as the number of variables increases. The pairwise nature of these values means that many interrelationships exist among a set of such correlations, and these interrelationships must be disentangled if any sense is to be made of the overall situation.

One simple technique is to calculate **partial correlations** between specified variables when others are considered to be held fixed. This will help to determine whether an apparently high association between a pair of variables is genuine or has been induced by their mutual association with the other variables. Another simple technique is to treat correlation r as ‘similarity’ between two variables, in which case a measure such as $1 - r^2$ or $(1 - r)/2$ can be viewed as a ‘distance’ between them. Applying some form of multidimensional scaling to a resulting set of distances will then represent the variables as points on a low-dimensional graph and will show up groups of ‘similar’ variables as well as highlighting those variables that are very different from the rest. If variables are qualitative/categorical, then **correspondence analysis** is a technique for obtaining low-dimensional representations that highlight relationships, both between variables and between groups of individuals. Finally, if the variables have some

form of *a priori* grouping structure on them, then **canonical correlation analysis** will extract essential association between these groups from the detail of the whole correlation matrix.

However, if a deeper understanding is to be gained into the nature of a set of dependencies, possibly leading on to a substantive explanation of the system, then some form of underlying model must be introduced. This area forms arguably the oldest single strand of multivariate research, that of **latent variable** models. The origins date back to the pioneering work of **Spearman** and others in the field of educational testing at the start of the twentieth century, where empirical studies showed the presence of correlations between scores achieved by subjects over a range of different tests. Initial work attempted to explain these correlations by supposing that each score was made up of the subject's value on an underlying, unobservable, variable termed 'intelligence', plus a 'residual' specific to the individual subject and test. First results seemed promising, but it soon became evident that such a single-variable explanation was not sufficient. The natural extension was then to postulate a model in which a particular test score was made up of a combination of contributions from a *set* of unobservable underlying variables in addition to 'intelligence,' say 'numerical ability,' 'verbal facility,' 'memory,' and so on, plus a residual specific to the individual subject and test. If these underlying *factors* could account for the observed inter-test correlations, then they would provide a characterization of the system. Determination of number and type of factors, and extraction of the combinations appropriate to each test thus became known as **factor analysis**. The ideas soon spread to other disciplines in the behavioral science, where unobservable (or *latent*) variables such as 'extroversion', 'political leanings', or 'degree of industrialization' were considered to be reasonable constructs for explanations of inter-variable dependencies.

The stumbling block in the early years was the lack of reliable computational methods for estimating the various model parameters and quantities. Indeed, the first half of the twentieth century was riven with conflicting suggestions and controversies, and many of the proposed methods often failed to find satisfactory solutions. The development of the maximum likelihood approach by Lawley and **Maxwell** in the 1940s was one major advance, but computational problems still remained and were not satisfactorily

resolved till the work of Jöreskog [19]. Now, it is a popular and much used technique throughout the behavioral sciences (although many technical pitfalls still remain, and successful applications require both skill and care). However, it is essentially only applicable to quantitative data. Development with qualitative or categorical variables was much slower, the first account being that by Lazarsfeld and Henry [20] under the banner of *latent structure analysis*. This area then underwent rapid development, and, now, there exist many variants such as *latent class analysis*, *latent trait analysis*, *latent profile analysis*, and others (see **Finite Mixture Distributions**). Further details of these techniques may be found in [4, 9].

Current Trends and Developments

The dramatic developments in computer power over the past twenty years or so have already been noted above, and it is these developments that drive much of the current progress in multivariate analysis. If the first half of the twentieth century can be described as years of mathematical foundations and development, and the succeeding twenty or thirty years as ones of widening practical applicability, then recent developments have centered firmly on computational power. Many of these developments have been brought on by the enormous data sets that arise in modern scientific research, due in no small measure to highly sophisticated and rapid data-collecting instruments. Thus, we have spectroscopy in chemistry, flow cytometry in medicine, and microarray experiments in biology (see **Microarrays**) which all yield multivariate data sets comprising many thousands of observations. Such data matrices bring their own problems and objectives, and this has led to many advances in computationally intensive multivariate analysis. Some specific techniques include: curve fitting through localized nonparametric smoothing and recursive partitioning, a flexible version of which is MARS ('multivariate adaptive regression splines') [12]; the use of wavelets for function estimation [1]; the analysis of data that themselves come in the form of functions [23]; the detection of pattern in two-way arrays by using biclustering [21]; and the decomposition of multivariate spatial and temporal data into meaningful components [3].

While much of the impetus for these developments has come from the scientific area, many of the

resultant techniques are, of course, applicable to the large data sets that are becoming more prevalent in behavioral science also. Indeed, a completely new field of study, **data mining**, has sprung up in recent years to deal with explorative aspects of large-scale multivariate data, drawing on elements of computer science as well as statistics. It is evident that the field is a rich one, and many more developments are likely in the future.

References

- [1] Abramovich, F., Bailey, T.C. & Sapatinas, T. (2000). Wavelet analysis and its statistical applications, *Journal of the Royal Statistical Society, Series D* **49**, 1–29.
- [2] Andrews, D.F. (1972). Plots of high-dimensional data, *Biometrics* **28**, 125–136.
- [3] Badcock, J., Bailey, T.C., Krzanowski, W.J. & Jonathan, P. (2004). Two projection methods for multivariate process data, *Technometrics* **46**, 392–403.
- [4] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd Edition, Arnold, London.
- [5] Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling, Theory and Applications*, Springer, New York.
- [6] Campbell, N.A. & Atchley, W.R. (1981). The geometry of canonical variate analysis, *Systematic Zoology* **30**, 268–280.
- [7] Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, University Press, Cambridge.
- [8] Denison, D.G.T., Holmes, C.C., Mallick, B.K. & Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*, Wiley, Chichester.
- [9] Everitt, B.S. (1984). *An Introduction to Latent Variable Models*, Chapman & Hall, London.
- [10] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th Edition, Arnold, London.
- [11] Fang, K.-T. & Zhang, Y.-T. (1990). *Generalized Multivariate Analysis*, Science Press, Beijing and Springer, Berlin.
- [12] Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**, 1–141.
- [13] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, New York.
- [14] Gordon, A.D. (1999). *Classification*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.
- [15] Gower, J.C. & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification* **3**, 5–48.
- [16] Hand, D.J. (1997). *Construction and Assessment of Classification Rules*, Wiley, Chichester.
- [17] Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd Edition, Springer, New York.
- [18] Jones, M.C. & Sibson, R. (1987). What is projection pursuit? (with discussion), *Journal of the Royal Statistical Society, Series A* **150**, 1–36.
- [19] Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis, *Psychometrika* **32**, 443–482.
- [20] Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent Structure Analysis*, Houghton-Mifflin, Boston.
- [21] Lazzeroni, L. & Owen, A. (2002). Plaid models for gene expression data, *Statistica Sinica* **12**, 61–86.
- [22] McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.
- [23] Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*, Springer, New York.
- [24] Seber, G.A.F. (1984). *Multivariate Observations (Appendix D)*, Wiley, New York.
- [25] Wegman, E.J. and Carr, D.B. (1993). Statistical graphics and visualization, in *Computational Statistics. Handbook of Statistics 9*, C.R. Rao, ed., North Holland, Amsterdam, pp. 857–958.
- [26] Young, F.W., Faldowski, R.A. & McFarlane, M.M. (1993). Multivariate statistical visualization, in *Computational Statistics. Handbook of Statistics 9*, C.R. Rao, ed., North Holland, Amsterdam, pp. 959–998.

(See also **Multivariate Multiple Regression**)

WOJTEK J. KRZANOWSKI

Multivariate Analysis of Variance

JOACHIM HARTUNG AND GUIDO KNAPP

Volume 3, pp. 1359–1363

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Analysis of Variance

The multivariate linear model is used to explain and to analyze the relationship between one or more independent or explanatory variables and $p > 1$ quantitative dependent or response variables that have been observed for each of n subjects. When all the explanatory variables are quantitative, the multivariate linear model is referred to as a *multivariate multiple linear regression model* (see **Multivariate Multiple Regression**). When all the explanatory variables are measured on a qualitative scale, or in other words, when the observations can only be made within a fixed number of factor level combinations, the model is called a *multivariate analysis of variance model*. The techniques used to draw inferences about the model parameters are typically referred to by the generic term multivariate analysis of variance or for short MANOVA.

The simplest MANOVA model is the one-way model in which the p quantitative variables are only affected by a single factor with r levels. The one-way MANOVA model can also be thought of as explaining p quantitative variables arising from r different populations. Typically, the null hypothesis of interest in such a model is that the mean responses of the groups corresponding to the different levels of the factor (or the r populations) are equal. The alternative hypothesis is that there is at least one pair of levels of the factor with different mean responses.

Consider the hypothetical example from Morrison [2] that has been changed slightly for the current presentation (Table 1). Assume that $s = 8$ rats were randomly allocated to each of $r = 3$ drug treatments (i.e., $n = 24$) and weight losses in grams were observed for each rat at two time points (after 1 week and after 2 weeks).

Let y_{ij1} denote the weight loss of rat j with drug i after 1 week, y_{ij2} the weight loss of the same rat after 2 weeks, and $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2})^T$ the vector of observations for rat j on drug i . Let μ_{i1} and μ_{i2} further denote the means of variables 1 and 2 for group i . Then, the expectation of $\mathbf{y}_{ij}$ is $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})^T$, and a one-way MANOVA model for this

Table 1 Weight losses in $n = 24$ rats (first week, second week)

Drug		
A	B	C
(5, 6)	(7, 6)	(21, 15)
(5, 4)	(7, 7)	(14, 11)
(9, 9)	(9, 12)	(17, 12)
(7, 6)	(6, 8)	(12, 10)
(7, 10)	(10, 13)	(16, 12)
(6, 6)	(8, 7)	(14, 9)
(9, 7)	(7, 6)	(14, 8)
(8, 10)	(6, 9)	(10, 5)

example is given by

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \mathbf{e}_{ij}, \quad i = 1, 2, \quad r = 3, \\ j = 1, \dots, s = 8, \quad (1)$$

with $\mathbf{e}_{ij} = (e_{ij1}, e_{ij2})^T$, a vector of errors that arises from a multivariate normal distribution (see **Catalogue of Probability Density Functions**) with mean vector $\mathbf{0}$ and (unknown) **covariance matrix** $\boldsymbol{\Sigma}$. Consequently, the vector of observations is multivariate normally distributed with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}$. Note that it is assumed that the covariance matrix $\boldsymbol{\Sigma}$ is identical for all vectors of observations (see **Covariance Matrices: Testing Equality of**). Moreover, it is assumed that the vectors of observations are stochastically independent.

To estimate the mean responses, the least squares principle (see **Least Squares Estimation**) can be employed. The least squares estimate, $\hat{\boldsymbol{\mu}}_i$, of the mean response vector $\boldsymbol{\mu}_i$ corresponds to the vector of least squares estimates from separate univariate **analysis of variance** (ANOVA) models, that is, $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}_{i1}, \hat{\mu}_{i2})^T$, with $\hat{\mu}_{i1} = 1/s \sum_{j=1}^s y_{ij1}$ and $\hat{\mu}_{i2} = 1/s \sum_{j=1}^s y_{ij2}$. In the rat example, the least squares estimates of the mean responses are $\hat{\boldsymbol{\mu}}_1 = (7, 7.25)^T$, $\hat{\boldsymbol{\mu}}_2 = (8, 8.5)^T$, and $\hat{\boldsymbol{\mu}}_3 = (14.75, 10.25)^T$.

Also of interest in a MANOVA model is a formal test of zero group difference. This translates into testing the null hypothesis of equal mean responses at each level of the factor, that is,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_r. \quad (2)$$

In univariate one-way ANOVA, a test for equality of group means is derived by partitioning the total sum of squares into a component that is due to the factor in question and an error sum of squares

2 Multivariate Analysis of Variance

(see **Analysis of Variance**). This principle can be generalized to the multivariate case to derive a test for the multivariate hypothesis H_0 . The *total sum-of-squares-and-products matrix* is given by

$$SST = \sum_{i=1}^r \sum_{j=1}^s (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})^T, \quad (3)$$

with $\bar{\mathbf{y}}_{..}$ the vector where each component is the arithmetic mean of all the observations of the corresponding quantitative variable. Note that SST is a symmetric $p \times p$ -dimensional matrix. The total sum-of-squares-and-products matrix SST can be partitioned into the sum-of-squares-and-products matrix according to the factor, say SSA , and the error sum-of-squares-and-products matrix, say SSE . It holds

$$SSA = s \sum_{i=1}^r (\hat{\boldsymbol{\mu}}_i - \bar{\mathbf{y}}_{..})(\hat{\boldsymbol{\mu}}_i - \bar{\mathbf{y}}_{..})^T \quad (4)$$

and

$$SSE = SST - SSA. \quad (5)$$

The $p \times p$ -error sum-of-squares-and-products matrix SSE is used to estimate the unknown covariance matrix $\boldsymbol{\Sigma}$. The estimate is given by SSE divided by the so-called error degrees of freedom $r(s-1)$. Note that the error degrees of freedom in MANOVA are identical to the error degrees of freedom in the univariate analysis of variance.

An equivalent assertion holds for the degrees of freedom that corresponds to the sum-of-squares-and-products matrix of the factor in question, here SSA . The degrees of freedom corresponding to SSA in MANOVA are $(r-1)$, which is the same value as in univariate analysis of variance. The degrees of freedom corresponding to a factor in question are often called *hypothesis degrees of freedom*.

In the rat example, the 2×2 -dimensional matrices SSA and SSE are given as

$$SSA = \begin{pmatrix} 301 & 97.5 \\ 97.5 & 36.333 \end{pmatrix} \text{ and } SSE = \begin{pmatrix} 109.5 & 98.5 \\ 98.5 & 147 \end{pmatrix}. \quad (6)$$

In MANOVA models, there are four statistics for testing H_0 , namely, *Pillai's Trace*, *Wilks' Lambda*, the *Hotelling-Lawley Trace*, and *Roy's Greatest Root*. All four test statistics can be calculated from the

eigenvalues $\xi_1, \xi_2, \dots, \xi_p$ (with $\xi_1 \geq \xi_2 \geq \dots \geq \xi_p$) of the matrix product $\mathbf{M}$ of SSA and SSE^{-1} . Specifically, the test statistics are given by

Pillai's Trace

$$\Lambda_P = \sum_{i=1}^p \frac{\xi_i}{1 + \xi_i},$$

Wilks' Lambda

$$\Lambda_W = \prod_{i=1}^p \frac{1}{1 + \xi_i},$$

Hotelling-Lawley Trace

$$\Lambda_{HL} = \sum_{i=1}^p \xi_i,$$

Roy's Greatest Root

$$\Lambda_R = \xi_1. \quad (7)$$

In the rat example, the eigenvalues of $\mathbf{M} = (SSA)(SSE)^{-1}$ are $\xi_1 = 4.4883$ and $\xi_2 = 0.0498$, so that the values of the four test statistics read as follows

Pillai's Trace

$$\Lambda_P = \sum_{i=1}^2 \frac{\xi_i}{1 + \xi_i} = 0.8652,$$

Wilks' Lambda

$$\Lambda_W = \prod_{i=1}^2 \frac{1}{1 + \xi_i} = 0.1735,$$

Hotelling-Lawley Trace

$$\Lambda_{HL} = \sum_{i=1}^2 \xi_i = 4.5381,$$

Roy's Greatest Root

$$\Lambda_R = \xi_1 = 4.4883. \quad (8)$$

When the number of quantitative variables p is only one – the univariate case – then all four test statistics yield equivalent results since the matrix product $(SSA)(SSE)^{-1}$ reduces to the scalar value SSA/SSE that can be simply transformed to the usual F-statistic in univariate analysis of variance. Also, when the hypothesis degrees of freedom are exactly one, as in a simple comparison of a factor with only

two levels, all four statistics yield equivalent results. For a given number of groups, the critical values of the exact distributions of the four test statistics under H_0 can be found in various textbooks, see [3] or [4]. In software packages, however, the values of the test statistics are transformed into F-values and the distributions of these F-values under H_0 are approximated by F distributions with suitable numerator and denominator degrees of freedom (*see Catalogue of Probability Density Functions*). In case of $p = 2$ variables, the distribution of the transformed Wilks' Lambda is exactly F-distributed. The F-value of Roy's Greatest Root is, in general, an upper bound, so that the reported P value is a lower bound and the test tends to be overly optimistic.

Table 2 shows the results of the four multivariate tests for the rat example. All four tests indicate

that the differences between the three drugs are statistically significant.

The one-way MANOVA model can be extended by including additively further factors. These factors can be nested factors, cross-classified factors, or interaction terms, and they usually reflect further structures in the different populations. As in univariate analysis of variance, the focus would be on determining that there was evidence of any factorial main effects or interactions (*see Analysis of Variance*). On the basis of the eigenvalues of the matrix product of a hypothesis sum-of-squares-and-products matrix according to the factor under consideration and the inverse of the error sum-of-squares-and-products matrix, the four tests described above can be applied for each hypothesis test.

Originally, Morrison [2] used the example to motivate a two-way MANOVA model with interactions.

Table 2 Multivariate tests in the hypothetical one-way MANOVA model

Statistic	Value	F-value	Num DF ^c	Den DF ^d	Approx. P value
Pillai's Trace	0.865	8.006	4	42	<0.0001
Wilks' Lambda	0.174	14.004 ^a	4	40	<0.0001
Hotelling–Lawley Trace	4.538	21.556	4	38	<0.0001
Roy's Greatest Root	4.488	47.127 ^b	2	21	<0.0001

^aF-statistic for Wilks' Lambda is exact.

^bF-statistic for Roy's Greatest Root is an upper bound.

^cNumerator degrees of freedom of the F distribution of the transformed statistic.

^dDenominator degrees of freedom of the F distribution of the transformed statistic.

Table 3 Multivariate tests in the two-way MANOVA model with interactions for the hypothetical example

Factor	Statistic	Value	F-Value	Num DF ^c	Den DF ^d	Approx. P value
Drug	Pillai's Trace	0.880	7.077	4	36	0.0003
	Wilks' Lambda	0.039	12.199 ^a	4	34	<0.0001
	Hotelling–Lawley Trace	4.640	18.558	4	32	<0.0001
	Roy's Greatest Root	4.576	41.184 ^b	2	18	<0.0001
Sex	Pillai's Trace	0.007	0.064 ^a	2	17	0.938
	Wilks' Lambda	0.993	0.064 ^a	2	17	0.938
	Hotelling–Lawley Trace	0.008	0.064 ^a	2	17	0.938
	Roy's Greatest Root	0.008	0.064 ^a	2	17	0.938
Drug $\times$ sex	Pillai's Trace	0.227	1.152	4	36	0.348
	Wilks' Lambda	0.774	1.159 ^a	4	34	0.347
	Hotelling–Lawley Trace	0.290	1.159	4	32	0.346
	Roy's Greatest Root	0.284	2.554 ^b	2	18	0.106

^aF-statistic is exact.

^bF-statistic is an upper bound.

^cNumerator degrees of freedom of the F distribution of the transformed statistic.

^dDenominator degrees of freedom of the F distribution of the transformed statistic.

The first four rats in each drug group in Table 1 are male and the remainder female. The design, therefore, allows assessment of the main effect of the three drugs, the main effect of sex, and the interactions between drug and sex. The interactions between drug and sex would mean that the differences between male and female rat responses varied across the drugs. Table 3 shows the results from the original two-way analysis.

Again, weight-loss differences between the drugs are statistically significant, whereas the main effect of factor sex and the interaction between drug and sex are not statistically significant. Since the comparison between male and female rats has exactly one hypothesis degree of freedom, the four multivariate tests yield equivalent results.

The question that may now arise is which test statistic should be used. One criterion for choosing a test statistic is the power of the tests. Another criterion is the robustness of the tests when assumptions of the basic data distributions, here independent multivariate normally distributed vectors of observation with equal covariance matrix, are violated. Hand and Taylor [1] extensively discuss the choice of the test statistic in MANOVA models. Wilks' Lambda is most popular in the literature and the statistic is a generalized likelihood ratio test statistic. Pillai's Trace is recommended in [1] on the basis of simulation results for power and robustness of the multivariate tests.

In standard (*fixed effects*) MANOVA, one is usually interested in analyzing the influence of the given

levels of the factors on the dependent variables, as described above. But sometimes, the levels of a factor are more appropriately considered a random sample from an (infinite) number of levels. This can be accommodated by treating the effects of the levels as *random effects*. In this case, besides the covariance matrix of the error term, further covariance matrices according to the random effects, also called *multivariate variance components*, contribute to the total covariance matrix of the dependent variables (for more details random effects MANOVA see [5]).

References

- [1] Hand, D.J. & Taylor, C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures*, Chapman & Hall, London.
- [2] Morrison, D.F. (1976). *Multivariate Statistical Methods*, 2nd Edition, McGraw-Hill, New York.
- [3] Rao, C.R. (2002). *Linear Statistical Inference and its Application*, 2nd Edition, John Wiley, New York.
- [4] Seber, G.A.F. (1984). *Multivariate Observations*, John Wiley, New York.
- [5] Timm, N.H. (2002). *Applied Multivariate Analysis*, Springer, New York.

(See also **Multivariate Analysis: Overview; Repeated Measures Analysis of Variance**)

JOACHIM HARTUNG AND GUIDO KNAPP

Multivariate Genetic Analysis

NATHAN A. GILLESPIE AND NICHOLAS G. MARTIN

Volume 3, pp. 1363–1370

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Genetic Analysis

Bivariate Heritability

The genetic and environmental components of covariation can be separated by a technique which is analogous to **factor analysis** [8]. Just as **structural equation modeling (SEM)** can be used to analyze the components of variance influencing a single variable in the case of univariate analysis, SEM can also be used to analyze the sources and structure of covariation underlying multiple variables [8]. When based on genetically informative relatives such as twins, this methodology allows researchers to estimate the extent to which genetic and environmental influences are shared in common by several traits or are trait specific. Information not only comes from the covariance between the variables but also from the cross-twin cross-trait covariances. More precisely, a larger cross-twin cross-trait correlation between monozygotic twin pairs as compared with dizygotic twin pairs suggests that covariance between the variables is partially due to genetic factors. A typical starting point in bivariate and multivariate analysis is the **Cholesky decomposition**. Other methods include common and independent genetic pathway models, as well as genetic simplex models, which will also be discussed.

Multivariate Genetic Analysis

Cholesky Decomposition

The most commonly used multivariate technique in the Classical Twin design (*see Twin Designs*) is the Cholesky decomposition. As shown in Figure 1, the Cholesky is a method of triangular decomposition where the first variable (y_1) is assumed to be caused by a latent factor (*see Latent Variable*) (η_1) that can explain the variance in the remaining variables ($y_2, \dots, y_n$). The second variable (y_2) is assumed to be caused by a second latent factor (η_2) that can explain variance in the second as well as remaining variables ($y_2, \dots, y_n$). This pattern continues until the final observed variable (y_n) is explained by a latent variable (η_n), which is constrained from explaining the variance in any of the previous observed

variables. A Cholesky decomposition is specified for each latent source of variance A, D, C, or E, and as in the univariate case, ACE, ADE, AE, DE, CE, and E models are fitted to the data (*see ACE Model*).

The expected variance–covariance matrix in the Cholesky decomposition is parameterized in terms of n latent factors (where n is the number of variables). All variables load on the first latent factor, $n - 1$ variables load on the second factor and so on, until the final variable loads on the n th latent factor only. Each source of phenotypic variation (i.e., A, C or D, and E) is parameterized in the same way. Therefore, the full factor Cholesky does not distinguish between common factor and specific factor variance and does not estimate a specific factor effect for any variable except the last.

Although symptoms of fatigue and somatic distress are frequently comorbid with anxiety and depressive disorders, a number of studies [6, 7, 14] have demonstrated that a significant proportion of patients with somatic disorders do not meet the criteria for other psychological disorders. As an example of how the Cholesky decomposition can be used

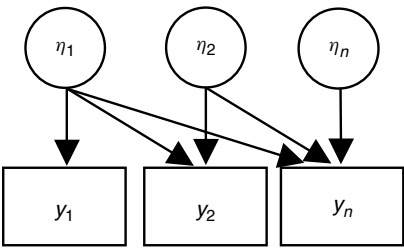


Figure 1 Multivariate Cholesky triangular decomposition, $y_1, \dots, y_n$ = observed phenotypic variables, η_{1-n} = latent factors

Table 1 Phenotypic factor correlations for the factor analytic dimensions of depression, phobic anxiety, and somatic distress. Male correlations appear below the diagonal (reproduced from [3, p. 455])

	Females ($n = 2219$)		
	1	2	3
1 Depression		0.64	0.52
2 Phobic anxiety	0.63		0.58
3 Somatic distress	0.54	0.58	
Males ($n = 1418$)			

2 Multivariate Genetic Analysis

Table 2 Univariate model-fitting for the factor analytic dimensions of depression, phobic anxiety, and somatic distress. The table includes standardized proportions of variance attributable to genetic and environmental effects (reproduced from [4, p. 1056])

A	C	E	$-2LL$	df	$\Delta-2LL$	Δdf	p
Depression							
0.33	0.00	0.67	10 720.59	7987			
0.33		0.67	10 720.59	7988	0.05	1	.82
	0.24	0.76	10 729.87	7988	9.28	1	^b
		1.00	10 791.02	7989	70.44	2	^c
Phobic Anxiety							
0.37	0.03	0.59	7968.50	7979			
0.41		0.59	7968.62	7980	0.11	1	.74
	0.30	0.70	7976.54	7980	8.04	1	^b
		1.00	8051.14	7981	82.64	2	^c
Somatic Distress							
0.11	0.17	0.72	9563.28	7969			
0.32		0.68	9566.50	7970	3.22	1	.07
	0.25	0.75	9564.08	7970	0.79	1	.37
		1.00	9624.57	7971	61.29	2	^c

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.

Results based on Maximum Likelihood.

$\Delta-2LL = -2 \log\text{-likelihood}$.

^a $p < 0.05$,

^b $p < 0.01$,

^c $p < 0.001$.

to resolve these sorts of questions, we administered self-report measures of anxiety, depression, and somatic distress to a community-based sample of 3469 Australian twin individuals aged 18 to 28 years. As shown in Table 1, the phenotypic correlations between somatic distress, depression, and phobic anxiety, for males and females alike, are all high.

The results for the univariate genetic analyses in Table 2 reveal that an additive genetic (*see Additive Genetic Variance*) and **nonshared environmental effects** model best explains individual differences in depression and phobic anxiety scores, for male and female twins alike. The same could not be said for somatic distress because there is insufficient power to choose between additive genetic or shared environment effects as the source of familial aggregation in somatic distress. This limitation can be overcome using multivariate genetic analysis, which has greater power to detect genetic and environmental effects by making use of all the covariance terms between variables. Moreover, it will allow us to determine whether somatic distress is etiologically distinct from self-report measures of depression and anxiety.

As shown in Table 3, an additive genetic and nonshared environment (AE) model best explained the sources of covariation between the three factors. This is illustrated in Figure 2, where 33% (i.e., $0.32^2/0.43^2 + 0.15^2 + 0.32^2$) of the genetic variance in somatic distress is due to specific gene action unrelated to depression or phobic anxiety. In addition, 74% of the individual environmental influence on somatic distress is unrelated to depression and phobic anxiety. These results support previous findings that somatic symptoms are partly etiologically distinct, both genetically and environmentally from the symptoms of anxiety and depression.

Common and Independent Genetic Pathway Models

Alternate multivariate methods can be used to estimate common factor and specific factor variance (*see Factor Analysis: Exploratory*). For instance, the common pathway model in Figure 3a assumes that the genetic and environmental effects (A, C and E) (*see ACE Model*) contribute to one or more latent

Table 3 Multivariate Cholesky decomposition model-fitting results. Results are based on combined male and female data adjusted for sex differences in the prevalence of depression, phobic anxiety, and somatic distress (reproduced from [4, p. 1056])

Model			$-2LL$	df	$\Delta-2LL$	Δdf	p
A	C	E	26 000.29	15 285			
A		E	26 003.82	15 291	3.53	6	.74
	C	E	26 015.20	15 291	14.91	6	^a
		E	26 147.97	15 297	147.68	12	^c

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.

Results based on Maximum Likelihood.

$\Delta-2LL = -2 \log\text{-likelihood}$.

^a $p < .05$,

^b $p < .01$,

^c $p < .001$.

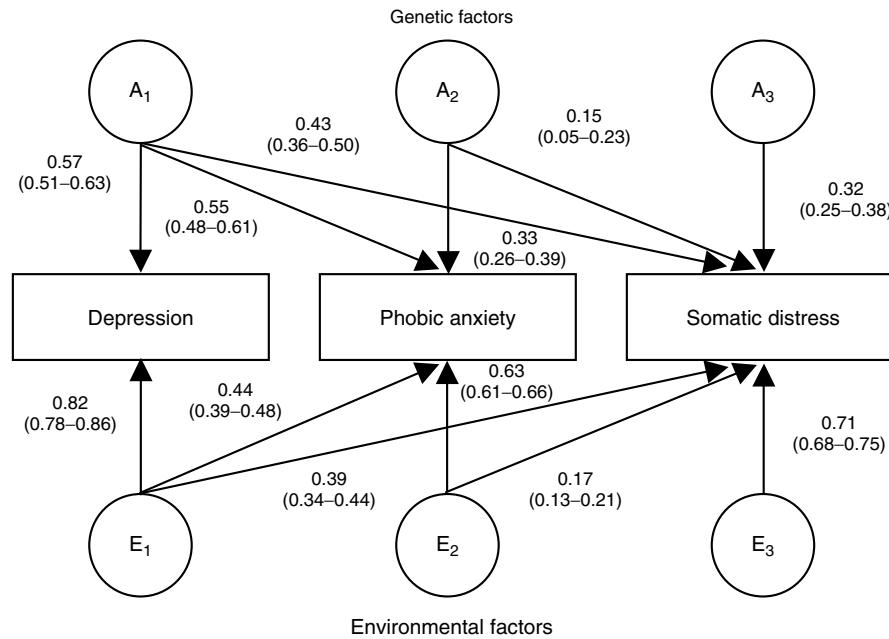


Figure 2 Path diagram showing standardized path coefficients and 95% confidence intervals for the latent genetic (A_1 to A_3) and environmental (E_1 to E_3) effect (reproduced from [4] p. 1057)

intervening variables (η), which in turn are responsible for the observed patterns of covariance between symptoms ($y_1, \dots, y_n$).

This is in contrast to the independent pathway model in Figure 3b, which predicts that genes and environment have different effects on the covariance between symptoms. Because it can be shown algebraically that the common pathway is nested within the independent pathway model, the two models can

be compared using a likelihood ratio chi-squared statistic (see **Goodness of Fit**).

Parker's 25-item Parental Bonding Instrument (PBI) [13] was designed to measure maternal and paternal parenting along the dimensions of Care and Overprotection [11, 12]. However, factor analysis of the short 14-item version based on 4514 females, aged 18 to 45, has yielded three correlated factors: Autonomy, Coldness, and Overprotection [5].

4 Multivariate Genetic Analysis

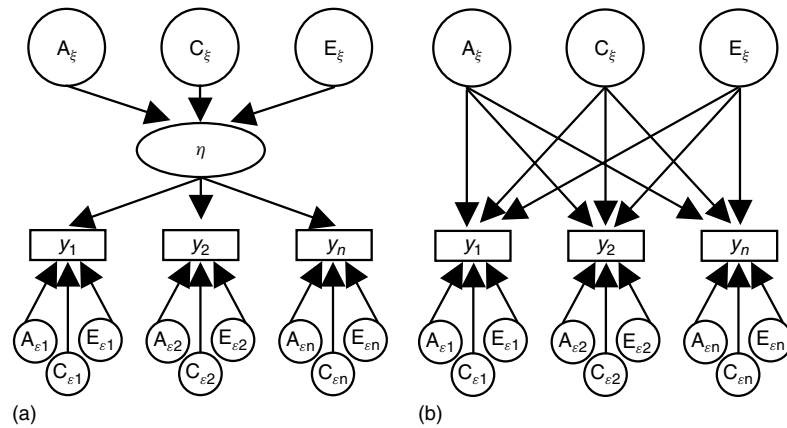


Figure 3 Common (a) and independent (b) genetic pathway models. η = common factor, $y_1, \dots, y_n$ = observed phenotypic variables, A_ξ , C_ξ & E_ξ = latent genetic & environmental factors, $A_{\varepsilon 1-n}$, $C_{\varepsilon 1-n}$ & $E_{\varepsilon 1-n}$ = latent genetic & environmental residual factors

Table 4 Best fitting univariate models with standardized proportions of variance attributable to genetic and environmental variance (reproduced from [5, p. 390])

PBI dimensions	A	C	E	$-2LL$	df
Coldness	.61	–	.39	5979.01	3603
Overprotection	.22	.24	.54	7438.84	3602
Autonomy	.33	.17	.51	9216.17	3599

Note: A, C & E = additive genetic, shared/common environment, and nonshared environment variance.

Results based on Maximum Likelihood.

$\Delta -2LL = -2 \log\text{-likelihood}$.

^a $p < .05$,

^b $p < .01$,

^c $p < .001$.

Table 5 Comparison of the common and independent pathway models for the PBI dimensions of Autonomy, Coldness, and Overprotection [5]

Model	χ^2 ^a	df	p	AIC
Independent pathway	9.17	15	.87	–20.83
Common pathway	13.39	19	.82	–24.61

Note: Results based on Weighted Least Squares.

AIC = Akaike Information Criterion.

Univariate analyses of the three dimensions, which are summarized in Table 4, reveal that variation in parental Overprotection and Autonomy can be best explained by additive genetic, shared, and nonshared environmental effects, whereas the best fitting model for Coldness includes additive genetic and non-shared environmental effects. As is shown in Table 5, when compared to an independent pathway model,

a common pathway genetic model provided a more parsimonious fit to the three PBI dimensions. The common pathway model is illustrated in Figure 4.

Genetic Simplex Modeling

When genetically informative longitudinal data are available, a multivariate Cholesky can again be

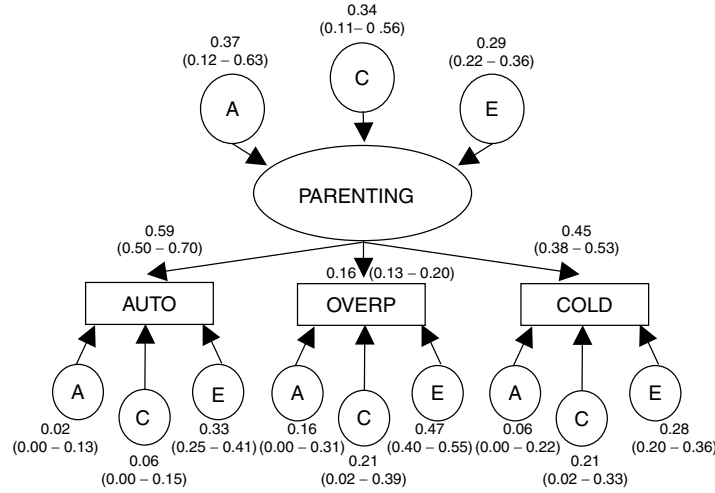


Figure 4 Common pathway genetic model (saturated) for the PBI dimensions with standardized proportions of variance and 95% confidence intervals (reproduced from 5, p. 391). AUTO = Autonomy, OVERP = Overprotection, COLD = Coldness, Results based on Weighted Least Squares

fitted to determine the extent to which genetic and environmental influences are shared in common by a trait measured at different time points. However, this approach is limited in so far as it does not take full advantage of the time-series nature of the data, that is, that causation is unidirectional through time [1].

One solution is to fit a simplex model, which explicitly takes into account the longitudinal nature of the data. As shown in Figure 5, simplex models are autoregressive, whereby the genetic and environmental latent variables at time i are causally related to the immediately preceding latent variables (η_{i-1}).

Eta (η_i) is a latent variable (i.e., A, C, E, or D) at time i , β_i is the regression of the latent factor on

the immediately preceding latent factor η_{i-1} , and ζ_i is the new input or innovation at time i . When using data from MZ and DZ twin pairs, structural equations can be specified for additive genetic sources of variation (A), common environmental (C), nonadditive genetic sources of variation such as dominance or epistasis (D), and unique environmental sources of variation (E).

Because measurement error does not influence observed variables at subsequent time points, simplex designs therefore permit discrimination between transient factors effecting measurement at one time point only, and factors that are continuously present or exert a long-term influence throughout the time series [1, 9]. Although denoted as error variance, the error parameters will also include variance attributable to short-term nonshared environmental effects.

We have used this model-fitting approach to investigate the stability and magnitude of genetic and environmental effects underlying major dimensions of adolescent personality across time [2]. The junior eyesenck personality questionnaire (JEPQ) was administered to over 540 twin pairs at ages 12, 14, and 16 years. Results for JEPQ Neuroticism are presented here.

The additive genetic factor correlations based on a Cholesky decomposition are shown in Table 6. These reveal that the latent additive genetic factors are

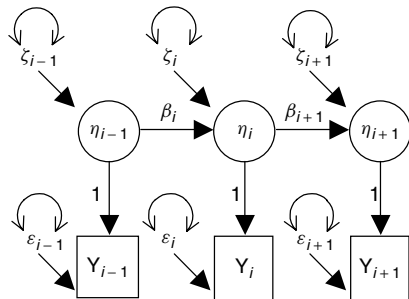


Figure 5 General simplex model. β = regression of the latent factor on the previous latent factor ζ = new input or innovation at time i η = latent variable at time i

6 Multivariate Genetic Analysis

Table 6 Additive genetic (above diagonal) and nonshared environmental latent factor correlations for JEPQ Neuroticism (reproduced from [2])

	Females			Males		
	1	2	3	1	2	3
1 12 years		0.76	0.68		0.86	0.79
2 14 years	0.40		0.94	0.24		0.74
3 16 years	0.36	0.41		0.12	0.53	

Table 7 Multivariate model-fitting results for JEPQ Neuroticism based on twins aged 12, 14, and 16 years (reproduced from [2])

	Neuroticism									
	Females					Males				
	$-2LL$	df	$\Delta 2LL$	Δdf	P	$-2LL$	df	$\Delta 2LL$	Δdf	p
Cholesky										
ACE	10424.88	1803				10016.48	1753			
Simplex										
ACE	10424.95	1805	0.07	2	.96	10016.88	1755	0.40	2	.82
AE	10425.34	1810	0.39	5	1.00	10021.26	1760	4.38	5	.50
Drop ζ_{a3}	10426.57	1811	1.23	1	.27	10058.59	1761	37.32	1	^c
CE	10432.45	1810	7.50	5	.19	10032.32	1760	15.44	5	^b
E	10663.09	1815	238.14	10	^c	10650.99	1765	634.11	10	^c

Note: Results based on Maximum Likelihood.

ζ_{a3} = Genetic innovation at time 3.

Best fitting models in bold.

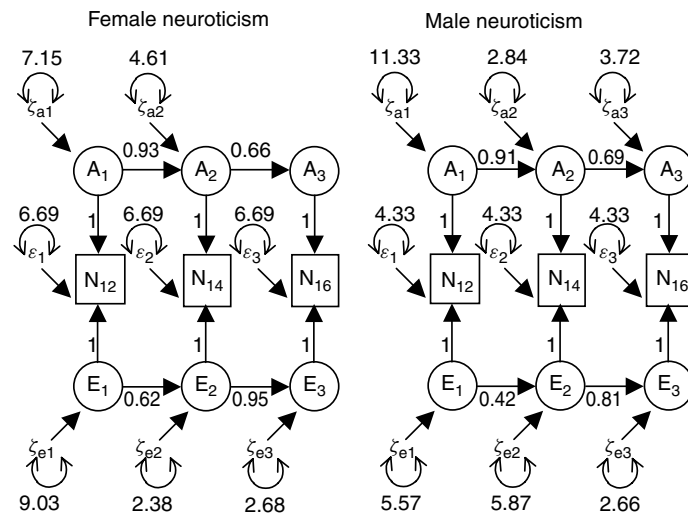


Figure 6 Best fitting genetic simplex model for female and male Neuroticism (reproduced from [2]). N_{12-16} = Neuroticism 12 to 16 yrs, A_{1-3} , E_{1-3} = additive genetic & nonshared environmental effects, ζ_{a1-3} , ζ_{e1-3} = additive genetic innovations & nonshared environmental innovations, ε_{1-3} = error parameters 12 to 16 yrs double/single headed arrows = variance components/path coefficients

highly correlated. This is consistent with a pleiotropic model of gene action, whereby the same genes explain variation across different time points. As shown in Table 7, the fit of the ACE simplex models provided a better explanation of the Neuroticism time-series data, in so far as the fit was no worse than the corresponding Cholesky decompositions. The final best-fitting AE simplex models for male and female Neuroticism are shown in Figure 6.

It is difficult to imagine that genetic variation in personality is completely determined by age 12. As shown in Figure 6, smaller genetic innovations are observed for male Neuroticism at 14 and 16, as well as female Neuroticism at 14. These smaller genetic innovations potentially hint at age-specific genetic effects related to developmental or hormonal changes during puberty and psychosexual development.

When data are limited to three time points, a common genetic factor model will also provide a comparable fit when compared to the genetic simplex model. Other possible modeling strategies include biometric growth models (see [10]). Despite these limitations, time-series data even when based on three time points still provides an opportunity to test explicit hypotheses of genetic continuity. Moreover, the same data are ideal for fitting univariate and multivariate linkage models to detect quantitative trait loci of significant effect.

The above sections have provided an introduction to bivariate and multivariate analyses and how these methods can be used to estimate the genetic and environmental covariance between phenotypic measures. This should give the reader an appreciation for the flexibility of SEM approaches to address more complicated questions beyond univariate decompositions. For a more detailed treatment of this subject, see [9].

References

- [1] Boomsma, D.I., Martin, N.G. & Molenaar, P.C. (1989). Factor and simplex models for repeated measures: application to two psychomotor measures of alcohol sensitivity in twins, *Behavior Genetics* **19**, 79–96.
- [2] Gillespie, N.A., Evans, D.E., Wright, M.M. & Martin, N.G. (submitted). Genetic simplex modeling of Eysenck's dimensions of personality in a sample of young Australian twins, *Journal of Child Psychology and Psychiatry*.
- [3] Gillespie, N., Kirk, K.M., Heath, A.C., Martin, N.G. & Hickie, I. (1999). Somatic distress as a distinct psychological dimension, *Social Psychiatry and Psychiatric Epidemiology* **34**, 451–458.
- [4] Gillespie, N.G., Zhu, G., Heath, A.C., Hickie, I.B. & Martin, N.G. (2000). The genetic aetiology of somatic distress, *Psychological Medicine* **30**, 1051–1061.
- [5] Gillespie, N.G., Zhu, G., Neale, M.C., Heath, A.C. & Martin, N.G. (2003). Direction of causation modeling between measures of distress and parental bonding, *Behavior Genetics* **33**, 383–396.
- [6] Hickie, I.B., Lloyd, A., Wakefield, D. & Parker, G. (1990). The psychiatric status of patients with chronic fatigue syndrome, *British Journal of Psychiatry* **156**, 534–540.
- [7] Kroenke, K., Wood, D.R., Mangelsdorff, D., Meier, N.J. & Powell, J.B. (1988). Chronic fatigue in primary care, *Journal of the American Medical Association* **260**, 926–934.
- [8] Martin, N.G. & Eaves, L.J. (1977). The genetical analysis of covariance structure, *Heredity* **38**, 79–95.
- [9] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.
- [10] Neale, M.C. & McArdle, J.J. (2000). Structured latent growth curves for twin data, *Twin Research* **3**, 165–177.
- [11] Parker, G. (1989). The parental bonding instrument: psychometric properties reviewed, *Psychiatric Developments* **4**, 317–335.
- [12] Parker, G. (1990). The parental bonding instrument. A decade of research, *Social Psychiatry and Psychiatric Epidemiology* **25**, 281–282.
- [13] Parker, G., Tupling, H. & Brown, L.B. (1979). A parental bonding instrument, *British Journal of Medical Psychology* **52**, 1–10.
- [14] Wessely, S. & Powell, R. (1989). Fatigue syndromes: a comparison of chronic postviral with neuromuscular and affective disorders, *Journal of Neurology, Neurosurgery, and Psychiatry* **52**, 940–948.

NATHAN A. GILLESPIE AND NICHOLAS
G. MARTIN

Multivariate Multiple Regression

JOACHIM HARTUNG AND GUIDO KNAPP

Volume 3, pp. 1370–1373

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Multiple Regression

The multivariate linear model is used to explain and to analyze the relationship between one or more explanatory variables and $p > 1$ quantitative dependent or response variables that have been observed at n subjects. In case all the explanatory variables are qualitative, the multivariate linear model is called the **Multivariate Analysis of Variance (MANOVA) model**. When all the explanatory variables are quantitative, that is a multivariate system of quantitative variables is given in which the relationships between p dependent quantitative variables and, say q , independent quantitative are of interest, then the model is referred to as a *multivariate regression model*. Multivariate regression analysis is used to investigate the relationships when the p dependent variables are correlated. In contrast, when the dependent variables are uncorrelated, relationships can be assessed by carrying out p univariate regression analyses (*see Regression Models*). Often, there is an implied predictive aim in the investigation, and the formulation of appropriate and parsimonious relationships among the variables is a necessary prerequisite.

Consider the small example given in [2]. The data in Table 1 show the four measurements: chest circumference (CC), midupper arm circumference (MUAC), height and age (in months) for a sample on nine girls. One practical objective would be to develop a predictive model for CC and MUAC from knowledge of height and age.

The dependent variables CC and MUAC are highly correlated with each other and Pearson's correlation coefficient is 0.77, so they should be incorporated in a single multivariate regression model for maximum efficiency as multiple regression analyses for each variable separately will ignore this correlation in the construction of hypothesis tests or confidence intervals.

Let y_{i1} denote the CC of the i th girl, y_{i2} the MUAC, x_{i1} the height, and x_{i2} the age; then, the univariate regression models for each variable are

$$y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2} + e_{i1},$$

$$i = 1, \dots, n = 9, \quad (1)$$

and

$$y_{i2} = \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2} + e_{i2},$$

$$i = 1, \dots, n = 9. \quad (2)$$

In the multivariate case, the observations y_{i1} and y_{i2} are put in a row vector so that the model has the form

$$(y_{i1} \ y_{i2}) = (1 \ x_{i1} \ x_{i2}) \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} + (e_{i1} \ e_{i2}), \quad i = 1, \dots, n = 9. \quad (3)$$

In general, the model for the observed data can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (4)$$

where $\mathbf{Y}$ is the $n \times p$ matrix whose i th row contains the values of the dependent variables for the i th subject, $\mathbf{X}$ is the $n \times (q + 1)$ -matrix, also called the

Table 1 Four measurements taken on each of nine young girls

	Chest circumference (cm)	Midupper arm circumference (cm)	Height (cm)	Age (months)
Individual	y_1	y_2	x_1	x_2
1	58.4	14.0	80	21
2	59.2	15.0	75	27
3	60.3	15.0	78	27
4	57.4	13.0	75	22
5	59.5	14.0	79	26
6	58.1	14.5	78	26
7	58.0	12.5	75	23
8	55.5	11.0	64	22
9	59.2	12.5	80	22

2 Multivariate Multiple Regression

regression matrix, whose i th row contains a one and the values of the explanatory variables for the i th subject, $\mathbf{B}$ is the $(q + 1) \times p$ -matrix of unknown regression parameters, and $\mathbf{E}$ is an $n \times p$ -matrix of random variables whose rows are independent observations from a multivariate normal distribution with mean zero and **covariance matrix** $\mathbf{\Sigma}$. Note that it is assumed that the covariance matrix is identical for each row of $\mathbf{E}$. The assumption of multivariate normality is only necessary when tests or confidence regions have to be constructed for the parameters or the predicted values. For other aspects, it is sufficient to assume that the rows of $\mathbf{E}$ are uncorrelated and have mean zero and covariance matrix $\mathbf{\Sigma}$.

The first step in multivariate regression analysis is to fit a model to the observed data. This requires the estimation of the unknown parameters in $\mathbf{B}$ and $\mathbf{\Sigma}$, which can be done by **maximum likelihood** when normality is assumed or by the least-squares approach (*see Least Squares Estimation*) when no distributional assumptions are made. For the parameter matrix $\mathbf{B}$, both approaches lead to the same estimate so that the actual assumptions are not critical to the outcome of the analysis. For standard application of the theory, it is necessary that the number of subjects n is greater than $p + q$, the total number of quantitative variables. Furthermore, the matrix $\mathbf{X}$ has to be of full rank $(q + 1)$ so that the inverse $(\mathbf{X}^T \mathbf{X})^{-1}$ exists. When these conditions are met, the estimator $\hat{\mathbf{B}}$ of the parameter matrix $\mathbf{B}$ is given by

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (5)$$

Note that the regression matrix $\mathbf{X}$ is identical in all the corresponding univariate regressions analyses. Consequently, the columns of the estimated parameter matrix $\hat{\mathbf{B}}$ are the estimators that would be obtained from the corresponding univariate regression analyses.

The estimator of the covariance matrix $\mathbf{\Sigma}$ is built from the matrix of the residuals

$$\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}. \quad (6)$$

The maximum likelihood estimator of $\mathbf{\Sigma}$ is given as $\hat{\mathbf{\Sigma}}_{\text{ML}} = \hat{\mathbf{E}}^T \hat{\mathbf{E}}/n$, while an unbiased estimator of $\mathbf{\Sigma}$ is $\hat{\mathbf{\Sigma}}_{\text{unbiased}} = \hat{\mathbf{E}}^T \hat{\mathbf{E}}/(n - q - 1)$. Details of the maximum likelihood derivation can be found in [3] while [4] covers the least-squares approach.

In the example, the estimated parameter matrix is given by

$$\hat{\mathbf{B}} = \begin{pmatrix} 36.7774 & -5.5258 \\ 0.2042 & 0.1405 \\ 0.2545 & 0.3477 \end{pmatrix}. \quad (7)$$

The residual sum-of-squares-and-products matrix is given by

$$\hat{\mathbf{E}}^T \hat{\mathbf{E}} = \begin{pmatrix} 2.4049 & -0.3558 \\ -0.3558 & 2.8713 \end{pmatrix}, \quad (8)$$

so that the maximum likelihood estimate of $\mathbf{\Sigma}$ has the form

$$\hat{\mathbf{\Sigma}}_{\text{ML}} = \frac{1}{9} \hat{\mathbf{E}}^T \hat{\mathbf{E}} = \begin{pmatrix} 0.2672 & -0.0395 \\ -0.0395 & 0.3190 \end{pmatrix}, \quad (9)$$

while the unbiased estimator of $\mathbf{\Sigma}$ is

$$\hat{\mathbf{\Sigma}}_{\text{unbiased}} = \frac{1}{6} \hat{\mathbf{E}}^T \hat{\mathbf{E}} = \begin{pmatrix} 0.4008 & -0.0593 \\ -0.0593 & 0.4786 \end{pmatrix}. \quad (10)$$

When the above assumptions of the general multivariate regression model are fulfilled, inference about the model parameters is possible. In analogy to univariate regression, the first test of interest is of the significance of the regression: do the explanatory variables help at all in predicting the dependent variables, or would simple prediction from the unconditional mean of $\mathbf{Y}$ be just as accurate? Formally, the null hypothesis can be formulated that all the parameters of $\mathbf{B}$ that multiply any of the explanatory variables are zero. Under this null hypothesis, the mean of $\mathbf{Y}$ is $\mathbf{1}\boldsymbol{\mu}^T$ where $\boldsymbol{\mu}$ is the population mean vector of the dependent variables and $\mathbf{1}$ is an n -vector of ones. The maximum likelihood estimator of $\boldsymbol{\mu}$ is $\bar{\mathbf{y}}$, the sample mean vector of the dependent variables. Then, the maximum likelihood estimate of $\mathbf{\Sigma}$ under the null hypothesis can be written as $n\hat{\mathbf{\Sigma}} = \mathbf{Y}^T \mathbf{Y} - n\bar{\mathbf{y}} \bar{\mathbf{y}}^T$. This *total sum-of-squares-and-products matrix* can be decomposed as

$$(\mathbf{Y}^T \mathbf{Y} - n\bar{\mathbf{y}} \bar{\mathbf{y}}^T) = (\hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - n\bar{\mathbf{y}} \bar{\mathbf{y}}^T) + \hat{\mathbf{E}}^T \hat{\mathbf{E}}. \quad (11)$$

Writing $\mathbf{H} = \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - n\bar{\mathbf{y}} \bar{\mathbf{y}}^T$, the decomposition of the total sum-of-squares-and-products matrix can be summarized in a table just like in MANOVA models (see Table 2).

The null hypothesis of no relationship between the explanatory and the dependent variables can be tested by any of the four multivariate test statistics,

Table 2 MANOVA table for multivariate regression

MANOVA Source	Sum-of-squares-and-products matrix	Degrees of freedom
Multivariate regression	$\mathbf{H}$	q
Residual	$\hat{\mathbf{E}}^T \hat{\mathbf{E}}$	$n - q - 1$
Total	$\mathbf{Y}^T \mathbf{Y} - n \bar{\mathbf{y}} \bar{\mathbf{y}}^T$	$n - 1$

namely, *Pillai's Trace*, *Wilks' Lambda*, the *Hotelling-Lawley Trace*, and *Roy's Greatest Root*. All the test statistics are based on the eigenvalues of the matrix product $(\hat{\mathbf{E}}^T \hat{\mathbf{E}})^{-1} \mathbf{H}$. (For more details see **Multivariate Analysis of Variance**).

Having established that a relationship does exist between the dependent variables and the explanatory variables for predictive purposes, the next objective often is to derive the most parsimonious relationship, that is, the one that has good predictive accuracy while involving the fewest number of explanatory variables. In this context, the null hypothesis to test is that a specified subset of explanatory variables provides no additional predictive information when the remaining explanatory variables are included in the model. Suppose that the null hypothesis is that the last q_2 explanatory variables have no predictive information when the first q_1 explanatory variables are included in the model, $q_1 + q_2 = q$. Then, the sum-of-squares-and-products matrix $\mathbf{H}$ can be partitioned in $\mathbf{H} = \mathbf{H}_{(q_1)} + \mathbf{H}_{(q_2|q_1)}$. The matrix $\mathbf{H}_{(q_1)}$ is given by $\mathbf{H}_{(q_1)} = \mathbf{T} - \tilde{\mathbf{E}}^T \tilde{\mathbf{E}}$, where $\tilde{\mathbf{E}}$ is the residual matrix when only the first q_1 explanatory variables are included in the multivariate regression,

$\mathbf{T}$ is the total sum-of-squares-and-product matrix, and $\mathbf{H}_{(q_2|q_1)} = \mathbf{H} - \mathbf{H}_{(q_1)}$. The tests statistics are then based on the eigenvalues of the matrix product $(\hat{\mathbf{E}}^T \hat{\mathbf{E}})^{-1} \mathbf{H}_{(q_2|q_1)}$.

In statistical software packages, usually the test of the null hypothesis that a specified explanatory variable has no predictive information when all the other explanatory variables are included in the multivariate regression model is reported by default. Since the hypothesis degrees of freedom are exactly one in this situation, the four multivariate tests yield equivalent results (for details see **Multivariate Analysis of Variance**).

Applying the multivariate tests to the example, the results put together in Table 3 are obtained. The analysis indicates that both explanatory variables height and age significantly affect the two dependent variables after adjusting for the effect of the other variable.

The fit of a model to a set of data involves a number of assumptions. These assumptions can relate either to the systematic part of the model, that is, to the form of the relationship between explanatory and dependent variables, or to the random part of the model, that is, to the distribution of the dependent variables. In a multivariate regression model, the matrix of residuals $\hat{\mathbf{E}}$ contains information that can be used to assess the adequacy of the model. If the model is correct and all distributional assumptions are valid, then the rows of $\hat{\mathbf{E}}$ will look approximately like independent observations from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ . A detailed discussion of procedures for checking model adequacy is provided in [1] and [2].

Table 3 Multivariate tests on the predictive information of height and age in the example

Variable	Statistic	Value	F Value	Num DF ^b	Den DF ^c	Approx. <i>P</i> value
Height	Pillai's Trace	0.839	12.981 ^a	2	5	0.0105
	Wilks' Lambda	0.161	12.981 ^a	2	5	0.0105
	Hotelling-Lawley Trace	5.193	12.981 ^a	2	5	0.0105
	Roy's Greatest Root	5.193	12.981 ^a	2	5	0.0105
Age	Pillai's Trace	0.785	9.148 ^a	2	5	0.0213
	Wilks' Lambda	0.215	9.148 ^a	2	5	0.0213
	Hotelling-Lawley Trace	3.659	9.148 ^a	2	5	0.0213
	Roy's Greatest Root	3.659	9.148 ^a	2	5	0.0213

^aF statistic is exact.

^bNumerator degrees of freedom of the F distribution of the transformed statistic.

^cDenominator degrees of freedom of the F distribution of the transformed statistic.

4 Multivariate Multiple Regression

References

- [1] Gnanadesikan, R. (1997). *Methods For Statistical Data Analysis of Multivariate Observations*, 2nd Edition, John Wiley, New York.
- [2] Krzanowski, W.J. (2000). *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford.
- [3] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1995). *Multivariate Analysis*, Academic Press, London.
- [4] Rao, C.R. (2002). *Linear Statistical Inference and its Application*, 2nd Edition, John Wiley, New York.

(See also **Multivariate Analysis: Overview**)

JOACHIM HARTUNG AND GUIDO KNAPP

Multivariate Normality Tests

H.J. KESELMAN

Volume 3, pp. 1373–1379

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Normality Tests

Behavioral science researchers frequently gather multiple measures in their research endeavors. That is, for many areas of investigation in order to capture the phenomenon under investigation numerous measures are gathered. For example, Harasym et al. [6] obtained scores on multiple personal traits using the Myers-Briggs inventory for nursing students with three different styles of learning. Knapp and Miller [11] discuss the measurement of multiple dimensions of healthcare quality as part of a system-wide evaluation program. Indeed, as Tabachnick and Fidell [29, p. 3] note, ‘Experimental research designs with multiple DVs (dependent variables) were unusual at one time. Now, however, with attempts to make experimental designs more realistic, . . . , experiments often have several DVs. It is dangerous to run an experiment with only one DV and risk missing the impact of the IV (independent variable) because the most sensitive DV is not measured’.

The assumption of multivariate normality (*see Catalogue of Probability Density Functions*) underlies many multivariate techniques, for example, **multivariate analysis of variance**, **discriminant analysis**, **canonical correlation** and maximum likelihood factor analysis (*see Factor Analysis: Exploratory*), to name but a few. That is to say, many areas of **multivariate analysis** involve parameter estimation and tests of significance and the validity of these procedures require that the data follow a multivariate normal distribution. As Timm [30, p. 118] notes, ‘The study of robust estimation for location and dispersion of model parameters, the identification of **outliers**, the analysis of multivariate residuals, and the assessment of the effects of model assumptions on tests of significance and power are as important in multivariate analysis as they are in univariate analysis.’ In the remainder of this paper, I review methods for assessing the validity of the multivariate normality assumption.

Multivariate Normality

Multivariate normality is most easily described by generalizing univariate normality. In particular, the

normal density function for a random variable y , having mean μ and variance σ^2 is

$$f(y) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}. \quad (1)$$

Notationally, when y has the density given in (1) we can write $y \sim N(\mu, \sigma^2)$. As is well known, the density function is depicted by a bell-shaped curve. Now if $\mathbf{y}$ has a multivariate normal distribution with mean μ and covariance Σ , its density is given as

$$g(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^p} \sqrt{|\Sigma|}} e^{-(\mathbf{y}-\mu)' \Sigma^{-1} (\mathbf{y}-\mu)/2}, \quad (2)$$

where p is the number of variables, $'$ is the transpose operator and $^{-1}$ is the inverse function. Again, notationally, we can say that when $\mathbf{y}$ has the density given in (2) that $\mathbf{y}$ is $\sim N_p(\mu, \Sigma)$. Note that the $p \times p$ population covariance matrix, Σ , is the generalization of σ^2 and the $p \times 1$ mean vector μ is the generalization of μ . The graph of a particular case of (2) is a $(p+1)$ -dimensional bell-shaped surface.

Another correspondence between multivariate and univariate normality is important to note. In the univariate case, the expression $(y - \mu)^2/\sigma^2 = (y - \mu)(\sigma^2)^{-1}(y - \mu)$ in the exponent of (1) is a distance statistic, that is, it measures the squared distance from y to μ in σ^2 units. Similarly, $(\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu)$ in the exponent of (2) is the squared generalized distance from $\mathbf{y}$ to μ . This squared distance is also referred to as **Mahalanobis's distance statistic**, and a sample estimate (D^2) can be obtained from

$$D^2 = (\mathbf{y} - \bar{\mathbf{y}})' S^{-1} (\mathbf{y} - \bar{\mathbf{y}}), \quad (3)$$

where $\bar{\mathbf{y}}$ is the $p \times 1$ vector of sample means and S is the $p \times p$ Sample covariance matrix (see [8], p. 55).

Properties of Multivariate Normal Distributions

The multivariate normal distribution has a number of important properties that bare on how one can assess whether data are indeed multivariate normally distributed.

- Linear combinations of the variables in $\mathbf{y}$ are themselves normally distributed.
- All marginal distributions for any subset of $\mathbf{y}$ are normally distributed. This fact means that

2 Multivariate Normality Tests

each variable in the set is normally distributed. That is, multivariate normality implies univariate normality for each individual variable in the set $\mathbf{y}$. However, the converse is not true; univariate normality of the individual variables does not imply multivariate normality.

- All conditional distributions (say $\mathbf{y}_1|\mathbf{y}_2$) are normally distributed.

Assessing Multivariate Normality

According to Timm [30, Section 3.7, p. 118], ‘The two most important problems in multivariate data analysis are the detection of outliers and the evaluation of multivariate normality’. Timm presents a very detailed strategy for detecting outlying values and for assessing multivariate normality. The steps he (as well as others) suggests are:

1. Evaluate each individual variable for univariate normality. A number of procedures can be followed for assessing univariate normality, such as checking the skewness and kurtosis measures for each individual variable (see [4, 13, 17], Chapter 4). As well, one can use a global test such as the Shapiro and Wilk [23] test. Statistical packages usually provide these assessments, for example, SAS’s [20] UNIVARIATE procedure (see **Software for Statistical Analyses**).
2. Use normal **probability plots** (Q-Q plots) for each individual variable, plots that compare each sample distribution to population quantiles of the normal distribution. Rencher [17] indicates that many researchers find this technique to be most satisfactory as a visual check for nonnormality.
3. Transform nonnormal variables. Researchers can use Box and Cox’s [3] power transformation. Timm [30, p. 130] and Timm and Mieczkowski [31, pp. 34–35] provide a SAS program for the Box and Cox power transformation (see **Transformation**).
4. Locate outlying values (see the SAS programs in Timm and Mieczkowski [31, pp. 26–31]. If outlying values are present they must be dealt with (see [30, p. 119]). See the references in [33] where they describe a number of (multivariate) outlier detection procedures.

These four steps are for evaluating normality and outlying values with respect to marginal (univariate)

normality. However, as previously indicated, univariate normality is not a sufficient condition for multivariate normality. As Looney [13, p. 64] notes, ‘Even if UVN (univariate normality) is tenable for each of the variables individually, however, this does not necessarily imply MVN (multivariate normality) since there are non-MVN distributions that have UVN marginals.’ Nonetheless, some authors believe that the results from assessing univariate normality are good indications of whether multivariate normality exists or not. For example, Gnanadesikan [5, p. 168] indicates ‘In practice, except for rare or pathological examples, the presence of joint (multivariate) normality is likely to be detected quite often by methods directed at studying the marginal (univariate) normality of the observations on each variable’. This view was also noted by Johnson and Wichern [10, p. 153] – ‘... for most practical work, one-dimensional and two-dimensional investigations are ordinarily sufficient. Fortunately, pathological data sets that are normal in lower dimensional representation but nonnormal in higher dimensions are not frequently encountered in practice’. These viewpoints suggest that it is sufficient to examine marginal and bivariate normality in order to assess multivariate normality. Procedures, however, are available for directly assessing multivariate normality.

Romeu and Ozturk [18] compared goodness-of-fit tests for multivariate normality and their results indicate that multivariate tests of **skewness** and **kurtosis** proposed by Mardia [14, 16] are most reliable for assessing multivariate normality. The skewness and kurtosis measures are defined respectively, as

$$\beta_{1,p} = E[(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})]^3 \quad (4)$$

and

$$\beta_{2,p} = E[(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})]^4. \quad (5)$$

The sample estimators, respectively, are

$$\hat{\beta}_{1,p} = \sum_{i=1}^n \sum_{j=1}^n \frac{[(\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_j - \bar{\mathbf{y}})]^3}{n^2} \quad (6)$$

and

$$\hat{\beta}_{2,p} = \sum_{i=1}^n \frac{[(\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})]^4}{n}. \quad (7)$$

The sampling distributions of $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ are given by Timm [30, p. 121]. Mardia [14,

15] provided percentage points of $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ for $p = 2, 3$, and 4; other values of p or when $n \geq 50$ can be obtained from approximations provided by Rencher [17, p. 113]. A macro provided by SAS (%MULTINORM – see Timm, [30]) can be used to compute these multivariate tests of skewness and kurtosis (see also [31, pp. 32–34]). The macro %MULTINORM can be downloaded from either the Springer-Verlag website – <http://www.springer-ny.com/detail.tpl?isbn=0387953477> or from Timm’s website – <http://www.pitt.edu/~timm>. As Timm indicates, rejection of either test suggests that the data are not multivariate normally distributed; that is, the data contains either multivariate outliers or are not multivariate normally distributed.

Other tests for multivariate normality have been proposed. Looney [13] discusses a number of tests: Royston’s [19] multivariate extension of the Shapiro–Wilk test, Small’s [24] multivariate extensions of the univariate skewness and kurtosis measures, Srivastava’s [26] measures of multivariate skewness and kurtosis, and Srivastava and Hui’s [27] Shapiro–Wilk’s tests. Commenting on these different tests, Looney [13, p. 69] notes that ‘Since no one test for MVN can be expected to be uniformly most powerful against all non-MVN alternatives, we maintain that the best strategy currently available is to perform a battery of tests and then base the assessment of MVN on the aggregate of results.’

If one rejects the multivariate normality hypothesis there are a number of paths that one may follow. First, researchers can search for **multivariate outliers**. However, detecting multivariate outliers is a much more difficult task than identifying univariate outlying values. Graphical procedures might help here (see [22]). The SAS [21] system offers a comprehensive graphing procedure [SAS/INSIGHT – see the discussion and examples in Timm [30, pp. 124–131]. Another approach that researchers frequently employ to detect multivariate outliers is based on Mahalanobis distance statistic (3). That is, one can plot the ordered D_i^2 values against the expected order statistics (q_i) from a chi-squared distribution [see also Timm and Mieczkowski [31, pp. 22–31]. As Timm points out, if the data are multivariate normal, the plotted pairs (D_i^2, q_i) should fall close to the straight line. Timm also notes that there are formal tests for detecting multivariate outliers (see [2]), suggesting however, that these tests have limited value.

Wilcox [32, Chapter 6], on the other hand, presents some newer, more satisfactory, methods for detecting multivariate outliers.

Methods for dealing with multivariate outlying values are also more complex than in the univariate case. As Timm [30] notes, the Box–Cox power transformation has been generalized to the multivariate case by Andrews et al. [1]; however, determining the appropriate transformation is complicated. As an alternative to seeking an appropriate transformation of the data, Timm suggests that researchers can adopt a data reduction transformation, a la principal components analysis, thereby only using a subset (hopefully which are multivariate normally distributed) of the data. Another strategy is to adopt robust estimators instead of the usual least squares estimates; that is, one can utilize trimmed means and a Winsorized variance-covariance matrix. These robust estimators can as well be used to detect outlying values (see [32, 33]) and/or in a multivariate test statistic (see e.g., [12]).

Grouped Data

For many multivariate investigations, data are categorized by group membership (e.g., one-way multivariate analysis of variance). Accordingly, the procedures previously described would be applied to the data within each treatment group (or cell, from a factorial design). Stevens [28] and Tabachnick and Fidell [29] present strategies for applying the previously enumerated procedures for grouped data. As well, they present data sets, indicating statistical software (SAS [20] and SPSS [25]) that can be used to obtain numerical results. It should be noted that when comparing group means it is more important to examine skewness than kurtosis. As DeCarlo [4, p. 297] notes ‘... a general finding for univariate and multivariate data is that tests of means appear to be affected by skew more than kurtosis’.

Numerical Example

To illustrate the use of the methods previously discussed, a hypothetical data set was created by generating data from a lognormal distribution (having skewness = 6.18 and kurtosis = 110.93). In particular, lognormal data were generated for each of two groups ($N_1 = 40$ and $N_2 = 60$), where there were

4 Multivariate Normality Tests

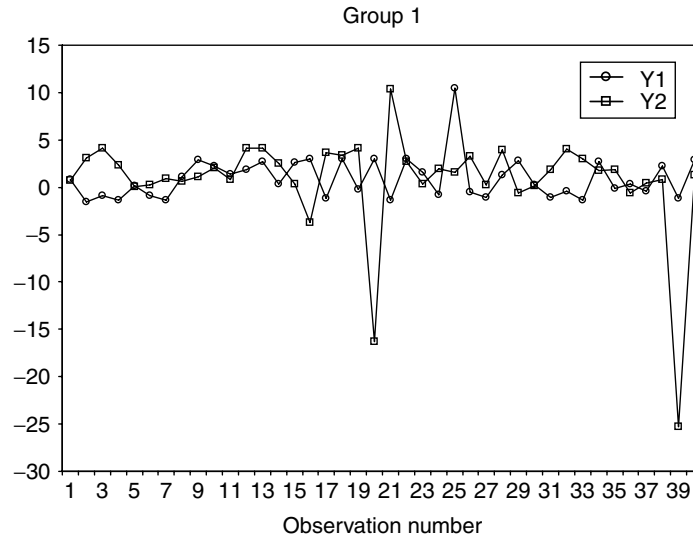


Figure 1 Scores from a lognormal distribution

Table 1 Univariate and multivariate normality tests

Variable	<i>N</i>	Test	Multivariate skewness & kurtosis	Test statistic value	<i>P</i> value
Group 1					
Y_1	40	Shapiro Wilk	.	0.806	.0000
Y_2	40	Shapiro Wilk	.	0.584	.0000
	40	Mardia skewness	15.2437	114.849	.0000
	40	Mardia kurtosis	26.6297	14.728	.0000
Group 2					
Y_1	60	Shapiro Wilk	.	0.699	.0000
Y_2	60	Shapiro Wilk	.	0.485	.0000
	60	Mardia skewness	22.0340	239.200	.0000
	60	Mardia kurtosis	39.4248	30.427	.0000

two dependent measures (Y_1 and Y_2) per subject (see Figures 1 and 2). As indicated, multivariate normality can be assessed by determining whether each of the dependent measures is normally distributed; if not, then multivariate normality is not satisfied. Researchers, however, can choose to bypass these univariate tests (i.e., Shapiro-Wilk) and merely proceed to examine the multivariate measures of skewness and kurtosis with the tests due to Mardia [14]. For completeness, results from both approaches are presented in Table 1. Within each group, both the univariate and multivariate tests for normality are statistically significant.

Having found that the data are not multivariate normally distributed, researchers have to choose some course of action to deal with this problem. As previously indicated, one can search for a transformation to the data in an attempt to achieve multivariate normality, locate outlying values and deal with them, or adopt some robust form of analysis, that is, a procedure (e.g., test statistic) whose accuracy is not affected by nonnormality. Part of the motivation for presenting a two-group problem was to demonstrate a robust solution for circumventing nonnormality. That is, the motivation for the hypothetical data set was that such data could have

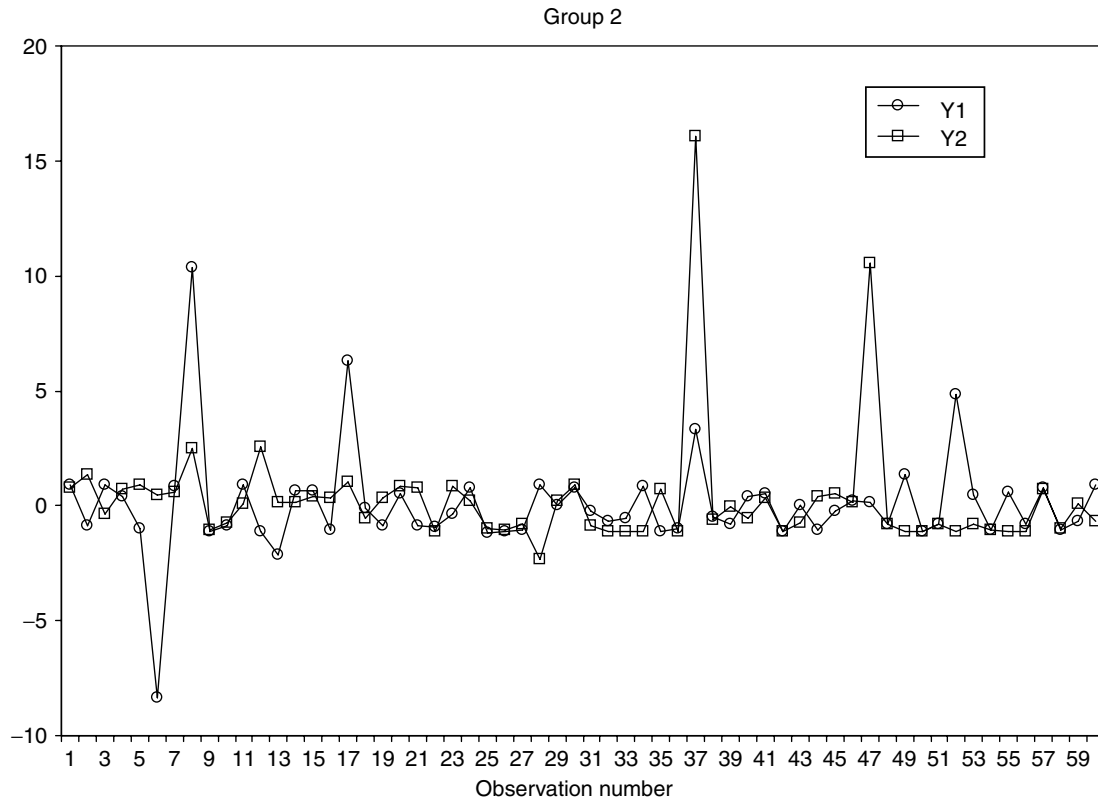


Figure 2 Scores from a lognormal distribution

been gathered in an experiment where a researcher intended to test a hypothesis that the two groups were equivalent on the set of dependent measures. The classical solution to this problem would be to test the hypothesis that the mean performance on each dependent variable is equal across groups, that is, $H_0: \mu_1 = \mu_2$, where $\mu_j = [\mu_{j1} \mu_{j2}]$, and $j = 1, 2$. Hotelling's [7] T^2 is the conventional statistic for testing this multivariate hypothesis; however, it rests on a set of derivational assumptions that are not likely to be satisfied. Specifically, this test assumes that the outcome measurements follow a multivariate normal distribution and exhibit a common covariance structure (covariance homogeneity). One solution for obtaining a robust test of the multivariate null hypothesis is to adopt a heteroscedastic statistic (one that does not assume covariance homogeneity) with robust estimators, estimators intended to be less affected by nonnormality. Trimmed means and Winsorized covariance matrices are robust estimators that

can be used instead of the usual least squares estimators of the group means and covariance matrices (*see Robust Statistics for Multivariate Methods*). Lix and Keselman [12] enumerated a number of robust test statistics for the two-group multivariate problem. For the data presented in Figures 1 and 2, the P value for Hotelling's T^2 statistic is .1454, a nonsignificant result, whereas the P value associated with any one of the robust procedures (e.g., Johanson's [9] test) is .0000, a statistically significant result. Consequently, the biasing effects of nonnormality and/or covariance heterogeneity have been circumvented by adopting a robust procedure, a procedure based on robust estimators and a heteroscedastic test statistic.

Summary

Evaluating multivariate normality is a crucial aspect of multivariate analyses. Though the task is considerably more involved than is the case when there is

but a single measure, researchers would be remiss to ignore the issue. Indeed, most classical multivariate procedures require multivariate normality in order to provide accurate tests of significance and confidence coefficients as well as safeguarding the power to detect multivariate effects.

Researchers should at a minimum assess each of the individual variables for univariate normality. As indicated, data cannot be multivariate normal if the marginal distributions are not univariate normal. Checking for univariate normality is straightforward. Statistical packages (e.g., SAS [20], SPSS [25]) provide measures of skewness and kurtosis as well as formal tests (e.g., Shapiro-Wilk). Perhaps, bivariate normality should be assessed as well. For some authors (e.g., [5]; [10]; [28]) these procedures should suffice to assess the multivariate normality assumption. However, others suggest that researchers go on to formally assessing multivariate normality [13, 17, 30]. (One could bypass tests of univariate normality and proceed directly to an examination of multivariate normality.)

The approach most frequently cited for assessing multivariate normality is the method due to Mardia [14]. Mardia presented measures and tests of multivariate skewness and kurtosis. If the data are found to contain multivariate outlying values and/or are not multivariate normal in form, researchers can adopt a number of ameliorative actions. Data can be transformed (not a straightforward task), outlying values can be detected and removed (see [30, 32, 33]) or researchers can adopt robust estimators with robust test statistics (see [12, 30, 32]). With regard to this last option, some authors would recommend that researchers abandon testing and/or examining the classical parameters (e.g., population least squares means and variance-covariance matrices, classical multivariate measures of effect size and confidence intervals) and examine and/or test robust parameters (e.g., population trimmed means and Winsorized covariance matrices) using robust estimators and thus bypass assessing multivariate normality altogether by exclusively adopting robust procedures (e.g., see [32]).

References

- [1] Andrews, D.F., Gnanadesikan, R. & Warner, J.L. (1971). Transformations of multivariate data, *Biometrics* **27**, 825–840.
- [2] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, Wiley, New York.
- [3] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B* **26**, 211–252.
- [4] DeCarlo, L.T. (1997). On the meaning and use of kurtosis, *Psychological Methods* **2**, 292–307.
- [5] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York.
- [6] Harasym, P.H., Leong, E.J., Lucier, G.E. & Lorscheider, F.L. (1996). Relationship between Myers-Briggs psychological traits and use of course objectives in anatomy and physiology, *Evaluation & The Health Professions* **19**, 243–252.
- [7] Hotelling, H. (1931). The generalization of student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.
- [8] Huberty, C.J. (1994). *Applied Discriminant Analysis*, Wiley, New York.
- [9] Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression, *Biometrika* **67**, 85–92.
- [10] Johnson, R.A. & Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis*, 3rd Edition, Prentice Hall, Englewood Cliffs.
- [11] Knapp, R.G. & Miller, M.C. (1983). Monitoring simultaneously two or more indices of health care, *Evaluation & The Health Professions* **6**, 465–482.
- [12] Lix, L.M. & Keselman, H.J. (2004). Multivariate tests of means in independent groups designs: Effects of covariance heterogeneity and non-normality, *Evaluation & The Health Professions* **27**, 45–69.
- [13] Looney, S.W. (1995). How to use tests for univariate normality to assess multivariate normality, *The American Statistician* **49**, 64–70.
- [14] Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika* **50**, 519–530.
- [15] Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies, *Sankhya, Series B* **36**, 115–128.
- [16] Mardia, K.V. (1980). Tests for univariate and multivariate normality, in *Handbook of Statistics*, Vol. 1, *Analysis of Variance*, P.R. Krishnaiah, ed., North-Holland, New York, 279–320.
- [17] Rencher, A.C. (1995). *Methods of Multivariate Analysis*, Wiley, New York.
- [18] Romeu, J.L. & Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality, *Journal of Multivariate Analysis* **46**, 309–334.
- [19] Royston, J.P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W, *Applied Statistics* **32**, 121–133.
- [20] SAS Institute Inc. (1990). *SAS Procedure Guide*, Version 6, 3rd Edition, SAS Institute, Cary.
- [21] SAS Institute Inc. (1993). *SAS/INSIGHT User's Guide*, Version 6, 2nd Edition, SAS Institute, Cary.
- [22] Serber, G.A.F. (1984). *Multivariate Observations*, Wiley, New York.

-
- [23] Shapiro, S.S. & Wilk, M.B. (1965). An analysis of variance test for normality, *Biometrika* **52**, 591–611.
- [24] Small, N.J.H. (1980). Marginal skewness and kurtosis in testing multivariate normality, *Applied Statistics* **29**, 85–87.
- [25] SPSS Inc. (1994). *SPSS Advanced Statistics 6.1*, Author Chicago.
- [26] Srivastava, M.S. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality, *Statistics and Probability Letters* **2**, 263–267.
- [27] Srivastava, M.S. & Hui, T.K. (1987). On assessing multivariate normality based on the Shapiro-Wilk W statistic, *Statistics and Probability Letters* **5**, 15–18.
- [28] Stevens, J.P. (2002). *Applied Multivariate Statistics for the Social Sciences*, Lawrence Erlbaum, Mahwah.
- [29] Tabachnick, B.G. & Fidell, L.S. (2001). *Using Multivariate Statistics*, 4th Edition, Allyn & Bacon, New York.
- [30] Timm, N.H. (2002). *Applied Multivariate Analysis*, Springer, New York.
- [31] Timm, N.H. & Mieczkowski, T.A. (1997). *Univariate and Multivariate General Linear Models: Theory and Applications Using SAS Software*, SAS Institute, Cary.
- [32] Wilcox, R.R. (in press). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.
- [33] Wilcox, R.R., & Keselman, H.J. (in press). A skipped multivariate measure of location: one- and two-sample hypothesis testing, *Journal of Modern Applied Statistical Methods*.

H.J. KESELMAN

Multivariate Outliers

NICK FIELLER

Volume 3, pp. 1379–1384

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Multivariate Outliers

The presence of outliers in univariate data is readily detected because they must be extreme observations that can easily be identified numerically or graphically. The outliers are amongst the largest or smallest observations and it is not difficult to identify and examine these and perhaps test them for discordancy in relation to the model proposed for the data (*see* **Outliers; Outlier Detection**). By contrast, outliers in multivariate data present special difficulties, primarily because of lack of a clear definition of the ‘extreme observations’ in such a data set. While the intuitive notion that the extreme observations are those that are ‘furthest from the main body of the data’, this does not help locate them since there are many possible ‘directions’ in which they could be separated. As many authors have commented, univariate outliers ‘stick out at one end or the other but multivariate outliers just stick out somewhere’.

Figure 1 shows a plot of two components, B and I, from a set of data containing measurements of nine trace elements (labelled A, B, . . . , I) measured in a collection of 269 archaic clay vessels. There is clearly one outlier ‘sticking out’ in a direction of forty-five degrees from the main body of the data and there are maybe some suspicious observations sticking out in the opposite direction.

Most proposed techniques for pinpointing any outliers that may then be tested formally for discordancy (or, maybe, just examined for deeper understanding of the data) rely on calculations of sample statistics such as the mean and covariance matrix, which

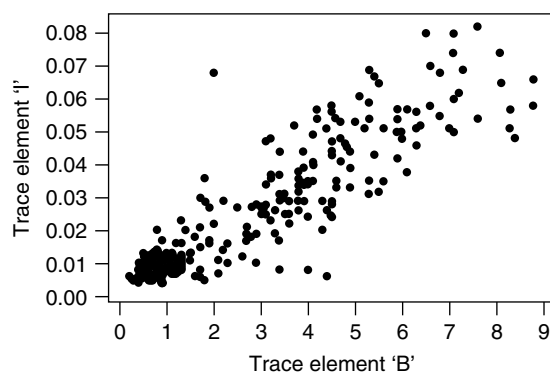


Figure 1 Outlier revealed on bivariate plot

can be seriously affected by the outliers themselves. With univariate data, inflation of sample variance and change in mean does not camouflage their hiding place; by contrast, distortion of the sample covariance can hide multivariate outliers effectively, especially when there are multiple outliers placed in different directions. This feature makes the use of robust estimates of the mean and covariance matrix a possibility to be considered, though this may or may not be effective. This is discussed further below.

As with all outlier problems in the whole range of contexts of statistical analysis, there are three distinct objectives: *identification*, *testing for discordancy* and *accommodation*. Irrespective of whether the first two have been considered, the third of these can be handled by the routine use of robust methods; specifically, use of robust estimates of mean and covariance (*see* **Robust Statistics for Multivariate Methods**). These methods are widely available in many statistical software packages, for example, R, S-Plus, and SAS and are not specifically discussed here (*see* **Software for Statistical Analyses**). Identification may be followed by a formal test of discordancy or it may be that the identification stage has merely revealed simple mistakes in the data recording or, maybe some unexpected feature of interest in its own right. Identification may proceed in step with formal tests in hunting for multiple outliers where *masking* and *swamping* are ever-present dangers.

Identification

Strategies for identification begin with simple graphical methods as an aid to visual inspection. For an $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ representing a random sample of n p -dimensional observations $\mathbf{x}_i$, $i = 1, \dots, n$, univariate plots of original components (or **boxplots** or listing the ordered values numerically) should reveal any simple extreme potential recording errors. Bivariate (or pseudo three-dimensional (*see* **Three Dimensional (3D) Scatterplots**) plots of original components are only worthwhile for low dimensional data, perhaps for $p \leq 10$ or 15, say (an effective upper limit for easy scanning of matrix plots of all pairwise combinations of components). These may not reveal anything other than simple recording errors in a single component which are not extreme enough to have caused concern on first examination of just that component alone. Further, they will not

2 Multivariate Outliers

reveal outliers that are not simple recording errors but are attributable to more interesting unexpected causes. Figure 1 above shows just such an example where the outlier could be a low value of element B or a high value of element I. They are nevertheless worthwhile, especially since such matrix plots are quick and easy to produce (see **Scatterplot Matrices**). For higher dimensional data sets, it is not easy to focus on all of the bivariate scatterplots individually.

A more sophisticated approach is to use an initial dimensionality reduction method as a preliminary step to reduce the combinations of components that need to be examined. The obvious candidate is **principal component analysis**, based on either the covariance or the correlation matrix (or on both in turn) and examining bivariate scatterplots of successive pairs of principal components. These methods have the key advantage in that they can be used even if $n < p$, that is, more dimensions than data points. Here, the rationale is that the presence of outliers will distort the covariance matrix resulting in ‘biasing’ a principal component by ‘pulling it towards’ the outlier, thus allowing it to be revealed as an extreme observation on a scatterplot including that principal component. Clearly, a gross outlier will be revealed on the high order principal components while a minor one will be revealed on those associated with the smaller **eigenvalues**. It is particularly sensible to examine plots of PCs around the ‘cutoff point’ on a scree plot of the eigenvalues since outliers might increase the apparent effective dimensionality of the data set and so be revealed in this ‘additional’ dimension. These methods can be effective for either low numbers of outliers in relation to the sample size or for ‘clusters’ of outliers that arise from a common cause and so ‘stick out’ in the same direction. For large numbers of heterogeneous outliers, then, a robust version of principal component analysis is a possibility, although this loses the rationale that the outliers distort the PCs towards them by distorting the covariance or correlation matrix.

Returning to the measurements of nine trace elements of the clay vessels, Figure 2 shows a scree plot of cumulative variance explained by successive principal components calculated from the correlation matrix of the nine variables on a subset of 53 of the vessels. This suggests that the inherent dimensionality of the data is around five, with the first five principal components containing 93% of the variation and so it could be that outliers would be revealed on plots

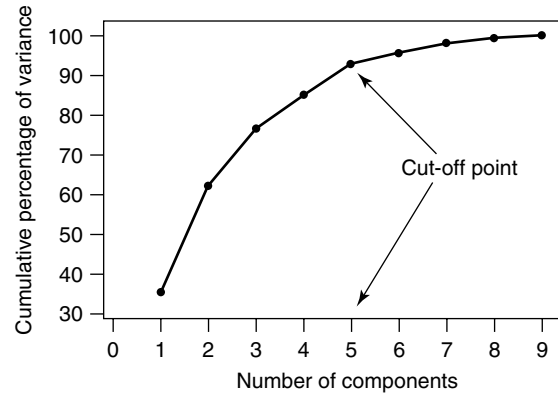


Figure 2 Scree plot of cumulative percentage of variance

of principal components around that value. Figure 3 below shows scatter plots of the data referred to successive pairs of the first five principal components.

Inspection of these suggests that the two observations on the right of the lower right-hand plot (i.e. with high scores on the fifth component) would be worth investigating further, though they are not exceptionally unusual. Closer inspection (with interactive brushing) reveals that one of this pair occurs as the extreme top right observation in the top left plot of PC1 versus PC2 in Figure 3, that is, it has almost the highest scores on both of the first two principal components. Further inspection with brushing reveals that the group of three observations on the top right of PC3 versus PC4 (i.e. those with the three highest scores on the fourth component) have very similar values on all other components and so might be considered a group of outliers from a common source.

A more intensive graphical procedure, which is only viable for reasonably modest data sets (say $n < \sim 50$) with more observations than dimensions ($n > p$) and where only a small number of outliers is envisaged is the use of outlier displaying components (ODCs). These are really just linear discriminant coordinates (sometimes called *canonical variate coordinates*) (see **Discriminant Analysis**). For a single outlier, these ODCs are calculated by dividing the data set into two groups, one consisting of just one observation $\mathbf{x}_j$ and the other of the $(n - 1)$ observations omitting $\mathbf{x}_j$, $j = 1, \dots, n$ (so that there is potentially a separate ODC for each observation). It is easily shown that the univariate likelihood ratio test statistic for discordancy of x_j under a Normal mean slippage model calculated from the values projected

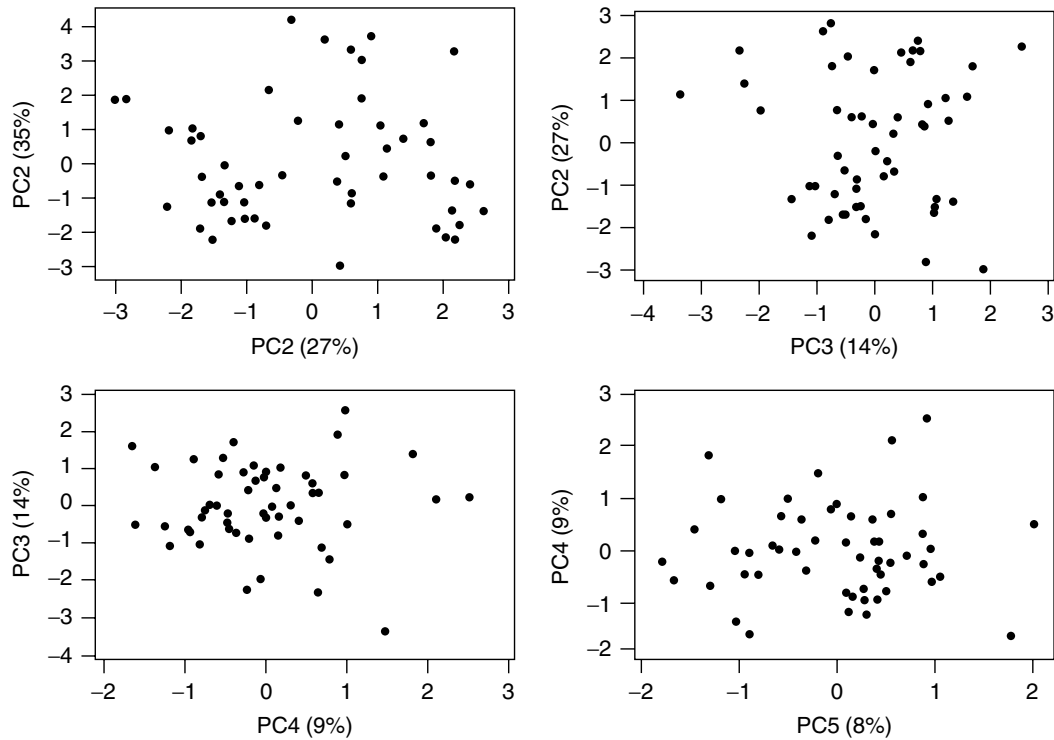


Figure 3 Pairwise plots on first five principal components

onto this ODC is numerically identical to the equivalent statistic (see below), calculated from the original p -dimensional data. Thus, the one-dimensional data projected onto the ODC capture all the numerical information on the discordancy of that observation (though the reference distribution for a formal test of discordancy does depend on the dimensionality p). Picking that with the highest value of the statistic will reveal the most extreme outlier. The generalization to two outliers and beyond depends on whether they arise from a common or two distinct slippages. If the former, then there is just one ODC for each possible pair and, if the latter, then there are two. For three or more, the number of possibilities increases rapidly and so the use is limited as a tool for pure detection of outliers to modest data sets with low contamination rates. However, the procedure can be effective for displaying outliers already identified and for plotting subsidiary data in the same coordinate system. Examination of loadings of original components may give information on the nature of the outliers in the spirit of union-intersection test procedures.

Returning to the data on trace elements in clay-pots, the plot in Figure 4 on the left displays the identified outlier in Figure 1 on the single outlier displaying component (ODC) for that observation as the horizontal axis. This has been calculated using all nine dimensions rather than just two as in Figure 1 and this contains all the information on the deviation of this observation from the rest of the data. The vertical axis has been chosen as the first principal component but other choices might be sensible, for example, a component around the cutoff value of four or five or else on a 'subprincipal component': that vector which maximizes the variance subject to the constraint of being orthogonal with the ODC. Examination of the loadings of the trace elements in the ODC show heavy (negative) weighting on element I with moderate contributions from trace elements F (negative) and E (positive). Trace element B, the horizontal axis in Figure 1, has a very small contribution and, thus, it might be suspected that any recording error is in the measurement of element I. A matrix plot of all components (not shown here) does not

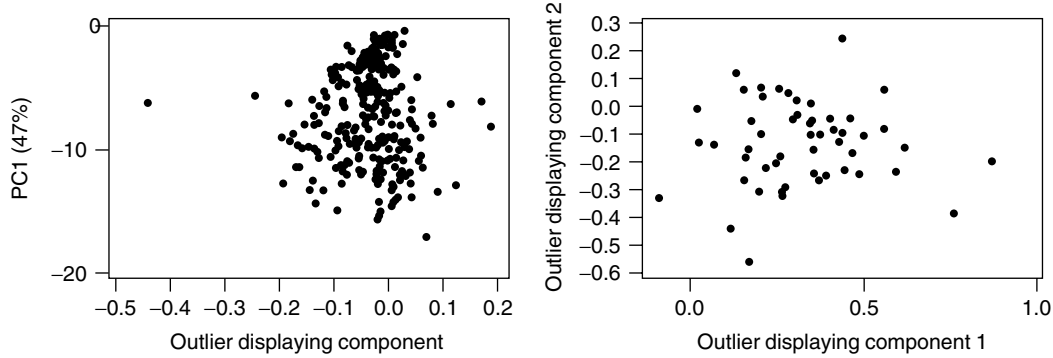


Figure 4 Illustrations of outlier displaying components

reveal this, primarily because this pair of elements is the only one exhibiting any strong correlation. It may be seen that there are several further observations separated from the main body of the data in the direction of the ODC, both with low values (as with the outlier) and with high values. Attention might be directed towards these samples. It should be noted, of course, that the sign of the ODC is arbitrary in the sense that all the coefficients could be multiplied by -1 (thus reflecting the plot about a vertical axis), in the same way that signs of principal components are arbitrary.

The figure on the right of Figure 4 displays the subset of the data discussed in relation to Figure 3. Two groups of outliers have been identified. One is the pair of extreme observation on the fifth principal component and these appear on the right of the plot. The other is the triple of extreme observations on the fourth principal component. These appear in the lower left of the plot. However, two further observations appear separated from the main body in this coordinate system (at the top of the plot) and further examination of these samples might be of interest. Informal examination of the loadings of these two ODCs suggests that the first is dominated by trace elements D and E and the second by E, F and I. Whether this information is of key interest and an aid in further understanding is, of course, a matter for the scientist involved rather than purely a statistical result.

The methods outlined above have the advantage that they display outliers in relation to the original data – they are methods for selecting interesting views of the raw data that highlight the outliers. For data sets with large numbers of observations and

dimensions, this approach may not be viable and various, more systematic approaches have been suggested based on some variant of probability plotting. The starting point is to consider ordered values of a measure of generalized squared distance of each observation from a measure of central location $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma}) = (\mathbf{x}_j - \mathbf{C})' \mathbf{\Gamma}^{-1} (\mathbf{x}_j - \mathbf{C})$, where $\mathbf{C}$ is a measure of location (e.g. the sample mean or a robust estimate of the population mean) and $\mathbf{\Gamma}$ is a measure of scale (e.g., the sample covariance or a robust estimate of the population value). The choice of the sample mean and variance for $\mathbf{C}$ and $\mathbf{\Gamma}$ yields the squared **Mahalanobis distance** of each observation from the mean. It is known that if the $\mathbf{x}_j$ comes from a multivariate Normal distribution (*see Catalogue of Probability Density Functions*), then the distribution of the $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$ is well approximated by a gamma distribution (*see Catalogue of Probability Density Functions*) with, dependent upon $\mathbf{\Gamma}$, some shape parameter that needs to be estimated. A plot of the ordered values against expected order statistics would reveal outliers as deviating from the straight line at the upper end. Barnett and Lewis [1] give further details and references. It should be noted that, typically, these methods are only available when $n > p$; in other cases, a (very robust) dimensionality reduction procedure might be considered first.

Further, these methods are primarily of use when the number of outliers is relatively small. In cases where there is a high multiplicity of outliers, the phenomena of *masking* and *swamping* make them less effective. An alternative approach, typified by Hadi [2, 3], is based upon finding a ‘clean’ interior subset of observations and successively adding more observations to the subset. Hadi recommends starting

with the $p + 1$ observations with smallest values of $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, with very robust choices for $\mathbf{C}$ and $\mathbf{\Gamma}$. An alternative might be to start by *peeling* away successive *convex hulls* to leave a small interior subset. Successive observations are added to the subset by choosing that with the smallest value of $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, calculated just from the clean initial subset (with $\mathbf{C}$ and $\mathbf{\Gamma}$ chosen as the sample mean and covariance) and with the ordering updated at each step. Hadi gives stopping criterion based on simulations from multivariate Normal distributions, the remaining observations not included in the clean subset being declared discordant.

Tests for Discordancy

Formal tests of discordancy rely on distributional assumptions; an outlier can only be tested for discordancy in relation to a formal statistical model. Most available formal tests presume multivariate Normality; exceptions are tests for single outliers in bivariate exponential and bivariate Pareto distributions. Barnett and Lewis [1] provide some details and tables of critical values for these two distributions. For multivariate Normal distributions (see **Catalogue of Probability Density Functions**, the likelihood ratio statistic for a test of discordancy of a single outlier arising from a similar distribution with a change in mean (i.e., a mean slippage model) is easily shown to be equivalent to the Mahalanobis distance from the sample mean (i.e., $R_j(\mathbf{x}_j, \mathbf{C}, \mathbf{\Gamma})$, with $\mathbf{C}$ and $\mathbf{\Gamma}$ taken as the sample mean and covariance). In one dimension, this is equivalent to the studentized distance from the sample mean. Barnett & Lewis [1] give tables of 5% and 1% critical values of this statistic for various values of $n \leq 500$ and $p \leq 5$, together with similar tables for critical values of equivalent statistics when the population covariance and also both population mean and covariance are known, as well as the case where an independent external estimate of the covariance is available. Additionally, they provide similar

tables for the appropriate statistic for a pair of outliers and references to further sources.

Implementation

Although the techniques described here are not specifically available as an ‘outlier detection and testing module’ in any standard package, they can be readily implemented using standard modules provided for principal component analysis, linear discriminant analysis and robust calculation of mean and covariance, provided there is also the capability of direct evaluation of algebraic expressions involving products of matrices and so on. Such facilities are certainly available in R, S-PLUS and SAS. For example, to obtain a single outlier ODC plot, a group indicator needs to be created, which labels the single observation as group 1 and the remaining observations as group 2 and then a standard linear discriminant analysis can be performed. Some packages may fail to perform standard linear discriminant analysis if there is only one observation in the group. However, it is easy to calculate the single outlier ODC for observation $\mathbf{x}_j$ directly as $\mathbf{\Gamma}^{-1}\mathbf{x}_j$, where $\mathbf{\Gamma}$ is the sample covariance matrix or a robust version of it. The matrix calculation facilities would probably be needed for implementation of the robust versions of the techniques.

References

- [1] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, Wiley, New York.
- [2] Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society Series B* **54**, 761–771.
- [3] Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal. Statistical Society Series B* **56**, 393–396.

NICK FIELLER

Neural Networks

FIONN MURTAGH

Volume 3, pp. 1387–1393

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Neural Networks

Introduction

The two types of algorithms to be described can be said to be neuro-mimetic. The human or animal brain is comprised of simple computing engines called neurons, and interconnections called *synapses* which seem to store most of the information in the form of electrical signals. Thus, neural network algorithms usually avail of neurons (also called *units*) that perform relatively simple computing tasks such as summing weighted inputs, and then thresholding the value found, usually in ‘soft’ manner. Soft thresholding is, loosely, approximate thresholding, and is counterposed to hard thresholding. A definition is discussed below. The second point of importance in neural network algorithms is that the interconnections between neurons usually play an important role. Since we are mostly dealing with software implementations of such algorithms, the role of electrical signals is quite simply taken up by a weight – a numeric value – associated with the interconnection.

In regard to the categorization or clustering objectives that these algorithms can cater for, there is an important distinction to be made between, on the one hand, the multilayer perceptron, and on the other hand the Kohonen self-organizing feature map.

The multilayer perceptron (MLP) is an example of a *supervised* method, in that a training set of samples or items of known properties is used. Supervised classification is also known as **discriminant analysis**. The importance of this sort of algorithm is that data from the real world is presented to it (of course, in a very specific form and format), and the algorithm learns from this (i.e., iteratively updates its parameters or essential stored values). So, reminiscent of a human baby learning its first lessons in the real world, a multilayer perceptron can adapt by learning. The industrial and commercial end-objective is to have an adaptable piece of decision-making software on the production line to undertake quality control, or carry out some other decision-based task in some other context.

The Kohonen self-organizing feature map is an example of an *unsupervised* method. **Cluster analysis** methods are also in this family of methods (see **k-means Analysis**). What we have in mind here is that the data, in some measure, sorts itself out or

‘self-organizes’ so that human interpretation becomes easier. The industrial and commercial end-objective here is to have large and potentially vast databases go some way towards facilitating human understanding of their contents.

In dealing with the MLP, the single perceptron is first described – it can be described in simple terms – and subsequently the networking of perceptrons, in a set of interconnected, multiple layers to form the MLP. The self-organizing feature map method is then described. A number of examples are given of both the MLP and Kohonen maps.

A short article on a number of neural network approaches, including those covered here, can be seen in [4]. A more in-depth review, including various themes not covered in this article, can be found in [7]. Taking the regression theme further are the excellent articles by Schumacher et al. [10] and Vach et al. [11].

Multilayer Perceptron

Perceptron Learning

Improvement of classification decision boundaries, or assessment of assignment of new cases, have both been implemented using neural networks since the 1950s. The perceptron algorithm is due to work by Rosenblatt in the late 1950s. The perceptron, a simple computing engine which has been dubbed a ‘linear machine’ for reasons which will become clear below, is best related to supervised classification. The idea of the perceptron has been influential, and generalization in the form of multilayer networks will be looked at later.

Let $\mathbf{x}$ be an input vector of binary values, o an output scalar, and $\mathbf{w}$ a vector of weights (or learning coefficients, initially containing arbitrary values). The perceptron calculates $o = \sum_j w_j x_j$. Let θ be some threshold.

If $o \geq \theta$, when we would have wished $o < \theta$ for the given input, then the given input vector has been incorrectly categorized. We therefore seek to modify the weights and the threshold. Set $\theta \leftarrow \theta + 1$ to make it less likely that wrong categorization will take place again.

If $x_j = 0$, then no change is made to w_j . If $x_j = 1$, then set $w_j \leftarrow w_j - 1$ to lessen the influence of this weight.

2 Neural Networks

If the output was found to be less than the threshold, when it should have been greater for the given input, then the reciprocal updating schedule is implemented.

If a set of weights exist, then the perceptron algorithm will find them. A counter example is the exclusive-or, XOR problem, shown in Table 1.

Here, it can be verified that no way exists to choose values of w_1 and w_2 to allow discrimination between the first and fourth vectors (on the one hand), and the second and third (on the other hand).

Network designs which are more sophisticated than the simple perceptron can be used to solve this problem. The network shown in Figure 1 is a feed-forward three-layer network. This type of network is the most widely used. It has directed links from all nodes at one level to all nodes at the next level. Such a network architecture can be related to a linear algebra formulation, as will be seen below. It has been found to provide a reasonably straightforward architecture

Table 1 The XOR, or exclusive-or problem. Linear separation between the desired output classes, 0 and 1, using just one separating line is impossible

Input vector	Desired output
0, 0	0
0, 1	1
1, 0	1
1, 1	0

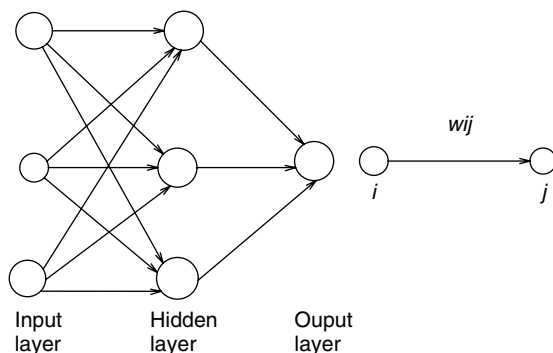


Figure 1 The MLP showing a three-layer architecture, on the left, and one arc on the right. The circles represent neurons (or units)

which can also carry out learning or supervised classification tasks quite well. For further discussion of the perceptron, Gallant [2] can be referred to for various interesting extensions and alternatives. Other architectures are clearly possible: directed links between all pairs of nodes in the network, as in the Hopfield model, directed links from later levels back to earlier levels as in recurrent networks, and so on.

In a feedforward multilayer perceptron, learning, weight estimation (Figure 1, right), takes place through use of gradient descent. This is an iterative sequence of updates to the weights along the directed links of the MLP. Based on the discrepancy between obtained output and desired output, some fraction of the overall discrepancy is propagated backwards, layer by layer, in the MLP. This learning algorithm is also known as backpropagation. Its aim is to bring the weights in line with what is desired, which implies that the discrepancy between obtained and desired output will decrease. Often a large number of feed forward, and backpropagation updates, are needed in order to fix the weights in the MLP. There is no guarantee that the weights will be optimal. At best, they are locally suboptimal in weight space.

It is not difficult to see that we can view what has been described in linear algebra terms. Consider the matrix W_1 as defining the weights between all input layer neurons, i , and all hidden layer neurons, j . Next consider the matrix W_2 as defining the weights between all hidden layer neurons, j , and all output layer neurons, k . For simplicity, we consider the case of the three-layer network but results are straightforwardly applicable to a network with more layers. Given an input vector, $\mathbf{x}$, the values at the hidden layer are given by $\mathbf{x}W_1$. The values at the output layer are then $\mathbf{x}W_1W_2$. Note that we can ‘collapse’ our network to one layer by seeking the weights matrix $W = W_1W_2$. If $\mathbf{y}$ is the target vector, then we are seeking a solution of the equation $\mathbf{x}W = \mathbf{y}$. This is linear regression. The backpropagation algorithm described above would not be interesting for such a classical problem. Backpropagation assumes greater relevance when nonlinear transfer functions are used at neurons.

Generalized Delta Rule for Nonlinear Units

Nonlinear transformations are less tractable mathematically but may offer more sensitive modeling of real data. They provide a more faithful modeling

of electronic gates or biological neurons. Now, we will amend the delta rule introduced in the preceding section to take into consideration the case of a nonlinear transfer function at each neuron.

Consider the accumulation of weighted values at any neuron. This is passed through a differentiable and nondecreasing transformation. In the matrix formalism discussed above, this is just a linear stretch or rescaling function.

For a nonlinear alternative, this transfer function is usually taken as a sigmoidal one. If it were a step function, implying that thresholding is carried out at the neuron, this would violate our requirement for a differentiable function. Taking the derivative of the transfer function, in other words, finding its slope (in a particular direction), is part and parcel of the backpropagation or gradient descent optimization commonly used.

A sigmoidal function is an elongated ‘S’ function, which is not allowed to buckle backwards. One possibility is the function $y = (1 + e^{-x})^{-1}$. Another choice is the hyperbolic tangent or tanh function: for $x > 20.0$, $y = +1.0$; for $y < -20.0$, $y = -1.0$; otherwise $y = (e^x - e^{-x}) / (e^x + e^{-x})$. Both of these functions are invertible and continuously differentiable. Both have semilinear zones which allow good (linear) fidelity to input data. Both can make ‘soft’ or fuzzy decisions. Finally, they are similar to the response curve of a biological neuron.

A ‘pattern’ is an input observation vector, for instance, a row of an input data table. The overall training algorithm is as follows: present pattern; feed forward through successive layers; backpropagate, that is, update weights; repeat.

An alternative is to determine the changes in weights as a result of presenting all patterns. This so-called ‘off-line’ updating of weights is computationally more efficient but loses out on the adaptivity of the overall approach.

We have discussed how discrepancy between obtained and desired output is used to drive or direct the learning. The discrepancy is called the *error*, and is usually a squared Euclidean distance, or a mean square error (i.e., mean squared Euclidean distance). Say that we are using binary data only. Such data could correspond to some qualitative or categorical data coding. In regard to outputs obtained, in practice, one must use approximations to hard-limiting values of 0, 1; for example, 0.2, 0.8 can be used. Any output value less than or equal to 0.2 is taken as tantamount

to 0, and any output value greater than or equal to 0.8 is taken as tantamount to 1.

If we are dealing instead with quantitative, real-valued data, then we clearly have a small issue to address: a sigmoidal ‘squashing’ transfer function at the output nodes means that output values will be between 0 and 1. To get around this limitation, we can use sigmoidal transfer functions at all neurons, save the output layer. At the output layer we use a linear transfer function at each neuron.

The MLP architecture using the generalized delta rule can be very slow for the following reasons. The compounding effect of sigmoids at successive levels can lead to a very flat energy surface. Hence movement towards fixed points is not easy. A second reason for slowness when using the generalized delta rule is that weights tend to try together to attain the same value. This is unhelpful – it would be better for weights to prioritize themselves when changing value. An approach known as cascade correlation, due to Scott Fahlman, addresses this by introducing weights one at a time.

The number of hidden layers in an MLP and the number of nodes in each layer can vary for a given problem. In general, more nodes offer greater sensitivity to the problem being solved, but also the risk of overfitting. The network architecture which should be used in any particular problem cannot be specified in advance. A three-layer network with the hidden layer containing, say, less than $2m$ neurons, where m is the number of values in the input pattern, can be suggested.

Other optimization approaches and a range of examples are considered in [4, 7]. Among books, we recommend the following: [1], [3], and [9].

Example: Inferring Price from Other Car Attributes

The source of the data used was the large machine learning datasets archive at the University of California Irvine. The origin of the data was import yearbook and insurance reports. There were 205 instances or records. Each instance had 26 attributes, of which 15 were continuous, 1 integer, and 10 nominal. Missing attribute values were denoted by ‘?’. Table 2 describes the attributes used.

Some explanations on Table 2 follow. ‘Symboling’ refers to the actuarial risk factor symbol associated with the price (+3: risky; –3: safe). Attribute

4 Neural Networks

Table 2 Description of variables in car data set

Attribute	Attribute Range
1. symboling	−3, −2, −1, 0, 1, 2, 3
2. normalized-losses	continuous from 65 to 256
3. make	alfa-romeo, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugeot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type	diesel, gas
5. aspiration	standard, turbo
6. num-of-doors	four, two
7. body-style	hardtop, wagon, sedan, hatchback, convertible
8. drive-wheels	4wd, fwd, rwd
9. engine-location	front, rear
10. wheel-base	continuous from 86.6 to 120.9
11. length	continuous from 141.1 to 208.1
12. width	continuous from 60.3 to 72.3
13. height	continuous from 47.8 to 59.8
14. curb-weight	continuous from 1488 to 4066
15. engine-type	dohc, dohc, l, ohc, ohc, ohcv, rotor
16. num-of-cylinders	8, 5, 4, 6, 3, 12, 2
17. engine-size	continuous from 61 to 326
18. fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
19. bore	continuous from 2.54 to 3.94
20. stroke	continuous from 2.07 to 4.17
21. compression-ratio	continuous from 7 to 23
22. horsepower	continuous from 48 to 288
23. peak-rpm	continuous from 4150 to 6600
24. city-mpg	continuous from 13 to 49
25. highway-mpg	continuous from 16 to 54
26. price	continuous from 5118 to 45 400

number 2: relative average loss payment per insured vehicle year, normalized for all cars within a particular size classification.

As a sample of the data, the first three records are shown in Table 3. The objective was to predict the

price of a car using all other attributes. The coding and processing carried out by us is summarized as follows.

Records with missing prices were ignored, leaving 201 records. All categorical attribute values were mapped onto 1, 2, Missing attribute values were set to zero. Price attribute (to be predicted) was discretized into 10 classes. All nonmissing attributes were normalized to lie in the interval [0.1, 1.0]. A test set of 28 uniformly sequenced cases was selected; a training set was given by the remaining 173 cases.

A three-layer MLP was used, with 25 inputs and 10 outputs. The number of units (neurons) in the hidden layer was assessed on the training set, and covered 4, 8, 12, . . . , 36, 40 neurons. For comparison, a k -Nearest Neighbors discriminant analysis was also carried out with values of $k = 1, 2, 4$. (K -Nearest Neighbors involves taking a majority decision for assignment, among the class assignments found for the k Nearest neighbors. Like MLP, it is a nonlinear mapping method.)

Best performance was found on the training set using 28 hidden layer neurons.

Results obtained were as follows, – learning on the training set and generalization on the test set.

MLP 25–28–10 network:

Learning: 96.53%

Generalization: 71.43%

k -Nearest Neighbors

$k = 1$: Learning: 97.64%, Generalization: 67.86%

$k = 2$: Learning: 85.54%, Generalization: 71.43%

$k = 4$: Learning: 73.99%, Generalization: 71.43%

Hence, the MLP's generalization ability was better than that of 1-Nearest Neighbor's, but equal to that of 2- and 3-Nearest Neighbor's. Clearly, further analysis (feature selection, data coding) could be useful. An advantage of the MLP is its undemanding requirements for coding the input data.

Table 3 Sample of three records from car data set

3, ?, alfa-romeo, gas, std, two, convertible, rwd, front, 88.60, 168.90, 64.10, 48.80, 2548, dohc, four, 130, mpfi, 3.47, 2.68, 9.00, 111, 5000, 21, 27, 13495
3, ?, alfa-romeo, gas, std, two, convertible, rwd, front, 88.60, 168.90, 64.10, 48.80, 2548, dohc, four, 130, mpfi, 3.47, 2.68, 9.00, 111, 5000, 21, 27, 16500
1, ?, alfa-romeo, gas, std, two, hatchback, rwd, front, 94.50, 171.20, 65.50, 52.40, 2823, ohcv, six, 152, mpfi, 2.68, 3.47, 9.00, 154, 5000, 19, 26, 16500

In the example just considered, we had 201 cars, with measures on 25 attributes. Let us denote this data as matrix X , of dimensions 201×25 . The weights between input layer and hidden layer can be represented by matrix U , of dimensions 25×28 . The weights between hidden layer and output layer can be represented by matrix W , of dimensions 28×10 . Let us take the sigmoid function used as $f(\mathbf{z}) = 1/(1 + e^{-\mathbf{z}})$, acting elementwise on its argument, $\mathbf{z}$. The sigmoid function can be straightforwardly extended for use on a matrix (i.e., set of vectors, $\mathbf{z}$). We will denote the outputs as Y , a 201×10 matrix, giving assignments – close to 0 or close to 1 – for all 201 cars, with respect to 10 price categories. Then, finally, our trained MLP gives us the following nonlinear relationship:

$$Y = f((f(XU))W) \quad (1)$$

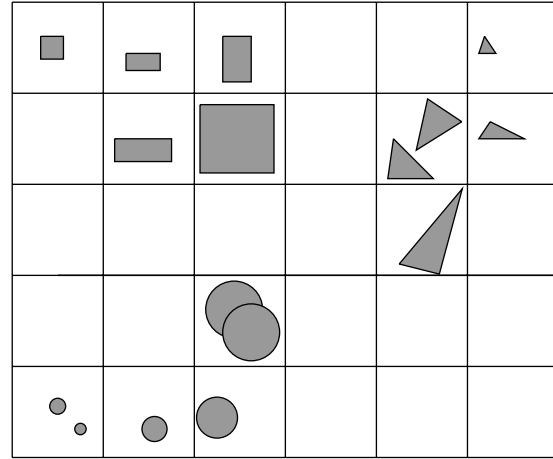
The Kohonen Self-organizing Feature Map

A self-organizing map can be considered as a display grid where objects are classified (Figure 2). In such a grid, similar objects are located in the same area. In the example (Figure 2, right), a global classification is shown: the three different shapes are located in three different clusters. Furthermore, the largest objects are located towards the center of the grid, and each cluster is ordered: the largest objects are at one side of a cluster, smaller shapes are at the other side.

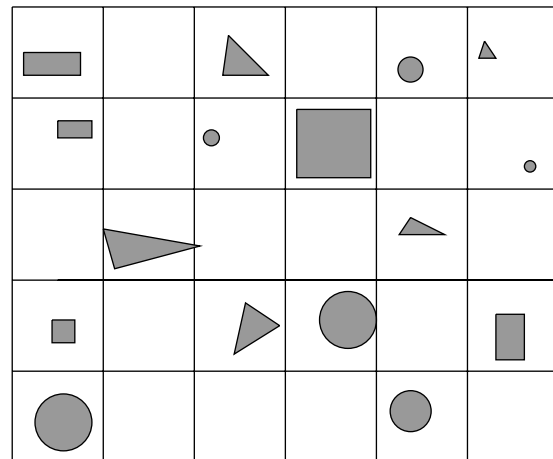
This leads to an interesting and effective use of the Kohonen self-organizing feature map for both the presentation of information, and as an interactive user interface. As such, it constitutes an active and responsive output from the Kohonen map algorithm.

A self-organizing map is constructed as follows.

- Each object is described by a vector. In Figure 2, the vector has two components: the first one corresponds to the number of angles, and the other one to the width of the area.
- Initially, a vector is randomly associated with each box (or node) of the grid.
- Each document is located in a box whose descriptive vector is the most similar to the object's vector.
- During an iterative learning process, the components of the nodes describing vectors are



(a)



(b)

Figure 2 Notional view of Kohonen self-organizing feature map. Object locations. (a) before learning. (b) after learning

modified. The learning process produces the classification.

The Kohonen self-organizing feature map (SOFM) with a regular grid output representational or display space, therefore involves determining cluster representative vectors, such that observation vector inputs are parsimoniously summarized (clustering objective); and in addition, the cluster representative vectors are positioned in representational space so that similar vectors are close (low-dimensional projection objective) in the representation space.

The output representation grid is usually a square, regular grid. ‘Closeness’ is usually based on the Euclidean distance. The SOFM algorithm is an iterative update one, not similar to other widely used clustering algorithms (e.g., **k-means clustering**). In representation space, a neighborhood of grid nodes is set at the start. Updates of all cluster centers in a neighborhood are updated at each update iteration. As the algorithm proceeds, the neighborhood is allowed to shrink. The result obtained by the SOFM algorithm is suboptimal, as also is the case usually for clustering algorithms of this sort (again, for example, k-means). A comparative study of SOFM performance can be found in [5]. An evaluation of its use as a visual user interface can be found in [6, 8].

References

- [1] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*, Oxford University Press.
- [2] Gallant, S.I. (1990). Perceptron-based learning algorithms, *IEEE Transactions on Neural Networks* **1**, 179–191.
- [3] MacKay, D.J.C. (2003). *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.
- [4] Murtagh, F. (1990). Multilayer perceptrons for classification and regression, *Neurocomputing* **2**, 183–197.
- [5] Murtagh, F. (1994). Neural network and related massively parallel methods for statistics: a short overview, *International Statistical Review* **62**, 275–288.
- [6] Murtagh, F. (1996). Neural networks for clustering, in *Clustering and Classification*, P., Arabie, L.J., Hubert & G., De Soete, eds, World Scientific.
- [7] Murtagh, F. & Hernández-Pajares, M. (1995). The Kohonen self-organizing feature map method: an assessment, *Journal of Classification* **12**, 165–190.
- [8] Poinçot, P., Lesteven, S. & Murtagh, F. (2000). Maps of information spaces: assessments from astronomy, *Journal of the American Society for Information Science* **51**, 1081–1089.
- [9] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.
- [10] Schumacher, M., Roßner, R. & Vach, W. (1996). Neural networks and logistic regression: Part I, *Computational Statistics and Data Analysis* **21**, 661–682.
- [11] Vach, W., Roßner, R. & Schumacher, M. (1996). Neural networks and logistic regression: Part II, *Computational Statistics and Data Analysis* **21**, 683–701.

Further Reading

- Murtagh, F., Taskaya, T., Contreras, P., Mothe, J. & Englemeier, K. (2003). Interactive visual user interfaces: a survey, *Artificial Intelligence Review* **19**, 263–283.

FIONN MURTAGH

Neuropsychology

YANA SUCHY

Volume 3, pp. 1393–1398

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Neuropsychology

Neuropsychology is a scientific and clinical discipline that is concerned with two related issues that are, at least theoretically, methodologically, and statistically, distinct. These are: (a) Understanding brain-behavior relationship, and (b) the development and application of methods (including statistical methods) that allow us to assess behavioral manifestations of various types of brain dysfunction.

To illustrate how these clinical and scientific interests come together practically, let us follow a hypothetical case scenario. Imagine a man, let us call him Mr. Fixit, who, while repairing a loose gutter 30 feet above the ground, falls from his ladder. Let us further imagine that this man is fortunate enough to fall into a pile of dry leaves. The man escapes this accident without fractures or internal injuries, although he does fall unconscious briefly. He regains consciousness within a few minutes, feeling woozy, shaky, and sick in his stomach. His wife takes him to the hospital, where he is diagnosed with a concussion and released, with the instruction to take it easy for a few days until the wooziness and nausea go away. Let us further imagine that this man is an owner of a small business, and that his daily activities require him to organize and keep in mind many details as he receives inquiries from potential customers, questions from his five employees, and quotes from suppliers.

The man returns to work after a few days of rest, looking forward to catching up on paperwork, phone calls, purchases, and business meetings. By the end of the first week, however, he feels frustrated, irritated, exhausted, and no closer to catching up than he was when he first returned to work. By the end of the second week, the situation is no better: In fact, not only is he nowhere close to catching up but he is also falling further behind. After a couple of more weeks, the man's wife convinces him that these difficulties seem to be linked to his fall, and that he should see his doctor.

Upon his wife's urging, Mr. Fixit goes back to the neurologist who first examined him, but his visit is not very satisfying: He is told to give it time, that his 'symptoms' are normal, and that he should not worry about it. To complicate matters further, Mr. Fixit has noticed that his ladder has a manufacturing defect that may have caused his fall, and is now considering legal action. And, so, to bring this story

to the topic of this article, Mr. Fixit ends up in a neuropsychologist's office, asking whether his fall might have permanently damaged his brain in such a way as to cause him the difficulties he has been experiencing.

How would a neuropsychologist go about answering such a question? To avoid an unduly complicated discussion, suffice it to say that the neuropsychologist first needs to have a working knowledge of the topics that are subsumed under the heading of 'brain-behavior relationship'. For example, the neuropsychologist must understand what type of damage *typically* occurs in a person's brain as a result of a fall such as the one suffered by Mr. Fixit, and what behaviors are *typically* affected by such damage. Second, the neuropsychologist must possess 'assessment instruments' that allow him to examine Mr. Fixit and interpret the results. The remainder of this chapter focuses on the statistical methods that allow neuropsychologists to both gather knowledge about brain-behavior relationship and to develop appropriate assessment instruments. We will return to the specifics of Mr. Fixit's case to illustrate methodological and statistical issues pertinent to both experimental and clinical neuropsychology.

Statistical Methods Used for Examination of Brain-behavior Relationship

The first question that a neuropsychologist examining Mr. Fixit needs to answer is whether it is realistic that Mr. Fixit's fall resulted in the symptoms that Mr. Fixit describes. To answer questions of this type, neuropsychologists conduct studies in which they compare test performances of individuals who sustained brain injury to healthy controls. Such studies typically use a simple independent *t* Test, or, if more than two groups are compared, a one-way **Analysis of Variance (ANOVA)**. Past studies utilizing such comparisons have demonstrated that patients who have sustained a mild traumatic brain injury (similar to the one sustained by Mr. Fixit) may exhibit difficulties with memory and attention, or difficulties in organization and planning.

Although simple in their design, comparisons of this kind continue to yield invaluable information about the specific nature of patients' deficits. For example, Keith Cicerone [4] recently examined different types of attention in a group of patients similar

to Mr. Fixit. Independent t Tests showed that, from among seven different measures of attention, only those that required vigilant processing of rapidly presented information were performed more poorly among brain-injured, as compared to healthy, participants. In contrast, brief focusing of attention was no different for the two groups.

However, because different measures have different discriminability, a failure to find differences between groups with respect to a particular variable (such as the ability to briefly focus attention) can be difficult to interpret. In particular, it is not clear whether such a failure truly suggests that patients are unimpaired, or whether it suggests inadequate measurement or inadequate statistical power. One way to address this question is to examine **interaction effects** using factorial ANOVA (*see Factorial Designs*). As an example, Schmitter-Edgecombe et al. [9] compared brain-injured and non-brain-injured participants on several indices of memory. They conducted a mixed factor ANOVA with group (injured vs. noninjured) as a between-subjects factor and different components of memory performance as within-subjects factors. Their results yielded a main effect of group, showing that brain-injured patients were generally worse on memory overall. However, the presence of an interaction between group and memory component also allowed the researchers to state with a reasonable degree of certainty that certain facets of memory processing were *not* impaired among brain-injured patients.

In addition to understanding what kinds of deficits are typical among patients who have sustained a brain injury, it is equally important to understand the time course of symptom development and symptom recovery? In other words, is it realistic that Mr. Fixit's symptoms would not have improved or disappeared by the time he saw the neuropsychologist? And, relatedly, should he be concerned that some symptoms will never disappear? Studies that address these questions generally rely on **Repeated Measures Analysis of Variance**, again using group membership as a between-subjects factor and time since injury (e.g., one week, one month, six months) as a within-subjects factor. This method tells us whether either, or both, groups improved in subsequent testing sessions (a within-subjects effect), as well as whether the two groups differ from one another at all, or on only some testing sessions (a between-subjects effect).

As an example, Bruce and Echemendia [3], in their study of sports-related brain injuries, applied this technique using group membership (brain-injured vs. non-brain-injured athletes) as the between-subjects factor and time since injury (2 hours, 48 hours, 1 week, and 1 month) as the within-subjects factor. They found that, whereas healthy participants' performance improved on second testing (due to practice effects), no such improvement was present for the brain-injured group. Essentially, this means that memory performance declined initially in the brain-injured group during the period of 2 to 48 hours, likely due to acute effects of injury such as brain swelling. The results further showed that brain-injured patients' memory abilities recovered significantly by the time of the one-month follow-up assessment. Thus, Mr. Fixit's physician was correct in encouraging Mr. Fixit to 'give it more time'.

Addressing Confounds. As is the case with much of clinical research, research in neuropsychology typically relies on **quasi-experimental designs**. It is well recognized that such designs are subject to a variety of confounds (*see Confounding Variable*). Because much research in clinical neuropsychology is based on datasets that come from actual patients who were examined initially for clinical purposes rather than research purposes, confounds in neuropsychology are perhaps even more troublesome than in some other clinical areas. Thus, for example, although the studies presented up to this point demonstrate that patients like Mr. Fixit exhibit poorer performances than healthy controls with respect to attention and memory, we cannot be absolutely certain that all of the observed group differences are due to the effects of brain injury. Instead, we might argue that people who fall from ladders or are otherwise 'accident prone' tend to suffer from attentional and memory problems even prior to their injuries. In fact, perhaps these accident-prone individuals have always been somewhat forgetful and inattentive, and perhaps these were the very characteristics that caused them to suffer their injuries in the first place.

Bruce and Echemendia [3], in their study of sports-related brain injury, found just such a scenario. Specifically, the athletes who sustained head injuries had lower premorbid (i.e., preinjury) SAT scores than those in a control group. This scenario poses the following questions: What is responsible for the poor

memory and attentional performances of the brain-injured group? Is it the lower premorbid functioning (reflected in their lower SAT scores), or is it the brain injury?

Addressing this question statistically is not as simple as one might think. On the one hand, some investigators attempt to address this problem by using **the Analysis of Covariance (ANCOVA)**, as did Bruce and Echemendia [3] in their study when they used the SAT scores as a **covariate**. On the other hand, Miller and Chapman [6] argue that the *only* appropriate use of ANCOVA is to account for *random* variation on the covariate, not to ‘control’ for differences between groups. However, while it is true that the results of an ANCOVA regarding premorbid differences are not easy to interpret, some investigators maintain that conducting an ANCOVA in this situation is appropriate, as it might at least afford ruling out the effect of the covariate. In particular, if group differences continue to be present even after a covariate is added to the analysis, it is clear that the dependent variable, not the covariate, accounted for such differences. If, on the other hand, adding a covariate abolishes a given effect, the results cannot be readily interpreted.

To avoid this controversy, some researchers match groups on a variety of potential confound variables, most commonly age, education, and gender. This approach, however, is not controversy-free either, as it may lead to samples that are unrealistically constrained or are not representative of the population of interest. For example, it is well understood that mild traumatic brain injury patients who are involved in litigation (as Mr. Fixit might become, due to his defective ladder) perform more poorly on neuropsychological tests than those who are not involved in litigation [8]. It is also well understood that a substantial subgroup of litigating patients consciously or unconsciously either feign or exaggerate their cognitive difficulties [1]. Thus, researchers who are interested in understanding the nature of cognitive deficits resulting from a traumatic brain injury often remove the litigating patients from their sample. However, it is possible that these patients litigate *because of* continued cognitive difficulties. For example, would Mr. Fixit even think of litigation (regardless of whether he noticed the defect on his ladder) if he were not experiencing difficulties at work? Thus, by excluding these patients from our samples, we may be underestimating the extent of

deleterious effects of traumatic brain injury on cognition. Because of these statistical and methodological controversies, accounting for confounds is a constant struggle both in neuropsychology research and in neuropsychology practice. In fact, these issues will figure largely in the mind of Mr. Fixit’s neuropsychologist who attempts to determine the true nature and severity of Mr. Fixit’s difficulties.

Neuropsychological Assessment Instruments

Understanding the realistic nature and course of Mr. Fixit’s difficulties allows the clinician to select a proper battery of tests for Mr. Fixit’s assessment. However, there are hundreds of neuropsychological instruments available on the market. How does a neuropsychologist choose the most appropriate ones for a given case? Research and statistical methods that address different aspects of this question are described below.

Validity

The first step in selecting assessment instruments is to select instruments that measure the constructs of interest (*see* **Validity Theory and Applications**). In this case, the various functional domains that might have been affected by Mr. Fixit’s fall, such as attention and memory, represent our constructs. While some constructs and the related assessment instruments are relatively straightforward and require relatively little validation past that conducted as part of the initial test development, others are complex or even controversial, requiring continual reexamination. An example of such a complex construct is a functional domain called *executive abilities*. Executive abilities allow people to make intelligent choices, avoid impulsive and disorganized actions, plan ahead, and follow through with plans. Research shows that executive abilities are sometimes compromised following a traumatic brain injury. Given Mr. Fixit’s difficulties managing his small business, it is quite likely that impaired executive functions are to blame. Assessment of executive abilities is arguably among the greatest challenges in clinical neuropsychology, with much research devoted to the validation of measures that assess this domain.

A common statistical method for addressing the question of a test’s construct validity is a factor

analysis (*see* **Factor Analysis: Exploratory**). Factor analysis can demonstrate that the items on a test follow a certain theoretically cogent structure. Typically, a test's factor structure is examined during the initial test development process. However, it is sometimes the case that factor analysis research may offer new insights even after a test has been in clinical use for some time. For example, based on their day-in and day-out experiences with patients, clinicians may sometimes become aware of certain qualitative aspects of test performance that may be related to specific disease processes. For these clinical impressions to be confirmed, a formal validation is required. An example of such a validation study was that of Osmon and Suchy [7] who examined several qualitative aspects of performance on the Wisconsin Card Sorting Test, a popular test of executive functioning. Factor analyses yielded three factors, demonstrating that different types of errors represent separate components of executive processing. Thus, the results support the notion that the Wisconsin Card Sorting Test assesses more than one executive ability, which is in-line with the conceptualization of executive functions as a nonunitary construct.

In addition to examining the factor structure of a test, validation can also be accomplished by examining whether a measure correlates with other instruments that are known to measure the same construct. A study conducted by Bogod and colleagues [2] is an example of this approach. Specifically, these researchers were interested in determining the validity of a measure of self-awareness (called SADI). Self-awareness, or awareness of deficits, is sometimes impaired in patients who have suffered a traumatic brain injury. Bogod and colleagues used such patients to examine whether SADI scores correlated with another index of self-awareness. These analyses yielded the first tentative support for the measure's construct validity. In particular, the analyses showed that, although the SADI did not correlate with the deficits reported by the patient or the patients' relatives, it did correlate with the degree of discrepancy between the patients' and the relatives' reports. In other words, the measure was unrelated to the severity of one's disability, but rather to one's inability to accurately appraise one's disability. Additionally, because deficits in self-awareness are, at least theoretically, related to deficits in executive functioning, Bogod and colleagues examined the relationship

between the SADI and measures of executive functions. Medium size correlations provided additional support for the measure, as well as for the theoretical conceptualization of the construct of self-awareness and its relationship to other cognitive abilities.

However, demonstrating that a particular test measures a particular cognitive construct does not address the question of whether a deficit on a test relates to real-life abilities. In other words, does a test predict how a person will function in everyday life? This is often referred to as the test's 'ecological validity'. As an example, Suchy and colleagues [10] determined whether an executive deficit identified during the course of hospitalization predicted patients' daily functioning following discharge. These researchers used a series of Stepwise Multiple Regressions (*see* **Multiple Linear Regression**) in which different behavioral outcomes (e.g., whether patients prepare their own meals, take care of their hygiene needs, or manage their own finances) served as criterion variables and three factors of the Behavioral Dyscontrol Scale (BDS), a measure of executive functioning, served as predictors. Because Stepwise Multiple Regressions can be used to generate parsimonious models of prediction, the authors showed not only that the BDS was a significant predictor of actual daily functioning, but also which components of the measure were the best predictors. A caution about Stepwise Multiple Regressions is that it can unduly capitalize on the idiosyncrasies of the sample. As a result, only a small difference in sample characteristics can lead to a variable's exclusion or inclusion in the final model. For that reason, it is a good idea to test the stability of generated models by repeating the analyses with random subsamples drawn from the original sample. Discrepancies in results must be resolved both statistically and theoretically prior to drawing final conclusions.

Finally, because patients and settings vary widely across different neuropsychological studies, it is common for a single measure to yield a variety of contradictory findings. For example, the sensitivity and specificity of the Wisconsin Card Sorting Test has been an ongoing source of controversy in neuropsychology research and practice. Some studies have found the measure to have good sensitivity and specificity for frontal lobe integrity (the presumed substrate of executive abilities), whereas others have shown no such relationship. Recently, Demakis [5] addressed the ongoing controversy about the test's

utility by conducting a **meta-analysis**. The composite finding of 24 studies yielded small but statistically significant effect sizes, providing validation for both sides of the argument: Yes, the measure is related to frontal lobe functioning, and, yes, the relationship is weak.

Clinical Utility

Research and statistical methods described up to this point allow us to determine that, on average, patients who have suffered a traumatic brain injury perform more poorly than healthy controls on validated measures of executive functioning, attention, and memory. This means that Mr. Fixit's difficulties might very well be real. However, we don't know which measures are the best for making decisions about Mr. Fixit's specific situation. To answer that question, researchers must examine how successful different measures are at patient classification.

Traditionally, **Discriminant Function Analysis (DFA)** has been the statistic of choice for examining a measure's classification accuracy. Although this procedure provides information about the overall hit rate and its statistical significance, these results do not always generalize to clinical settings. First, this method produces cutting scores that will maximize the overall hit rate, regardless of what the clinical needs are. In other words, in some clinical situations, it may be more useful to have a test with a high specificity, whereas, in other situations, a test of high sensitivity may be needed. Second, classifications are maximized by using discriminant function coefficients that require a conversion of the raw data. Such conversions are not always practical, and, as a result, are generally of little clinical use.

In recent years, the **Receiver Operating Characteristic (ROC) Curve** has gained popularity over DFA in neuropsychology research. This procedure allows determination of exact sensitivity and specificity associated with every possible score on a given test. Based on ROC data, one can choose a cutting score with a high sensitivity to 'rule in' patients who are at risk for a disorder, high specificity to 'rule out' patients who are at low risk for a disorder, or an intermediate sensitivity and specificity to maximize the overall hit rate. Additionally, the ROC analysis allows determination of whether the classification rates afforded by the test are statistically significant.

To conduct an ROC analysis, one needs a sample of subjects for whom an external validation of 'impaired' is available. Depending on the setting in which a test is employed, a variety of criteria can be used. Some examples of external criteria of impairment include inability to succeed at work (as was the case with Mr. Fixit), inability to live independently (a common issue with the elderly), or an external diagnosis of a condition based on neuroradiology or other medical data. By conducting an ROC analysis with such a sample, empirically based cutting scores can be applied in clinical settings to make decisions about patients for whom external criteria of impairment are not known.

Conclusions

A variety of statistical techniques are needed to conduct research that allows a clinician to address even one clinical scenario. In our hypothetical scenario, *t* Tests, ANOVAs, and ANCOVAs helped our clinician to determine whether Mr. Fixit's complaints were realistic and what cognitive difficulties might account for his complaints. Factor analyses, correlational analyses, meta-analyses, DFA, and ROC curve analyses helped determine what measures are most appropriate for assessing the exact nature of Mr. Fixit's difficulties. Multiple regressions allowed predictions about the impact of his deficits on his everyday functioning.

References

- [1] Binder, L. (1993). Assessment of malingering after mild head trauma with the Portland digit recognition Test, *Journal of Clinical and Experimental Neuropsychology* **15**, 170–182.
- [2] Bogod, N.M., Mateer, C.A. & MacDonald, S.W.S. (2003). Self-awareness after traumatic brain injury: a comparison of measures and their relationship to executive functions, *Journal of International Neuropsychological Society* **9**, 450–458.
- [3] Bruce, J.M. & Echemendia, R.J. (2003). Delayed-onset deficits in verbal encoding strategies among patients with mild traumatic brain injury, *Neuropsychology* **17**, 622–629.
- [4] Cicerone, K.D. (1997). Clinical sensitivity of four measures of attention to mild traumatic brain injury, *The Clinical Neuropsychologist* **11**, 266–272.
- [5] Demakis, G. (2003). A meta-analytic review of the sensitivity of the Wisconsin card sorting Test to frontal

- and lateralized frontal brain damage, *Neuropsychology* **17**, 255–264.
- [6] Miller, G.M. & Chapman, J.P. (2001). Misunderstanding analysis of covariance, *Journal of Abnormal Psychology*, **110**(1), 40–48.
- [7] Osmon, D.C. & Suchy, Y. (1996). Fractionating frontal lobe functions: factors of the Milwaukee card sorting Test, *Archives of Clinical Neuropsychology* **11**, 541–552.
- [8] Reitan, R.M. & Wolfson, D. (1996). The question of validity of neuropsychological test scores among head-injured litigants: Development of a dissimulation index, *Archives of Clinical Neuropsychology* **11**, 573–580.
- [9] Schmitter-Edgecombe, M., Marks, W., Wright, M.J. & Ventura, M. (2004). Retrieval inhibition in directed forgetting following severe closed-head injury, *Neuropsychology* **18**, 104–114.
- [10] Suchy, Y., Blint, A. & Osmon, D.C. (1997). Behavioral dyscontrol scale: criterion and predictive validity in an inpatient rehabilitation unit sample, *The Clinical Neuropsychologist* **11**, 258–265.

YANA SUCHY

New Item Types and Scoring

FRITZ DRASGOW AND KRISTA MATTERN

Volume 3, pp. 1398–1401

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

New Item Types and Scoring

Computers have been used to administer tests since the 1960s [2]. At first, computerized tests were just that; tests were still comprised of the traditional multiple-choice questions, but examinees read the items from a computer screen or teletype rather than a printed test booklet. Improvements in the technological capabilities of the computer over the decades have revolutionized the field of testing and assessment. (*see Computer-Adaptive Testing*) Advances in hardware capabilities (e.g., memory, hard disks, video adaptor cards, sound cards) and software (e.g., graphical user interfaces, authoring tools) have enabled test developers to create a variety of new item types. Researchers and test developers have devised these new item types to assess skills and performance abilities not well measured by traditional item types, as well as to measure efficiently characteristics traditionally assessed by a multiple-choice format.

In this entry, a review of many of these new item types is presented. Moreover, issues related to scoring innovative items are described. Because the area is rapidly growing and changing, it is impossible to give an exhaustive review of all developments; however, some of the most important advancements in new item types and related scoring issues are highlighted.

Simulations

Recently, several licensing and credentialing boards have concluded that traditional multiple-choice items may adequately assess candidates' factual knowledge of a given domain, but fail to measure the skills and performance abilities necessary for good job performance. For example, once candidates pass the Uniform Certified Public Accountant Examination (UCPAE), they, as accountants, do not answer multiple-choice questions on the job; instead they must often address vague and ambiguous problems posed to them by clients. As a result, many licensing boards have begun incorporating simulated workplace situations in their exams. These simulations are designed to faithfully reflect important challenges faced by employees on the job; adequate responses are required for satisfactory performance.

In April of 2004, the American Institute of Certified Public Accountants (AICPA) implemented a new computerized UCPAE comprising multiple-choice items and simulations. The simulations require examinees to perform activities that are required to answer questions posed by clients. Consequently, they may enter information into spreadsheets, complete cash flow reports, search the definitive literature (which is available as a on-line resource), and write a letter to the client that provides their advice and states its justification. Clearly, the advantage of this assessment format is that it provides an authentic test of the skills and abilities necessary to be a competent accountant.

One disadvantage of the simulation format is that many of the items are highly interdependent due to the nature of the task. For example, a typical simulation, which might require about 20 responses, might require an examinee to enter several values in a balance sheet. If an examinee's first entry in the balance sheet is incorrect, it is unlikely that his/her entry in the next cell will be correct. Having two highly intercorrelated items (i.e., the interitem correlation may exceed 0.90) is problematic because, essentially, the same question is asked twice. The examinee who correctly answers the first item is rewarded twice, whereas the candidate who answers the first item incorrectly is penalized twice. As a result, the AICPA is considering alternative scoring methods, such as not scoring redundant items or giving examinees one point if they correctly answer all parts of redundant sets of items and zero points if they do not.

In 1999, National Board of Medical Examiners (NBME) changed its licensing test to a computerized format, which included simulated patient encounters [5]. These simulations were developed to assess critical diagnostic skills required to be a competent physician. The candidate assumes the role of the physician and must treat simulated patients with presenting symptoms. On the basis of these symptoms, the candidate can request medical history, order a test, request a consultation, or order a treatment; virtually any action of a practicing physician can be taken. As in the real world, these activities take time. Therefore, if the candidate orders a test that requires an hour to complete, the candidate must advance an on-screen clock by one hour in order to obtain the results. Simultaneously, the patient's symptoms and condition also progress by one hour. Clearly, this simulation format provides a highly authentic test of the

2 New Item Types and Scoring

skills required to be a competent physician; however, scoring becomes very complicated because candidates can take any of literally thousands of actions.

Items that Incorporate Media

Early computerized test development efforts focused on multiple-choice items were presented in a format that included text and simple graphics [7]. However, incorporating media within a test has improved measurement in tests such as Vispoel's [10] tonal memory computer adaptive test (CAT) that assesses musical aptitude, Ackerman, Evans, Park, Tamassia, and Turner's [1]. Test of Dermatological Skin Disorders that assesses the ability to identify and diagnose different types of skin disorders, and Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, and Donovan's [8] Conflict Resolution Skills Assessment that measures the interpersonal skill of conflict resolution.

Vispoel's [10] tonal memory CAT improves on traditional tests of music aptitude that are administered via paper and pencil with a tape recorder controlled by the test administrator at the front of the room. Biases such as seat location and disruptive neighbors are eliminated because each examinee completes the test via computer with a pair of headphones, which offers superior sound quality to that of a tape cassette. Vispoel [10] demonstrated that fewer questions have to be administered because the test is adaptive, which reduced problems of fatigue and boredom; furthermore, the test pace is controlled by the examinee rather than the administrator. Many of the benefits of Vispoel's tonal memory CAT are analogous to the benefits of Ackerman et al.'s Test of Dermatological Skin Disorders [1]. For example, this test, which scores examinees on their ability to identify skin disorders, is improved when computerized because the high-resolution color image on a computer monitor is superior to that of a slide projector. Moreover, the computerized test allows examinees to zoom on an image. Note that examinees seated at the back of the room are no longer at a visual disadvantage, and, the respondent, not the test administrator, controls the pace of item administration.

Olson-Buchanan et al.'s [8] Conflict Resolution Skills Assessment presents video clips of realistic workplace problems a manager might encounter and then asks the examinee what he/she would do in each situation. Its developers believe that examinees become more psychologically involved in the

situation when watching the video clips than when reading passages. Chan and Schmidt [4] found that reading comprehension correlated 0.45 with a paper-and-pencil assessment of an interpersonal skill, but only 0.05 with a video-based version. Thus, it appears that the video format enables a more valid assessment of interpersonal skill by reducing the influence of cognitive ability. On the other hand, scoring these types of items presents a challenge. Is it best to use subject matter experts to determine the scoring or should empirical keying be used? Are the wrong options equally wrong or are some better than others? If so, how do you score them? Test developers are investigating the best way to score these item types.

Innovative Mathematical Item Types

Traditional standardized tests assessing mathematical ability such as the American College Testing (ACT) and the Scholastic Assessment Test (SAT) rely solely on a multiple-choice format and, consequently, examinees have a high probability (25% if there are four options) of answering correctly by simply guessing. However, three novel mathematics item types, *mathematical expressions*, *generating examples*, and *graphical modeling*, require examinees to construct their own response [3], which reduces the probability of answering these problems correctly by guessing to roughly zero.

Mathematical expressions are items that require examinees to construct a numerical equation to represent the problem [3]. For example, if x is the mean of 7, 8, and y , what is the value of y in terms of x ? The obvious solution is $y = 3x - 7 - 8$; however, $y = 3x - 15$, $y = 3x - (7 + 8)$, and $y = (6x - 14 - 16)/2$ are also correct. In fact, there are an infinite number of correct solutions, which makes scoring problematic. Fortunately, Bennett et al. [3] developed a scoring algorithm, which converts answers to their simplest form. In a test of the scoring algorithm, Bennett et al. (2000) found the accuracy to be 99.6%.

Generating examples are items that require examinees to construct examples that fulfill the constraints given in the problem [3]. For example, suppose x and y are positive integers and $2x + 4y = 20$; provide two possible solution sets for x and y (Solution: $x = 4$, $y = 3$ and $x = 6$, $y = 2$). Again, there is more than one right answer, and therefore scoring is not as straightforward as with multiple-choice items.

Graphical modeling items require examinees to construct a graphical representation of the information provided in the problem [3]. For example, examinees may be asked to use a grid to draw a triangle that has an area of 12. As before, there are many solutions to this problem. These problems require the test developer to create and validate more extensive scoring rules than a simple multiple-choice key. Nonetheless, by requiring examinees to construct responses, Bennett et al. [3] showed that these item types have enriched the assessment of mathematical ability.

Innovative Verbal Item Types

Like standardized tests of mathematical ability, most tests of verbal ability rely on multiple-choice items. However, recent innovations in the assessment of verbal ability may change the status quo. The following section describes two examples of innovation in verbal assessment: *passage editing* and *essay with automated scoring*.

A traditional item type used to assess verbal ability requires examinees to read a sentence with a section underlined and choose from a list of options the alternative that best corrects the section grammatically. The problem with this type of item is that it points the examinee to the error by underlining it. Would examinees know there was an error if it was not pointed out for them? Passage editing bypasses this problem because it requires examinees to locate an error within a passage (nothing is underlined) as well as correct it [6, 11]. This results in a more rigorous test of verbal ability than the traditional format because it assesses both the ability to locate errors as well as fix them.

It is difficult to dispute the argument that having an examinee write an essay is the most authentic test of writing ability. Unfortunately, essays are often omitted from standardized tests because they are costly and time-consuming to score. However, innovations in computerized scoring such as *e-rater* [9] make the inclusion of essays a realistic option. *E-rater* is a computerized essay grader that utilizes scoring algorithms based on the rules of natural language processing. In an interesting study examining the validity of *e-rater*, Powers et al. [9] invited a wide variety of people to submit essays that might be scored incorrectly by the computer. A Professor of Computational Linguistics received the highest possible score by

writing a paragraph and then repeating it 37 times. *E-rater* has since been improved, and Powers et al. [9] state that *e-rater* is now capable of correctly scoring most of the essays that had previously tricked it.

Conclusion

Psychological testing and assessment has been revolutionized by the advent of the computer, and, specifically, improvements in the capabilities of modern computers. As a result, test developers have explored a wide variety of innovative approaches to assessment; however, there is a cost. Scoring innovative items types is rarely as straightforward as a multiple-choice key and often quite complex. This review presents some of the newest innovations in item types along with related scoring issues. These innovations in item types and concomitant scoring algorithms represent genuine improvement in psychological testing.

References

- [1] Ackerman, T.A., Evans, J., Park, K.S., Tamassia, C. & Turner, R. (1999). Computer assessment using visual stimuli: a test of dermatological skin disorders, in *Innovations in Computerized Assessment*, F. Drasgow & J.B. Olson-Buchanan, eds, Lawrence Erlbaum, Mahwah, pp. 137–150.
- [2] Bartram, D. & Bayliss, R. (1984). Automated testing: past, present and future, *Journal of Occupational Psychology* **57**, 221–237.
- [3] Bennett, R.E., Morely, M. & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests, *Applied Psychological Measurement* **24**, 294–309.
- [4] Chan, D. & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions, *Journal of Applied Psychology* **82**, 143–159.
- [5] Clyman, S.G., Melnick, D.E. & Clauser, B.E. (1999). Computer-based case simulations from medicine: assessing skills in patient management, in *Innovative Simulations for Assessing Professional Competence*, A. Tekian, C.H. McGuire & W.C. McGahie, eds, University of Illinois, Department of Medical Education, Chicago, pp. 29–41.
- [6] Davey, T., Godwin, J. & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment, *Journal of Educational Measurement* **34**, 21–41.
- [7] McBride, J.R. (1997). The Marine Corps exploratory development project: 1977–1982, in *Computerized*

4 New Item Types and Scoring

- Adaptive Testing: From Inquiry to Operation*, W.A. Sands, B.K. Waters & J.R. McBride, eds, American Psychological Association, Washington, pp. 59–67.
- [8] Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A. & Donovan, M.A. (1998). Interactive video assessment of conflict resolution skills, *Personnel Psychology* **51**, 1–24.
- [9] Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E. & Kukich, K. (2002). Stumping *e-rater*: challenging the validity of automated essay scoring, *Computers in Human Behavior* **18**, 103–134.
- [10] Vispoel, W.P. (1999). Creating computerized adaptive tests of musical aptitude: problems, solutions, and future directions, in *Innovations in Computerized Assessment*, F. Drasgow & J.B. Olson-Buchanan, eds, Lawrence Erlbaum, Mahwah, pp. 151–176.
- [11] Zenisky, A.L. & Sireci, S.G. (2002). Technological innovations in large-scale assessment, *Applied Measurement in Education* **15**, 337–362.

FRITZ DRASGOW AND KRISTA MATTERN

Neyman, Jerzy

DAVID C. HOWELL

Volume 3, pp. 1401–1402

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Neyman, Jerzy

Born: April 16, 1894, Bendery, Russia.

Died: August 5, 1981, Berkeley, CA.

Jerzy Neyman spent the first half of his life moving back and forth between Russia, Poland, England, and France but in the last half of his life firmly ensconced in Berkeley, California. In the process, he helped lay the foundation of modern statistical theory and built one of the major statistics departments in the United States.

Jerzy Neyman was born to a Polish family in Russia in 1894. His father was an attorney, and Neyman was educated at home until he was 10, at which point he was fluent in five languages. When his father died, the family moved to Kharkov, where he attended school and eventually university. While at university, he read the work of Lebesgue and proved five theorems on the Lebesgue integral on his own. At the end of the First World War, he was imprisoned briefly. He remained at Kharkov, teaching mathematics until 1921, at which point he moved to Poland. He took a position as a senior statistical assistant at the Agricultural Institute but soon moved to Warsaw to pursue his PhD., which he received in 1924.

In 1925, Neyman won a Polish Government Fellowship to work with **Karl Pearson** in London, but he was disappointed with Pearson's knowledge of mathematical statistics. In the next year, he won a Rockefeller Fellowship to study pure mathematics in Paris.

While in London, Neyman had formed a friendship with Pearson's son **Egon Pearson**, and they began to collaborate in 1927 while Neyman was still in Paris. This collaboration had profound effects on the history and theory of statistics.

In 1928, Neyman returned again to Poland to be head of the Biometric Laboratory at the Nencki Institute of Warsaw. He remained there until returning to London in 1934 but continued his collaboration with Egon Pearson throughout that period.

The Neyman–Pearson approach to hypothesis testing is outlined elsewhere (*see* **Neyman–Pearson Inference**) and will not be elaborated on here. But it is worth pointing out that their work developed the concepts of the *alternative hypothesis*, **power**,

Type II errors, 'critical region', and the 'size' of the critical region, which they named the *significance level*. Neyman had also developed the concept of a '**confidence interval**' while living in Poland, though he did not publish that work until 1937.

In 1934, Neyman returned to London to take a position in Pearson's department, and they continued to publish together for another four years. They clashed harshly with **Fisher**, who was at that time the head of the Eugenics department at University College, which was the other half of Karl Pearson's old department. Those battles are well known to statisticians and were anything but polite – especially on Fisher's side.

In 1938, Neyman accepted a position of Professor of Mathematics at Berkeley and moved to the United States, where he was to remain. He established and directed the Statistics Laboratory at Berkeley, and in 1955 he founded and chaired the Department of Statistics. O'Connor and Robinson [2] note that his efforts at establishing a department were not always appreciated by all at the University. An assistant to the President of Berkeley once wrote:

Here, a willful, persistent and distinguished director has succeeded, step by step over a fifteen-year period, against the wish of his department chairman and dean, in converting a small 'laboratory' or institute into, in terms of numbers of students taught, an enormously expensive unit; and he argues that the unit should be renamed a 'department' because no additional expense will occur.

That quotation will likely sound familiar to those who are members of academic institutions.

In 1949, he published the first report of the Symposium of Mathematical Statistics and Probability, and this symposium became the well-known Berkeley Symposium, which continued every five years until 1971.

Additional material on Neyman's life and influence can be found in [1], [3], and [4].

References

- [1] Chiang, C.L. & Jerzy Neyman, (1894–1981). *Found on the Internet at* <http://www.amstat.org/about/statisticians/index.cfm?fuseaction=biobio&BioID=11>.
- [2] O'Connor, J.J. & Robertson, E.F. Neyman, J. (2004). *Found on the Internet at* <http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Neyman.html>.

2 Neyman, Jerzy

- [3] Pearson, E.S. The Neyman-Pearson story: 1926–1934. Historical sidelights on an episode in Anglo-Polish collaboration, Festschrift for J Neyman, New York, 1966.
- [4] Reid, C. (1998). *Neyman*, Springer Verlag, New York.

DAVID C. HOWELL

Neyman–Pearson Inference

RAYMOND S. NICKERSON

Volume 3, pp. 1402–1408

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Neyman–Pearson Inference

Jerzy Neyman [9] credits **Pierre Laplace** [5] with being the first to attempt to test a statistical hypothesis. Prominent among the many people who have contributed to the development of mathematical statistics from the time of Laplace until now are the developers of schools of thought that have dominated the use of statistics by psychologists for the analysis of experimental data. One of these schools of thought derives from the work of **Ronald A. Fisher**, the other from that of Neyman and **Egon Pearson** (son of **Karl Pearson**, for whom the Pearson correlation coefficient was named). A third school, emanating from the work of the **Reverend Thomas Bayes**, has had considerable impact on the theorizing of psychologists, especially in their efforts to develop normative and descriptive models of decision making and choice, but it has been applied much less to the analysis of experimental data, despite eloquent pleas by proponents of **Bayesian analysis** for such an application.

The Collaboration

The collaboration between Neyman and E. Pearson began when the former spent a year as a visitor in Karl Pearson's laboratory at University College, London in 1925. It lasted for the better part of a decade during which time the two co-authored several papers. The collaboration, which has been recounted by Pearson [15] in a Festschrift for Neyman, was carried on primarily by mail with Neyman in Poland and Pearson in England.

In their first paper on hypothesis testing, Neyman and Pearson [10] set the stage for the ideas they were to present this way: 'One of the most common as well as most important problems which arise in the interpretation of statistical results, is that of deciding whether or not a particular sample may be judged as likely to have been randomly drawn from a certain population, whose form may be either completely or only partially specified. We may term Hypothesis A¹ the hypothesis that the population from which the sample $\sum$ has been randomly drawn is that specified, namely Π . In general the method of procedure is to apply certain tests or criteria, the results of which will

enable the investigator to decide with a greater or less degree of confidence whether to accept or reject Hypothesis A, or, as is often the case, will show him that further data are required before a decision can be reached' (p. 175).

In this paper, they work out the implications of certain methods of testing, assuming random sampling from populations with normal, rectangular, and exponential distributions. They emphasize the importance of the assumption one makes about the shape of the underlying distribution in applying statistical methods and about the randomness of sampling (the latter especially in the case of small samples). Later, Pearson [15] described this paper as having 'some of the character of a research report, putting on record a number of lines of inquiry, some of which had not got very far. The way ahead was open; for example, five different methods of approach were suggested for the test of the hypothesis that two different samples came from normal populations having a common mean!' (p. 463).

Neyman and Pearson versus Fisher

As Macdonald [6] has pointed out, both the Fisherian and Neyman–Pearson approaches to statistical inference were concerned with establishing that an observed effect could not plausibly be attributed to sampling error. Perhaps because of this commonality, the blending of the ideas of Fisher with those of Neyman and Pearson has been so thorough, and so thoroughly assimilated by the now conventional treatment of statistical analysis within psychology, that it is difficult to distinguish the two schools of thought. What is generally taught to psychology students in introductory courses on statistics and experimental design makes little, if anything, of this distinction. Typically, statistical analysis is presented as an integrated whole and who originated specific ideas is not stressed. Seldom is it pointed out that the Fisherian approach to statistical analysis and that of Neyman and Pearson were seen by their proponents to be incompatible in several ways and that the rivalry between these pioneers is a colorful chapter in the history of applied statistics. An account of the disagreements between Fisher and Karl Pearson, which were especially contentious, and how they carried over to the relationship between Fisher and Neyman and Egon Pearson has been told with flair by Salsburg [16].

Neyman and Pearson considered Fisher’s method of hypothesis testing, which posed questions that sought yes–no answers – yes, the hypothesis under test is rejected, no it is not – to be too limited. ‘It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from Π would not in itself justify the rejection of the hypothesis that it had been so drawn, if there were no other more probable hypotheses conceivable’ [10, p. 178]. They proposed an approach to take account of how the weight of evidence contributes to the relative statistical plausibility of each of two mutually exclusive hypotheses.² So, in distinction from Fisher’s yes–no approach, Neyman and Pearson’s may be characterized as either–or, because it leads to a judgment in favor of one or the other of the candidate hypotheses. ‘[I]n the great majority of problems we cannot so isolate the relation of $\sum$ to Π ; we reject Hypothesis A not merely because $\sum$ is of rare occurrence, but because there are other hypotheses as to the origin of $\sum$ which it seems more reasonable to accept’ [10, p. 183].

Neyman and Pearson’s approach to statistical testing differed from that of Fisher in other respects as well. Fisher prohibited acceptance of the null hypothesis (the hypothesis of no difference between two samples, or no effect of an experimental manipulation); if the results did not justify rejection of the null hypothesis, the most that could be said was that the null hypothesis could not be rejected. (One of the concerns that have been expressed about the effect of this aspect of null-hypothesis testing is that too often ‘failure’ to reject the null hypothesis has been equated, inappropriately, with failure of a research effort.) The Neyman–Pearson approach permitted acceptance of either of the hypotheses under consideration, including the null if it was one of them, although it provided for making the acceptance (and hence erroneous acceptance) of one of the hypotheses more difficult than acceptance of the other. Other distinguishing features of the Neyman–Pearson approach include the distinction between error types, the conception of sample spaces, the pre-experiment specification of decision criteria, and the notion of the power of a test.

As to the kinds of conclusions that can be drawn from the results of statistical tests, Neyman and Pearson describe them as ‘rules of behaviour’, and they explicitly contrast this with the idea that one can take the results as indicative of the probability that a hypothesis of interest is true or false. In the

first of two 1933 papers, Neyman and Pearson [12] address the question: ‘Is it possible that there are any efficient tests of hypotheses based upon the theory of probability, and if so, what is their nature?’ (p. 290). Their answer is a carefully qualified yes. ‘Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a “rule of behaviour”: to decide whether a hypothesis, H , of a given type be rejected or not, calculate a specified character, x , of the observed facts; if $x > x_0$ reject H , if $x \leq x_0$ accept H . Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false’ (p. 291).

Error Types and Level of Significance

Neyman and Pearson recognized that there are two possible ways to be right, and two possible ways to be wrong, in drawing a conclusion from a test, and that the two ways of being right (wrong) may not be equally desirable (undesirable). In particular, they stressed that deciding in favor of H_0 when H_1 is true may be a more (or less) acceptable error than deciding in favor of H_1 when H_0 is correct. ‘These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. . . From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator’ [12, p. 296]. The idea that the main purpose of statistical testing is to identify and learn from error has been expounded in some detail in [7].

Assuming that in most cases of hypothesis testing the two types of error would not be considered equally serious, Neyman and Pearson defined as ‘the error of the first kind’ (generally referred to as Type I error) the error that one would most wish to avoid. ‘The first demand of the mathematical theory is to

deduce such test criteria as would ensure that the probability of committing an error of the first kind would equal (or approximately equal, or not exceed) a preassigned number α , such as $\alpha = 0.05$ or 0.01 , etc. This number is called the level of significance' [9, p. 161].

It is important to stress that α is a conditional probability, specifically the probability of rejecting H_0 – here referring to the null hypothesis in the sense of the hypothesis under test, not necessarily the hypothesis of no difference – conditional on its being true. And its use is justified only when certain assumptions involving random sampling and the parameters of the distributions from which the sampling is done are met. This conditional-probability status of α is often not appreciated and it is treated as the absolute probability of occurrence of a Type I error. But, by definition, a Type I error can be made only when H_0 is true; the unconditional probability of the occurrence of a Type I error is the product of the probability that H_0 is true and the probability of a Type I error given that it is true.

A similar observation pertains to Type II error (failure to reject H_0 when it is false, i.e., failure to accept H_1 when it is true), which, by definition, can be made only when H_0 is false. The probability of occurrence of a Type II error, conditional on H_0 being false, is usually referred to as β (beta), and the unconditional probability of a Type II error is the product of β and the probability that H_0 is false. Computing the absolute unconditional probabilities of Type I and Type II errors generally is not possible, because the probability of H_0 being true (or false) is not known and, indeed, whether to behave as though it were true, to use Neyman and Pearson's term, is what one wants to determine by the application of statistical procedures. Failure to recognize the conditionality of α and β has led to much confusion in the interpretation of results of statistical techniques [14].

Sample Spaces and Critical Areas

Neyman and Pearson conceptualize the outcome of an experiment in terms of the sample spaces that represent the possible values a specified statistic could take under each of the hypotheses of interest. A point in a sample space represents a particular experimental outcome, if the associated hypothesis is true. To test a hypothesis, one divides the sample

space for that hypothesis into two regions and then applies the rule of accepting the hypothesis if the statistic falls in one of these regions and rejecting it if it falls in the other. In null-hypothesis testing, the convention is to define a region, known as the **critical region**, and to reject the hypothesis if the statistic falls in that region.

The situation may be represented as in Figure 1. Each distribution represents the theoretical distribution – the sample space – of some statistic assuming the specified hypothesis is true. The vertical line (decision criterion) represents the division of the space into two regions, the critical region being to the right of the line.³ The area under the curve representing H_0 and to the right of the criterion represents

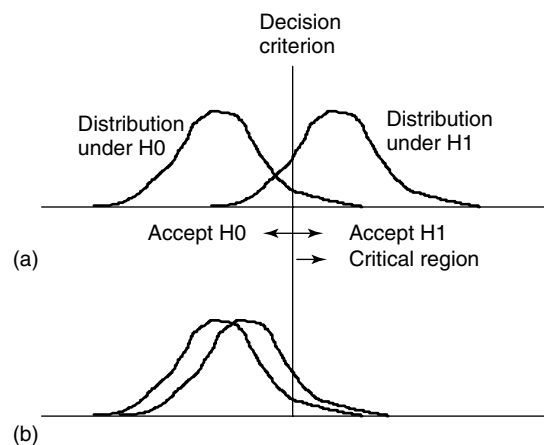


Figure 1 In panel a, the distribution to the left represents the distribution under H_0 ; that to the right represents the distribution under H_1 . The area under the H_0 distribution and to the right of the decision criterion represents α , the probability of a Type I error conditional on H_0 being true; that under the H_1 distribution and to the left of the criterion represents β , the probability of a Type II error conditional on H_1 being true; the area under the H_1 distribution and to the right of the criterion, $1 - \beta$, is said to be the power of the test, which is the probability of deciding in favor of H_1 conditional on its being true. One decreases the conditional probability of a Type I error by moving the decision to the right, but in doing so, one necessarily increases the conditional probability of a Type II error. Panel b shows the situation assuming a much greater overlap between the two distributions, illustrating that a greater degree of overlap means necessarily a greater error rate. In this case, holding the decision criterion so as to keep the conditional probability of a Type I error constant means greatly increasing the conditional probability of a Type II error

the probability of a Type I error on the condition that H_0 is true; the area under the curve representing H_1 and to the left of the decision criterion represents the probability of a Type II error, on the condition that H_1 is true. As is clear from the figure, given the distributions shown, the probability of a Type I error, conditional on H_0 being true, can be decreased only at the expense of an increase in the probability of a Type II error, conditional on H_1 being true.

Figure 1 has been drawn to represent cases in which the theoretical distributions of H_0 and H_1 overlap relatively little (a) and relatively greatly (b). Obviously, the greater the degree of overlap, the more a stringent criterion for minimizing the conditional probability of a Type I error will cost in terms of allowance of a high conditional probability of a Type II error. We should also note that, as drawn, the figures show the distributions having the same variance and indeed the same Gaussian shape. Such are the default assumptions that are often made in null-hypothesis testing; however, the theoretical distributions used need not be Gaussian, and procedures exist for taking unequal variances into account.

Neyman and Pearson [12] devote most of their first 1933 paper to a discussion of the nontrivial problem of specifying a ‘best critical region’ for controlling Type II error, for a criterion that has been selected to yield a specified (typically small) conditional probability of a Type I error. Such a selection must depend, of course, on what the alternative hypothesis(es) is (are). When there are more alternative hypotheses than one, the best critical region with regard to one of the alternative hypotheses will not generally be the best critical region with respect to another. Neyman and Pearson show that in certain problems it is possible to specify a ‘common family of best critical regions for H_0 with regard to the whole class of admissible alternative hypotheses’ (p. 297). They develop procedures for identifying best critical regions in the sense of maximizing the probability of accepting H_1 when it is true, given a fixed criterion for erroneously rejecting H_0 when it is true.

Power

In setting the decision criterion so as to make the conditional probability of a Type I error very small, one typically ensures that the conditional probability of a Type II error, β , is relatively large, which means

that the chance of failing to note a real effect, if there is one, is fairly high. Neyman and Pearson [13] define the **power** of a statistical test with respect to the alternative hypothesis as $1-\beta$, which is the probability of detecting an effect (rejecting the null) if there is a real effect to be detected (if the null is false).

The power of a statistical test is a function of three variables, as can be seen from Figure 1: the **effect size** (the difference between the means of the distributions), relative variability (the variance or standard deviation of the distributions), and the location of the decision criterion. Often, sample size is mentioned as one of the determinants of power, but this is because the relative variability of a distribution decreases as sample size is increased. Variability can be affected in other ways as well (*see Power*).

Statistical Tests and Judgment

In view of the tendency of some textbook writers to present hypothesis testing procedures, including those of Neyman and Pearson, as noncontroversial formulaic methods that can be applied in cook-book fashion to the problem of extracting information from data, it is worth noting that Neyman and Pearson were sensitive to the need to exercise good judgment in the application of statistical methods to the testing of hypotheses and they characterized the methods as themselves being aids to judgment; they would undoubtedly have been appalled at how the techniques they developed are sometimes applied with little thought as to whether they are really appropriate to the situation and the results interpreted in a categorical fashion.

In general, Neyman and Pearson saw statistical tests as attempts to quantify commonsense notions about the interpretation of probabilistic data. ‘In testing whether a given sample, $\sum$, is likely to have been drawn from a population Π , we have started from the simple principle that appears to be used in the judgments of ordinary life – that the degree of confidence placed in an hypothesis depends upon the relative probability or improbability of alternative hypotheses. From this point of view any criterion which is to assist in scaling the degree of confidence with which we accept or reject the hypothesis that $\sum$ has been randomly drawn from Π should be one which decreases as the probability (defined in some definite manner) of alternative hypotheses becomes

relatively greater. Now it is of course impossible in practice to scale the confidence with which we form a judgment with any single numerical criterion, partly because there will nearly always be present certain *a priori* conditions and limitations which cannot be expressed in exact terms. Yet though it may be impossible to bring the ideal situation into agreement with the real, some form of numerical measure is essential as a guide and control' [11, p. 263].

So, while Neyman and Pearson argue the need for quantitative tests, they stress that they are to be used to aid judgment and not to supplant it. '[T]ests should only be regarded as tools which must be used with discretion and understanding, and not as instruments which in themselves give the final verdict' [10, p. 232]. Again, 'The role of sound judgment in statistical analysis is of great importance and in a large number of problems common sense may alone suffice to determine the appropriate method of attack' [12, p. 292].

They note, in particular, the need for judgment in the interpretation of likelihood ratios, the introduction of which they credit to Fisher [2]. The likelihood ratio is the ratio of the probability of obtaining the sample in hand if H_1 were true to the probability of obtaining the sample if H_0 were true. But, the likelihood ratio is to be weighed along with other, not necessarily numerical considerations in arriving at a conclusion. 'There is little doubt that the criterion of likelihood is one which will assist the investigator in reaching his final judgment; the greater be the likelihood of some alternative Hypothesis A' (not ruled out by other considerations), compared with that of A, the greater will become his hesitation in accepting the latter. It is true that in practice when asking whether Σ can have come from Π , we have usually certain *a priori* grounds for believing that this may be true, or if not so, for expecting that Π' differs from Π in certain directions only. But such expectations can rarely be expressed in numerical terms. The statistician can balance the numerical verdict of likelihood against the vaguer expectations derived from *a priori* considerations, and it must be left to his judgment to decide at what point the evidence in favour of alternative hypotheses becomes so convincing that Hypothesis A must be rejected' [10, p. 187].

Often Neyman and Pearson speak of taking *a priori* beliefs into account in selecting testing procedures or interpreting the results of tests. They

sometimes note that more than one approach to statistical testing can be taken with equal justification, making the choice somewhat a matter of personal taste. And they emphasize the need for discretion or an 'attitude of caution' in the drawing of conclusions from the results of tests.

They viewed statistical analysis as only one of several reasons that should weigh in an investigator's decision as to whether to accept or reject a particular hypothesis, and noted that the combined effect of those reasons is unlikely to be expressible in numerical terms. 'All that is possible for [an investigator] is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon *a priori* or *a posteriori* considerations. The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision; one man may prefer to use one method, a second another, and yet in the long run there may be little to choose between the value of their conclusions' [10, p. 176].

Neyman and Pearson give examples of situations in which they would consider it reasonable to reject a hypothesis on the basis of common sense despite the outcome of a statistical test favoring acceptance. They stress the importance of a 'clear understanding [in the mind of a user of a statistical test] of the process of reasoning on which the test is based' [10, p. 230], and note that that process is an individual matter: 'we do not claim that the method which has been most helpful to ourselves will be of greatest assistance to others. It would seem to be a case where each individual must reason out for himself his own philosophy' [10, p. 230]. In sum, Neyman and Pearson presented their approach to statistical hypothesis testing in considerably less doctrinaire terms than methods of hypothesis testing are often presented today.

Other Sources

Other sources of information on the Neyman–Pearson school of thought regarding statistical significance testing and the differences between it and the Fisherian school include [1, 3, 4, 7], and [8].

Notes

1. Neyman and Pearson use a variety of letters and subscripts to designate hypotheses under test. Except

when quoting Neyman and Pearson, I here use H_0 and H_1 to represent respectively the main (often null) hypothesis under test and an alternative, incompatible, hypothesis. In practice, H_1 is often the negation of H_0 .

2. In [11], Neyman and Pearson distinguish between a simple hypothesis (e.g., ‘that $\sum$ has been drawn from an exactly defined population Π ’) and a composite hypothesis (e.g., ‘that $\sum$ has been drawn from an unspecified population belonging to a clearly defined subset ω of the set Ω ’). The distinction is ignored in this article because it is not critical to the principles discussed.
3. In ‘two-tailed’ testing, appropriate when the alternative hypothesis is that there is an effect (e.g., a difference between means), but the direction is not specified, the critical region can be composed of two subregions, one on each end of the distributions.

References

- [1] Cowles, M.P. (1989). *Statistics in Psychology: A Historical Perspective*, Lawrence Erlbaum, Hillsdale.
- [2] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A* **222**, 309–368.
- [3] Gigerenzer, G. & Murray, D.J. (1987). *Cognition as Intuitive Statistics*, Lawrence Erlbaum, Hillsdale.
- [4] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*, Cambridge University Press, New York.
- [5] Laplace, P.S. (1812). *Théorie Analytique Des Probabilités*, Académie Française, Paris.
- [6] Macdonald, R.R. (1997). On statistical testing in psychology, *British Journal of Psychology* **88**, 333–347.
- [7] Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago.
- [8] Mulaik, S.A., Raju, N.S. & Harshman, R.A. (1997). There is a time and a place for significance testing: appendix, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum, Mahwah, pp. 103–111.
- [9] Neyman, J. (1974). The emergence of mathematical statistics: a historical sketch with particular reference to the United States, in *On the History of Statistics and Probability*, D.B. Owen, ed., Marcel Dekker, New York, pp. 147–193.
- [10] Neyman, J. & Pearson, E.S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika* **20A**, 175–240.
- [11] Neyman, J. & Pearson, E.S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II, *Biometrika* **20A**, 263–294.
- [12] Neyman, J. & Pearson, E.S. (1933a). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society, A* **231**, 289–337.
- [13] Neyman, J. & Pearson, E.S. (1933b). The testing of statistical hypotheses in relation to probability a priori, *Proceedings of the Cambridge Philosophical Society* **29**, 492–510.
- [14] Nickerson, R.S. (2000). Null hypothesis statistical testing: a review of an old and continuing controversy, *Psychological Methods* **5**, 241–301.
- [15] Pearson, E.S. (1970). The Neyman-Pearson story: 1926–1934: historical sidelights on an episode in Anglo-Polish collaboration, in *Studies in the History of Statistics and Probability*, E.S. Pearson & M. Kendall, eds, Charles Griffin, London, pp. 455–477. (Originally published in 1966.).
- [16] Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W. H. Freeman, New York.

(See also **Classical Statistical Inference: Practice versus Presentation**)

RAYMOND S. NICKERSON

Nightingale, Florence

DAVID C. HOWELL

Volume 3, pp. 1408–1409

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nightingale, Florence

Born: May 12, 1820, in Florence, Italy.

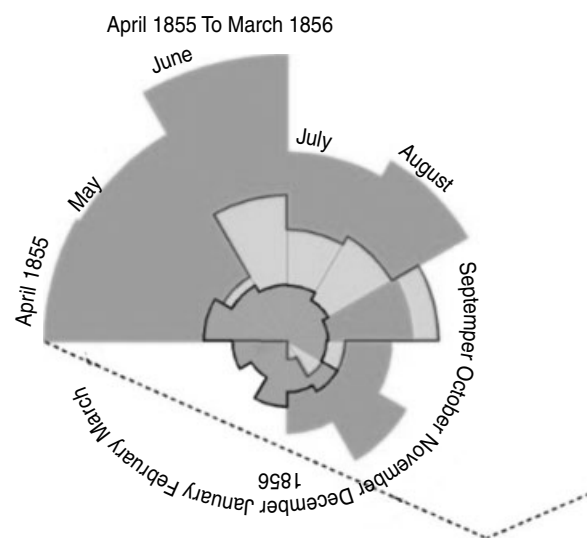
Died: August 13, 1910, in London, England.

Florence Nightingale was born into a wealthy English family early in the nineteenth century and was educated at home by her father. That education included mathematics at some level. She continued her studies of mathematics under James Joseph Sylvester. She was also later influenced by the work of **Quetelet**, who was already well known for the application of statistical methods to data from a variety of fields (see [1] for the influence of Quetelet's writing and a short biography of Nightingale).

Nightingale received her initial training in nursing in Egypt, followed by experience in Germany and France. On returning to England, she took up the position of Superintendent of the *Establishment for Gentlewomen during Illness*, on an unpaid basis.

When the Crimean War broke out in 1854, Nightingale was sent by the British Secretary of War

to Turkey as head of a group of 38 nurses. *The Times* had published articles on the poor medical facilities available to British soldiers, and the introduction of nurses to military hospitals was hoped to address the problem. Nightingale was disturbed by the lack of sanitary conditions and fought the military establishment to improve the quality of care. Most importantly to statistics, she collected data on the causes of death among soldiers and used her connections to publicize her results. She was able to show that soldiers were many times more likely to die from illnesses contracted as a result of poor sanitation in the hospitals or from wounds left untreated than to die from enemy fire. To illustrate her point, she plotted her data in the form of polar area diagrams, whose areas were proportional to the cause of deaths. Such a diagram is shown below (Figure 1), and can be found in [2]. Diagrams such as this have since been referred to as *Coxcombs*, although Small [3] has argued that Nightingale used that term to refer to the collection of such graphics. This graphic was published 10 years before the famous graphic by Minard on the size of Napoleon's army in the invasion of Russia. Unlike



The area of the medium, light, and dark gray wedges are each measured from the center as the common vertex.

The medium wedges measured from the center of the circle represent area for area the deaths from Preventable or Mitigable Zymotic Diseases, the light wedges measured from the center the deaths from wounds, & the dark wedges measured from the center the deaths from all other causes.

Figure 1 Causes of death among military personnel in the Crimean War. (The figure was adapted from <http://www.florence-nightingale-avenging-angel.co.uk/Coxcomb.htm>)

2 Nightingale, Florence

Minard's graphic, it carries a clear social message and played an important role in the reform of military hospitals. It represents an early, though certainly not the first, example of the use of statistical data for social change.

Among other graphics, Nightingale created a simple line graph that showed the death rates of civilian and military personnel during peacetime. The data were further broken down by age. The implications were unmistakable and emphasize the importance of controlling confounding variables. Such a graph can be seen in [3].

Because of her systematic collection of data and use of such visual representations, Nightingale played an important role in the history of statistics. Though she had little use for the theory of statistics, she was influential because of her ability to use statistical data to influence public policy.

Nightingale was elected as the first female Fellow of the Royal Statistical Society in 1858, and was made an honorary member of the American Statistical Association, in 1874. During the 1890s she worked unsuccessfully with **Galton** on the establishment of a Chair of Statistics at Oxford.

For many years of her life she was bed-ridden, and she died in 1910 in London, having published

over 200 books, reports, and pamphlets. Additional material on her life can be found in a paper by Conquest in [4].

References

- [1] Diamond, M. & Stone, M. (1981). Nightingale on Quetelet, *Journal of the Royal Statistical Society, Series A* **144**, 66–79, 176–213, 332–351.
- [2] Nightingale, F.A. (1859). *Contribution to the Sanitary History of the British Army During the Late War with Russia*, Harrison, London.
- [3] Small, H. (1998). Paper from *Stats & Lamps* Research Conference organized by the Florence Nightingale Museum at St. Thomas' Hospital, 18th March 1998. London. Constable. A version of this paper is available at <http://www.florence-nightingale.com/muse.uk/small.htm>
- [4] Stinnett, S. (1990). Women in statistics: sesquicennial activities, *The American Statistician* **44**, 74–80.

(See also **Graphical Methods pre-20th Century**)

DAVID C. HOWELL

Nonequivalent Group Design

CHARLES S. REICHARDT

Volume 3, pp. 1410–1411

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonequivalent Group Design

A nonequivalent group design is used to estimate the relative effects of two or more treatment conditions. In the simplest case, only two treatment conditions are compared. These two conditions could be alternative interventions such as a novel and a standard treatment, or one of the treatment conditions could be a control condition consisting of no intervention. The participants in a nonequivalent group design can be either individuals or aggregates of individuals, such as classrooms, communities, or businesses. Each of the treatments being compared is given to a different group of participants. After the treatment conditions are implemented, the relative effects of the treatments are assessed by comparing the performances across the groups on an outcome measure.

The nonequivalent group design is a quasi-experiment (as opposed to a randomized experiment) (see **Quasi-experimental Designs**) because the participants are assigned to the treatment conditions nonrandomly. Assignment is nonrandom, for example, when participants self-select the treatments they are to receive on the basis of their personal preferences. Assignment is also nonrandom when treatments are assigned by administrators on the basis of what is most convenient or some other nonrandom criterion. Nonequivalent group designs also arise when treatments are assigned to preexisting groups of participants that were created originally for some other purpose without using random assignment.

Differences in the composition of the treatment groups are called *selection differences* [7]. Because the groups of participants who receive the different treatments are formed nonrandomly, selection differences can be systematic and can bias the estimates of the treatment effects. For example, if the participants in one group tend to be more motivated than those in another, and if motivation affects the outcome that is being measured, the groups will perform differently even if the two treatments are equally effective. The primary task in the analysis of data from a nonequivalent group design is to estimate the relative effects of the treatments while taking account of the potentially biasing effects of selection differences.

A variety of statistical methods have been proposed for taking account of the biasing effects

of selection differences [3], [8]. The most common techniques are change-score analysis, **matching** or blocking (using either **propensity scores** or other covariates), **analysis of covariance** (with or without correction for unreliability in the covariates), and selection-bias modeling. These methods all require that one or more pretreatment measures have been collected. In general, the best pretreatment measures are those that are operationally identical to the outcome (i.e., posttreatment) measures. The statistical methods differ in how they use pretreatment measures to adjust for selection differences. Change-score analysis assumes that the size of the average posttreatment difference will be the same as the size of the average pretreatment difference in the absence of a treatment effect. Matching and blocking adjusts for selection differences by comparing participants who have been equated on their pretreatment measures from the treatment conditions. Analysis of covariance is similar to matching, except that equating is accomplished mathematically rather than by physical pairings. Unreliability in the pretreatment measures can be taken into account in the analysis of covariance using **structural equation modeling** techniques [2]. When there are multiple pretreatment measures, **propensity scores** are often used in either matching or analysis of covariance, where a participant's propensity score is the probability that the participant is in one rather than another treatment condition, which is estimated using the multiple pretreatment measures [6]. Selection-bias modeling corrects the effects of selection differences by modeling the nonrandom selection process [1].

Unfortunately, there is no guarantee that any of these or any other statistical methods will properly remove the biases due to selection differences. Each method imposes different assumptions about the effects of selection differences, and it is difficult to know which set of assumptions is most appropriate in any given set of circumstances. Typically, researchers should use multiple statistical procedures to try to 'bracket' the effects of selection differences by imposing a range of plausible assumptions [5]. Researchers can often improve the credibility of their results by adding design features (such as pretreatment measures at multiple points in time and nonequivalent dependent variables) to create a pattern of outcomes that cannot plausibly be explained as a result of selection differences [4]. In addition, because uncertainty about the size of treatment effects

2 Nonequivalent Group Design

is reduced to the extent selection differences are small, researchers should strive to use groups (such as cohorts) that are as initially similar as possible. Nonetheless, even under the best of conditions, the results of nonequivalent group designs tend to be less credible than the results from randomized experiments, because random assignment is the optimal way to make treatment groups initially equivalent.

References

- [1] Barnow, B.S., Cain, G.C. & Goldberger, A.S. (1980). Issues in the analysis of selectivity bias, in *Evaluation Studies Review Annual*, Vol. 5, E.W. Stromsdorfer & G. Farkas, eds, Sage Publications, Newbury Park.
- [2] Magidson, J. & Sobom, D. (1982). Adjusting for confounding factors in quasi-experiments: another reanalysis of the Westinghouse head start evaluation, *Educational Evaluation and Policy Analysis* **4**, 321–329.
- [3] Reichardt, C.S. (1979). The statistical analysis of data from nonequivalent group designs, in *Quasi-experimentation: Design and Analysis Issues for Field Settings*, T.D. Cook & D.T. Campbell, eds, Rand McNally, Chicago, pp. 147–205.
- [4] Reichardt, C.S. (2000). A typology of strategies for ruling out threats to validity, in *Research Design: Donald Campbell's Legacy*, Vol. 2, L. Bickman, ed., Sage Publications, Thousand Oaks, pp. 89–115.
- [5] Reichardt, C.S. & Gollob, H.F. (1987). Taking uncertainty into account when estimating effects, in *Multiple Methods for Program Evaluation*. New Directions for Program Evaluation No. 35, M.M. Mark & R.L. Shotland, eds, Jossey-Bass, San Francisco, pp. 7–22.
- [6] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.
- [7] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
- [8] Winship, C. & Morgan, S.L. (1999). The estimation of causal effects from observational data, *Annual Review of Sociology* **25**, 659–707.

CHARLES S. REICHARDT

Nonlinear Mixed Effects Models

MARY J. LINDSTROM AND HOURI K. VORPERIAN

Volume 3, pp. 1411–1415

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonlinear Mixed Effects Models

Linear mixed-effects (LME) models were developed to model hierarchical (or nested) data. However, before the publication of [2], methods for estimating LME models were not widely available to handle hierarchical data where the observations within a subcategory (often subject) are not equally correlated. This situation usually arises because the observations have been taken over time. This type of hierarchical data is often referred to as growth curve, longitudinal, or (more generally) repeated measures data. We will begin with a quick review of LME and nonlinear regression models to set up ideas and notation.

Linear Mixed-effects Models (LME)

Example: Heights of Girls

The Berkeley growth data [5] includes heights of boys and girls measured between ages 1 to 18 years. In Figure 1, we show a subset of the data for the girls over a range of ages where one might postulate a linear model for the relationship between age and height, for example:

$$y_{ij} = \beta_1 + \beta_2 \text{age}_j + e_{ij} \quad i = 1, \dots, M \quad (1)$$

where M is the number of subjects, y_{ij} is the j th height on the i th subject, β_1 is the intercept, age_j is the j th age, β_2 is the slope for age, e_{ij} is the error term for the j th age for the i th subject.

LME Modeling

We might be tempted to fit the model described in 1 to the data relating to heights of girls. However, the data do not fulfill the assumptions of the typical regression model, that is, $\text{Cov}(e_{ij}, e_{kl}) = 0$ except when $i = k$ and $j = l$. The error terms will not be independent since some come from the same subject and some come from different subjects. In other words, this data has a hierarchical structure that we must take in to account. One additional complication is that the error terms within subject may not be equally correlated since they are observed over time (see **Linear Multilevel Models**).

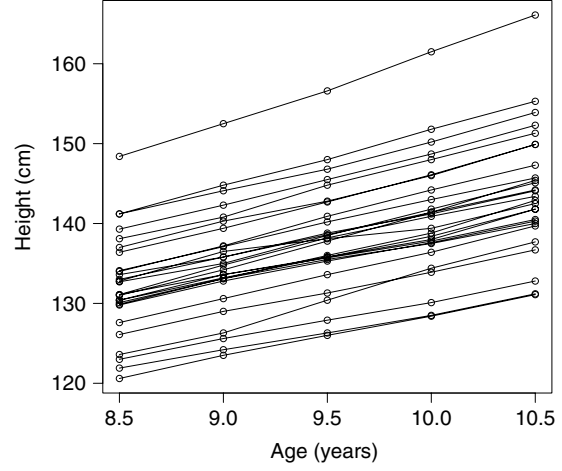


Figure 1 Heights of 27 girls measured at 8.5, 9, 9.5, 10, and 10.5 years

Figure 1 indicates that the intercept and may be the slope will vary across subjects. We can modify the model to accommodate this as

$$y_{ij} = \beta_{i,1} + \beta_{i,2} \text{age}_j + e_{ij} \quad i = 1, \dots, M \quad (2)$$

Where $\beta_{i,1}$ is the intercept for the i th subject and $\beta_{i,2}$ is the slope for age for the i th subject.

We still have the problem of the dependence of the error terms plus we have a model where the number of parameters grows as the number of subjects does. Common assumptions for the error terms are normality and conditional independence, that is, within subjects, the errors are independent. We write this as $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ where $\mathbf{e}_i$ is the vector of all the error terms for the i th subject and $\mathbf{I}$ is the identity matrix. Note that the assumption of **conditional independence** is a working assumption that we know may be violated. We will address this issue later in the Section ‘Parameter estimation’.

A popular approach to reducing the number of parameters is to assume that the subjects are a random sample from some underlying population of subjects and that the parameter values for the subjects follow a distribution. A typical choice is

$$\begin{bmatrix} \beta_{i,1} \\ \beta_{i,2} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{D}) \quad (3)$$

where $\boldsymbol{\mu}^T = [\mu_1, \mu_2]$ and where μ_1 and μ_2 are the mean (fixed) intercept and slope values for the population of subjects and where $\mathbf{D}$ is a 2×2 covariance

2 Nonlinear Mixed Effects Models

matrix. This assumed distribution on the parameters makes this a mixed-effects model. Models of this type are also commonly called hierarchical, two-stage, empirical-Bayes, growth-curve (*see Growth Curve Modeling*), and multilevel-linear model (*see Linear Multilevel Models*).

Parameter Estimation

Typically, we estimate the fixed effects and variance parameter by computing the marginal distribution of the data by integrating out the random effects. This solves the problem of too many parameters. The resulting distribution has the form

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\mu}, \boldsymbol{\Sigma}_i) \quad \text{where } \boldsymbol{\Sigma}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \sigma^2 \mathbf{I} \quad (4)$$

and where $\mathbf{Z}_i$ is the design matrix for the random effects and $\mathbf{X}_i$ is the design matrix for the fixed effects (for individual i). In our example they are the same:

$$\mathbf{X}_i = \mathbf{Z}_i = \begin{bmatrix} 1 & 8.5 \\ 1 & 9.0 \\ 1 & 9.5 \\ 1 & 10.0 \\ 1 & 10.5 \end{bmatrix} \quad \text{for all } i \quad (5)$$

The fixed effects ($\boldsymbol{\mu}$) and variance components ($\mathbf{D}$ and σ) are estimated by maximizing the likelihood of the data under this marginal model (4). The marginal variance–covariance matrix $\boldsymbol{\Sigma}_i$ has a rich structure that depends on $\mathbf{D}$ and $\mathbf{Z}$ and that may be sufficient to provide accurate inference for the fixed effects even if our assumption of conditional independence within subject (the $\sigma^2 \mathbf{I}$ term in 4) is not valid. If the term $\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T$ is not sufficient to model the marginal variance–covariance structure, it may be necessary to assume a more complex within-subject correlation model. For example $\mathbf{e}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is a correlation matrix corresponding to an autoregressive model.

While this approach requires that $\mathbf{D}$ must be specified and estimated, inference about the fixed effects will not be sensitive to misspecification of $\mathbf{D}$ if the marginal distribution of $\mathbf{y}_i$ is rich enough. This should also hold for sensitivity to the assumption of conditional independence within subject. Since the marginal distribution for LME models has a closed form, the theory for finding **maximum likelihood** and restricted maximum likelihood (*see Maximum Likelihood Estimation*) is straightforward [3]. In

practice, one would use existing software packages to do the estimation.

If estimates for the random effects are desired, the best linear unbiased predictors (BLUPs) are used where ‘Best’ in this case means minimum squared error of prediction (or loss). The BLUPs are also the empirical-Bayes predictors or posterior means of the random effects (*see Random Effects in Multivariate Linear Models: Prediction*).

Nonlinear Regression

Our first task is to define the class of **nonlinear regression model**. Let us consider linear regression models first (*see Multiple Linear Regression*). The term linear regression has two common meanings. The first is straight line regression where the model has the form

$$y_j = \beta_1 + \beta_2 x_j + e_j \quad (6)$$

The subscript j indexes observations as in equation (1). We reserve the subscript i for subjects used in the next section.

The second meaning is a linear model that may, for instance, have additional polynomial terms or terms for groups, for example,

$$y_j = \beta_1 + \beta_2 \text{tmt}_j + \beta_3 x_j + \beta_4 x_j^2 + e_j \quad (7)$$

where tmt_j would be, for example, 0 for observations in the control group and 1 for observations in the treated group.

The definition of a linear model is one that can be written in the following form

$$y_j = \sum_k \beta_k x_{kj} + e_j \quad (8)$$

where the β s are the only parameters to be estimated. Equations 6 and 7 can be written in this form and are linear but the following nonlinear regression model is not linear

$$y_j = \beta_1 + \beta_2 x_j^{\beta_3} + e_j \quad (9)$$

A second definition of a linear model is a model for which the derivative of the right-hand side of the model equation with respect to any individual parameter contains no parameters at all.

Nonlinear models are most commonly used for data that show single or double asymptotes, inflection

points, and other features not well fit by polynomials. Often, there is no mechanistic interpretation for the parameters (other than intercepts, asymptotes, inflection points, etc.), they simply describe the relationship between the predictor and outcome variables via the model function. There are certain nonlinear models where scientists have attempted to interpret the parameters as having physical meaning. One such class of models is compartment models [1].

Equation 9 is a simple (fixed effects only) nonlinear regression model that can be used for data where there is no nested structure but where a linear model does not adequately describe the relationship between the predictors and the outcome variable. The general form of the model for the j th observation is

$$y_j = f(\boldsymbol{\beta}, \mathbf{x}_j) + e_j \quad e_j \sim \mathcal{N}(0, \sigma^2) \quad (10)$$

where $f(\boldsymbol{\beta}, \mathbf{x}_j) = \beta_1 + \beta_2 x_j^{\beta_3}$ is the assumed model with parameters $\boldsymbol{\beta}$ that relates the mean response y_j to the predictor variables $\mathbf{x}_j$. The error terms are assumed to be independent. Note that $\mathbf{x}_j$ can contain more than one predictor variable although in this model there is only one. The parameters $\boldsymbol{\beta}$ are typically estimated via nonlinear least squares [1].

Nonlinear Mixed-effects Models

Nonlinear mixed-effects (NLME) models are used when the data have a hierarchical structure as described above for LME models but where a nonlinear function is required to adequately describe the relationship between the predictor(s) and the outcome variable.

Example: Tongue Lengths

Figure 2 shows measured tongue length from magnetic resonance images for 14 children [6]. Note that in this data set, not every subject has been measured at each age. An important feature of mixed effects models (linear and nonlinear) is that all subjects do not have to be measured at the same time points or the same number of times.

One model that fits these data well is the asymptotic regression model. It has many formulations but the one we have chosen is

$$y_{ij} = \alpha - \beta \exp(-\gamma \text{age}_{ij}) + e_{ij} \quad (11)$$

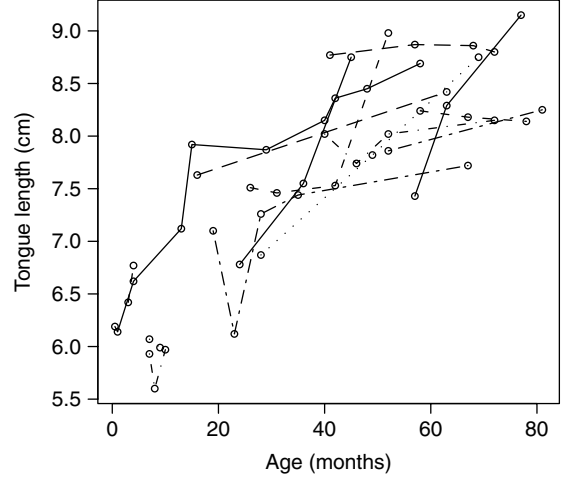


Figure 2 Tongue length at various ages for 14 children. Each child's data are connected by a line

In this version, α is the asymptote as age goes to infinity, β is the range between the response at age = 0 and α , and γ controls the rate at which the curve achieves its asymptote.

Once again it is sensible to postulate a model that includes parameters that are specific to subject. It turns out that in this case only α needs to vary between subjects. The model then takes on the following form:

$$y_{ij} = \alpha_i - \beta \exp(-\gamma \text{age}_{ij}) + e_{ij} \quad (12)$$

We model the parameters as random in the population of subjects with a fixed mean μ and assume $\alpha_i \sim \mathcal{N}(\mu, \sigma_\alpha^2)$ and $e_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

It is also possible to add additional predictors such as gender. If gender is expected to affect the intercept, we can rewrite the model for the intercepts by modifying the model for the random intercepts $\alpha_i \sim \mathcal{N}(\mu_1 + \mu_2 \text{gender}, \sigma_\alpha^2)$ or equivalently by modifying the model for the responses (but not both)

$$y_{ij} = \alpha_i + \mu_2 \text{gender} - \beta \exp(-\gamma \text{age}_{ij}) + e_{ij} \quad (13)$$

Parameter Estimation

Just as in the LME case, estimation for NLME models is accomplished by finding the maximum likelihood estimates that correspond to the marginal distribution for the data. However, unlike in the linear case, there is usually no closed form expression

for the marginal distribution of an NLME model. Thus, direct maximum likelihood estimation requires either numerical or Monte Carlo integration (*see Markov Chain Monte Carlo and Bayesian Statistics*). Alternatively, estimation can be accomplished using approximate maximum likelihood or restricted maximum likelihood [3].

We can write down an approximate marginal variance covariance–matrix which has the form

$$\hat{\Sigma}_i = \hat{\mathbf{Z}}_i \hat{\mathbf{D}} \hat{\mathbf{Z}}_i^T + \hat{\sigma}^2 \mathbf{I}. \quad (14)$$

Here $\hat{\mathbf{Z}}_i$ is the derivative matrix of the right-hand side of the subject specific model (12) with respect to the random effects. Unlike in the linear model, it will, in general, depend on the estimated parameter values [3]. The term $\hat{\mathbf{Z}}_i \hat{\mathbf{D}} \hat{\mathbf{Z}}_i^T$ may capture the marginal covariance of the data and assuming independence within subject may still result in a valid inference on the fixed effects. If not (as in this case where only the ‘intercept’ is random and $\mathbf{Z}_i$ is a column vector of 1s), various more general forms for the within-subject covariance can be assumed (as described under LME models above).

The maximum likelihood estimates (and standard errors) for the tongue data main effects are $\hat{\mu} = 12.16(0.65)$, $\hat{\beta} = 4.65(0.57)$, and $\hat{\gamma} = 0.023(0.0069)$. The estimated variance components are $\hat{\sigma}_2 = 0.203$ (standard deviation of the intercepts) and $\hat{\sigma} = 0.478$ (within-subject standard deviation). Figure 3 shows the fitted population average and subject specific curves for the tongue length data. The thick line corresponds to the fitted values when the estimated mean value $\hat{\mu}$ is substituted in for α_i . The lighter lines each correspond to a subject and are computed from the BLUP estimates of α_i . These estimates were obtained using the NLME package for the R statistical software package [4]. There are other packages that implement NLME models, including the SAS procedure NLMIXED (*see Software for Statistical Analyses*).

Testing in NLME models is a bit more complex than in standard regression or **analysis of variance**. Simulation studies have found that likelihood ratio tests perform well for testing for the need for variance components (like adding a random effect) even though they are theoretically invalid because the parameter value being tested is on the edge of the parameter space. Methods for testing fixed

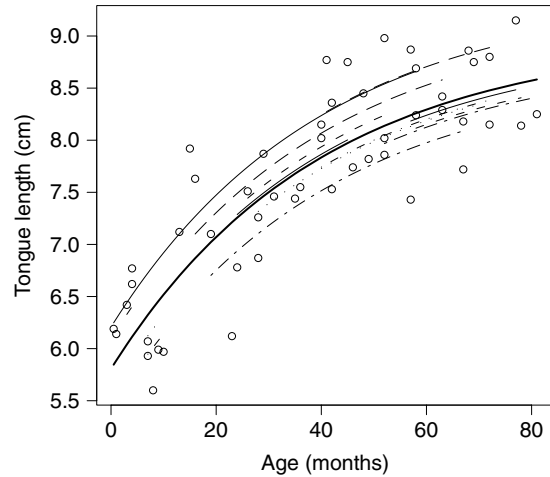


Figure 3 Population average (thick line) and subject level (lighter lines) fitted curves for the tongue length data

effects are still controversial but it is known that LR tests are much too liberal and approximate F tests [3] seem to do a better job. More evaluation is required.

References

- [1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, New York.
- [2] Laird, N.M. & Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4), 963–974.
- [3] Pinheiro, J.C. & Bates, D.M. (2000). *Mixed Effects Models in S and S-PLUS*, Springer, New York.
- [4] R Development Core Team. (2004). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.
- [5] Tuddenham, R.D. & Snyder, M.M. (1954). Physical growth of California boys and girls from birth to eighteen years, *University of California Publications in Child Development* **1**, 183–364.
- [6] Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R. & Yandell, B.S. (2005). Development of vocal tract length during early childhood: a magnetic resonance imaging study, *Journal of the Acoustical Society of America* **117**, 338–350.

(See also **Generalized Linear Mixed Models; Marginal Models for Clustered Data**)

MARY J. LINDSTROM AND
HOURI K. VORPERIAN

Nonlinear Models

L. JANE GOLDSMITH AND VANCE W. BERGER

Volume 3, pp. 1416–1419

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonlinear Models

A nonlinear model expresses a relationship between a response variable and one or more predictor variables as a nonlinear function. The stochastic, or random, model is used to indicate that the expected response is obscured somewhat by a random error element. The nonlinear model may be expressed as follows:

$$Y = f(X, \theta) + \varepsilon, \quad (1)$$

where X denotes a vector (perhaps of length one) of independent, or predictor, variables and θ denotes a vector of one or more (unknown) parameters in the model. $f(X, \theta)$ expresses the functional relationship between the expectation of Y , denoted $E(Y)$, and X and θ . That is, $E(Y) = f(X, \theta)$. ε denotes the random error that prevents measurement of $E(Y)$ directly. ε is assumed to have a Gaussian distribution centered upon 0 and with constant variance for each independent observation of Y and X . That is, $\varepsilon \sim N(0, \sigma^2)$.

An example of a nonlinear model is

$$Y = \frac{\theta_1 X}{\theta_2 + X} + \varepsilon, \quad (2)$$

the so-called Michaelis–Menten model. This model can be transformed to a linear form for $E(Y)$, but the error properties of ε usually dictate the use of nonlinear estimation techniques. The Michaelis–Menten model has been used to model the velocity of a chemical reaction as a function of the underlying concentrations. Thus, a mnemonic representation is given by

$$V = \frac{\theta_1 C}{\theta_2 + C} + \varepsilon. \quad (3)$$

A graph of this model is presented in Figure 1:

Here, we see the nonlinear relationship expected between V and C , with $\theta_1 = 30$ and $\theta_2 = 0.5$.

Advantages of Nonlinear Models

An obvious advantage of a nonlinear model is that the methodology allows modeling and prediction of relationships that are more complicated in nature than the familiar linear model. Nonlinear

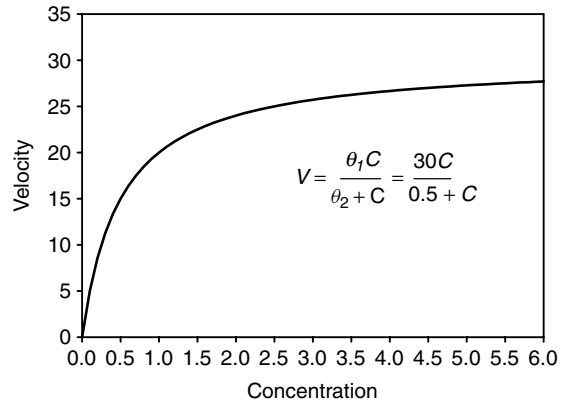


Figure 1 A Michaelis–Menten model

modeling also allows for special interpretation of parameters. In some cases, the parameters or functions of them describe special attributes of the nonlinear relationship that are apparent in graphs, such as (a) intercepts, (b) asymptotes, (c) maxima, (d) minima, and (e) inflection points. The following graph (Figure 2) demonstrates a graphical interpretation of $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ in the Michaelis–Menten model:

This demonstrates that θ_1 is the maximum attainable value, expressed as an asymptote, for V , and θ_2 is the so-called ‘half-dose,’ or the concentration C at which V attains half its theoretical maximum. Another nonlinear model is graphed in Figure 3:

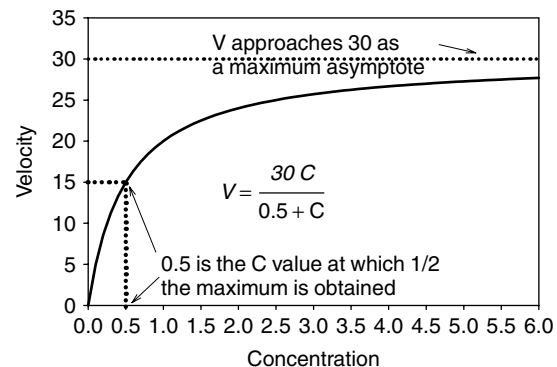


Figure 2 Parameter interpretation in a Michaelis–Menten model

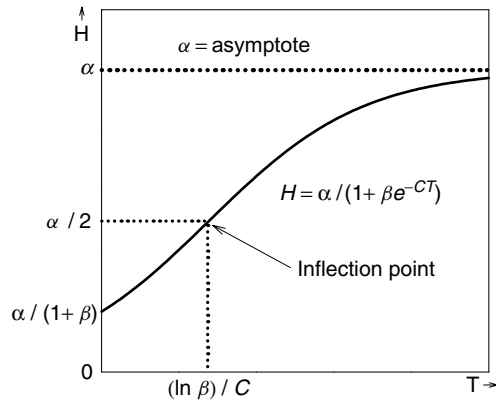


Figure 3 The autocatalytic growth model

The ‘autocatalytic growth model’ postulates that height H cannot exceed its maximum α , that it occurs as an S-shaped curve with inflection point $((\ln \beta)/C, \alpha/2)$, and that the starting height at $T = 0$ is $\alpha/(1 + \beta)$. The formula $H = \alpha/(1 + \beta e^{-CT})$, with its logistic, S-shaped curve, assures that the curve is skew-symmetric about the inflection point.

As we can see, a researcher familiar with data properties through graphs may be able to deduce or find a nonlinear model that reflects desired properties. Another advantage of nonlinear models is that the model can reflect properties based on relevant differential equations. For instance, the autocatalytic growth model is the solution to a specific differential equation. This rate equation states that the growth rate relative to the attained height is proportional to the amount of height yet unattained. In other words, as height approaches its maximum, the growth slows. The differential equation for which $H = \alpha/(1 + \beta e^{-CT})$ is a solution is

$$\frac{dH}{dT} \frac{1}{H} = \frac{C(\alpha - H)}{\alpha}. \quad (4)$$

Often nonlinear models arise when researchers can specify a rate-of-change relationship, which can be expressed as either a single differential equation or as a system of differential equations. Choosing a candidate model can be a creative, enjoyable process in which researchers use their knowledge and experience to find the most promising model for their project.

Parameter Estimation

To estimate the parameters in a proposed nonlinear model $Y = f(X, \theta) + \varepsilon$, data observations must be made on both the response variable Y and its predictors X . The number of observations made is denoted by n , or the ‘sample size.’ An iterative process is often used to find the parameter estimates $\hat{\theta}$. To find the ‘best’ estimates, or the ones likely to be close to the true, unknown θ , the iterative method tries to find the $\hat{\theta}$ that minimizes the sum of squared differences between the n observed Y ’s and the predicted Y ’s, the $\hat{Y}$ ’s. That is, the least-squares solution (see **Least Squares Estimation**) will be the $\hat{\theta}$ that minimizes the sum of squared residuals

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \quad (5)$$

In other words, the nY_i ’s obtained using the least-squares optimal $\hat{\theta}$ will minimize the sum of squared residuals. There is a vast literature dealing with methods of optimization using computer methods (see **Optimization Methods**). Some techniques developed to minimize $\sum_{i=1}^n r_i^2$ are the steepest-descent or gradient method, the Newton method, the modified Gauss–Newton method, the Marquardt method, and the multivariate secant method. Most of these methods require an initial starting value for $\hat{\theta}$. This can often be obtained through graphical analysis of the data, using any known relationships between a θ component θ_i and graphical properties such as asymptotes, intercepts, and so on.

If there is a linearizing transformation that is not being used due to error distribution considerations (as was the case for the Michaelis–Menten model above), then the linear least-squares solution for the transformed linear model may provide a good starting $\hat{\theta}$. Another possibility might be to use an initial $\hat{\theta}$ from the literature if the model has been used successfully in a previous study.

Since there is not a closed-form solution, a criterion must be chosen to determine when a satisfactory solution has been found and the computer minimization program should stop iterating. Usually, the stopping rule is to abandon searching for a better solution and report the current $\hat{\theta}$ when the reduction of $\sum_{i=1}^n r_i^2$ is small in the most recent step. Sometimes there is difficulty obtaining convergence, with the computer program ‘cycling’ around a potential solution, but not homing in on the best answer due

to technical problems with the method chosen or a poorly chosen starting value. Trying again with a new starting $\hat{\theta}$ or choosing a new iterative optimization technique are obvious suggestions when convergence difficulties occur. This approach will also help ensure that if convergence is reached, then what was found was a true global minimum, and not just a local one.

Confidence Regions

In addition to the point estimate $\hat{\theta}$, which is generally the least-squares solution, it is also desirable to report 95% (or some such high confidence level) confidence regions for θ . Researchers often desire confidence intervals for individual θ_i , the components of θ . Through advanced programming techniques, it is possible to compute exact confidence regions for θ . However, if the sample size is small, then it may be extremely difficult to compute the regions, or the regions themselves may be unsatisfactorily complicated. For example, a 95% exact confidence region for θ may consist of several disjoint, asymmetric areas in R^k , where k is the dimension of θ . Approximate confidence ellipsoids for θ can be computed, along with approximate confidence intervals for the θ_i . The accuracy and precision of these approximate confidence regions depend upon two factors: the estimation properties of the chosen model and the sample size.

Estimation Properties

The accuracy (unbiasedness) and precision (small size) of the approximate confidence ellipsoids and confidence intervals mentioned above depend upon an estimation property of a particular nonlinear model termed ‘close-to-linear behavior.’ As it turns out, a model with close-to-linear behavior will tend to produce good confidence regions and confidence intervals with relatively small sample sizes. Conversely, those models without this behavior may not produce similarly desirable confidence intervals. The properties of various types of nonlinear models have been rigorously studied, and researchers seeking an efficient nonlinear model, which can accomplish accurate inference with small sample sizes, can search the literature of nonlinear models to find a model with desirable estimation properties.

Model Validation

The nonlinear model validation process is similar in some ways to that for linear regression (*see Multiple Linear Regression*). Scatter plots of residuals versus predicted values can be used to detect model fit. Small confidence intervals for the θ_i indicate precision. An important *caveat* is that R^2 , the familiar coefficient of determination from linear regression, is not meaningful in the context of a nonlinear model.

A Fitted Nonlinear Model

To illustrate these points, simulated data, with a sample size of $n = 32$, were generated using the Michaelis–Menten model discussed above:

$$V = \frac{\theta_1 C}{\theta_2 + C} + \varepsilon, \text{ with } \theta_1 = 30, \text{ and } \theta_2 = 0.5. \\ \varepsilon \sim N(0, (1.5)^2) \quad (6)$$

The nonlinear regression program (SAS version 8.02) converged in five iterations to the following estimates:

Parameter	Estimate	Estimated S.E.	Approximate 95% confidence limits	
θ_1	29.36	0.66	28.02	30.70
θ_2	0.44	0.05	0.34	0.54

A graph is presented in Figure 4.

There are some excellent texts dealing with nonlinear models. See, for example, [1], [2], [3], [4], [5], and [6].

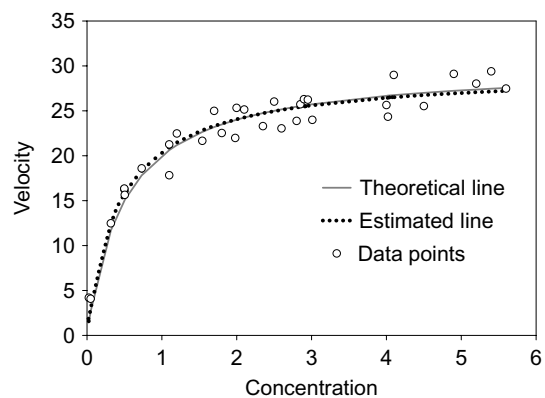


Figure 4 A Fitted Michaelis-Menton Model

4 Nonlinear Models

References

- [1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, New York.
- [2] Dixon, W.J. chief editor. (1992). *BMDP Statistical Software Manual*, Vol. 2, University of California Press, Berkeley.
- [3] Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition, Wiley, New York.
- [4] Ratkowsky, D.A. (1990). *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York.
- [5] SAS (Statistical Analysis System). *Online Documentation, Release 8.02*, (1999). SAS Institute, Cary.
- [6] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*, Wiley, New York.

L. JANE GOLDSMITH AND VANCE W. BERGER

Nonparametric Correlation (r_s)

DAVID C. HOWELL

Volume 3, pp. 1419–1420

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonparametric Correlation (r_s)

Spearman's Correlation for Ranked Data

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$ be random samples of sizes n from the two random variables, X and Y . The scale of measurement of the two random variables is at least ordinal and, to avoid problems with ties, ought to be strictly continuous. Rank the data within each variable separately and let $(r_{x1}, r_{x2}, \dots, r_{xn})$ and $(r_{y1}, r_{y2}, \dots, r_{yn})$ represent the ranked data.

Our goal is to correlate the two sets of ranks and obtain a test of significance on that correlation. Spearman provided a formula for this correlation, assuming no ties in the ranks, as

$$r_s = 1 - \frac{6\sum D_i^2}{N(N^2 - 1)}, \quad (1)$$

where D_i is defined as the set of differences between r_{xi} and r_{yi} . Spearman derived this formula (using such equalities as the sum of the first N integers = $N(N + 1)/2$) at a time when calculations were routinely carried by hand. The result of using this formula is exactly the same as applying the traditional **Pearson product-moment correlation** coefficient to the ranked data, and that is the approach commonly used today. The Pearson formula is correct regardless of whether there are tied ranks, whereas the Spearman formula requires a correction for ties.

Example

The Data and Story Library website (DASL) (<http://lib.stat.cmu.edu/DASL/Stories/AlcoholandTobacco.html>) provides data on the average weekly spending on alcohol and tobacco products for each of 11 regions in Great Britain. The data follow in Table 1. Columns 2 and 3 contain expenditures in pounds, and Columns 4 and 5 contain the ranked data for their respective expenditures.

Though it is not apparent from looking at either the alcohol or tobacco variable alone, in a bivariate plot it is clear that Northern Ireland is a major outlier.

Table 1 Expenditures for alcohol and tobacco by region of Great Britain

Region	Alcohol	Tobacco	Rank A	Rank T
North	6.47	4.03	11	9
Yorkshire	6.13	3.76	9	7
Northeast	6.19	3.77	10	8
East Midlands	4.89	3.34	4	4
West Midlands	5.63	3.47	6	5
East Anglia	4.52	2.92	2	2
Southeast	5.89	3.20	7	3
Southwest	4.79	2.71	3	1
Wales	5.27	3.53	5	6
Scotland	6.08	4.51	8	10
Northern Ireland	4.02	4.56	1	11

Similarly, the distribution of alcohol expenditures is decidedly nonnormal, whereas the ranked data on alcohol, like all ranks, is rectangularly distributed.

The Spearman correlation coefficient (r_s) for this sample is 0.373, whereas the Pearson correlation (r) on the raw-score variables is 0.224.

There is no generally accepted method for estimation of the standard error of r_s for small samples. This makes the computation of confidence limits problematic, but with very small samples the width of the confidence interval is likely to be very large in any event. Kendall [1] has suggested that for sample sizes greater than 10, r_s can be tested in the same manner as the Pearson correlation (*see* **Pearson Product Moment Correlation**). For these data, because of the extreme outlier for Northern Ireland, the correlation is not significant using the standard approach for Pearson's r ($p = 0.259$, two tailed) or using a randomization test ($p = 0.252$, two tailed).

There are alternative approaches to rank correlation. See **Nonparametric Correlation (tau)** for an alternative method for correlating ranked data.

Reference

- [1] Kendall, M.G. (1948). *Rank Correlation Methods*, Griffen Publishing, London.

(See also **Kendall's Tau – τ**)

DAVID C. HOWELL

Nonparametric Correlation (τ)

DAVID C. HOWELL

Volume 3, pp. 1420–1421

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonparametric Correlation (tau)

Kendall's Correlation for Ranked Data (τ)

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_n)$ be random samples of sizes n from the two random variables, X and Y . The scale of measurement of the two random variables is at least ordinal and, to avoid problems with ties, ought to be strictly continuous. Rank the data within each variable separately and let $(r_{x1}, r_{x2}, \dots, r_{xn})$ and $(r_{y1}, r_{y2}, \dots, r_{yn})$ represent the ranked data.

Our goal is to correlate the two sets of ranks and obtain a test of significance on that correlation. The correlation coefficient is based on the number of concordant and discordant rankings, or, equivalently, the number of inversions in rank.

We will rank the data separately for each variable and let C represent the number of concordant pairs of observations, which is the number of pairs of observations whose rankings on the two variables are ordered in the same direction. We will let D represent the number of discordant pairs (pairs where the rankings on the two variables are in the opposite direction), and let $S = C - D$. Then

$$\tau = \frac{2S}{n(n-1)}. \quad (1)$$

(It is not essential that we actually do the ranking, though it is easier to work with ranks than with raw scores.)

Example

The Data and Story Library website (DASL) (<http://lib.stat.cmu.edu/DASL/Stories/AlcoholandTobacco.html>) provides data on the average weekly spending on alcohol and tobacco products for each of 11 regions in Great Britain. The data follow. Columns 2 and 3 contain expenditures in pounds, and Columns 4 and 5 contain the ranked data for their respective expenditures as in Table 1.

Though it is not apparent from looking at either the Alcohol or Tobacco variable alone, in a bivariate plot it is clear that Northern Ireland is a major outlier.

Table 1 Alcohol and tobacco expenditures in Great Britain, broken down by region

Region	Alcohol	Tobacco	Rank A	Rank T
North	6.47	4.03	11	9
Yorkshire	6.13	3.76	9	7
Northeast	6.19	3.77	10	8
East Midlands	4.89	3.34	4	4
West Midlands	5.63	3.47	6	5
East Anglia	4.52	2.92	2	2
Southeast	5.89	3.20	7	3
Southwest	4.79	2.71	3	1
Wales	5.27	3.53	5	6
Scotland	6.08	4.51	8	10
Northern Ireland	4.02	4.56	1	11

Similarly, the distribution of Alcohol expenditures is decidedly nonnormal, whereas the ranked data on alcohol, like all ranks, are rectangularly distributed.

There are $n(n-1)/2 = 11(10)/2 = 55$ pairs of rankings. 37 of those pairs are concordant, while 18 are discordant. (For example, Yorkshire ranks lower than the North of England on both Alcohol and Tobacco, so that pair of rankings is concordant. On the other hand, Scotland ranks lower than the North of England on Alcohol expenditures, but higher on Tobacco expenditures, so that pair of rankings is discordant.)

Then

$$C = 37$$

$$D = 18$$

$$S = C - D = 37 - 18 = 19$$

$$\tau = \frac{2S}{N(N-1)} = \frac{38}{110} = 0.345. \quad (2)$$

For these data Kendall's $\tau = 0.345$.

Unlike Spearman's r_s , there is an accepted method for estimation of the standard error of Kendall's τ [1].

$$s_\tau = \sqrt{\frac{2(2N+5)}{9N(N-1)}}. \quad (3)$$

Moreover, τ is approximately normally distributed for $N \geq 10$. This allows us to approximate the sampling distribution of Kendall's τ using the normal approximation.

$$z = \frac{\tau}{s_\tau} = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} = \frac{0.345}{\sqrt{\frac{2(27)}{9(11)(10)}}}$$

2 Nonparametric Correlation (tau)

$$= \frac{0.345}{0.2335} = 1.48. \quad (4)$$

For a two-tailed test $p = .139$, which is not significant.

With a standard error of 0.2335, the **confidence limits** on Kendall's τ , assuming normality of τ , would be

$$\begin{aligned} CI &= \tau \pm 1.96s_\tau = \tau \pm 1.96 \left(\sqrt{\frac{2(2N+5)}{9N(N-1)}} \right) \\ &= \tau \pm 1.96(0.2335) \end{aligned} \quad (5)$$

For our example this would produce confidence limits of $-0.11 \leq t \leq 0.80$.

There are alternative approaches to rank correlation. See **Nonparametric Correlation (r_s)** for an alternative method for correlating ranked data. Kendall's τ has generally been given preference over Spearman's r_s because it is a better estimate of the corresponding population parameter, and its standard error is known.

Reference

- [1] Kendall, M.G. (1948). *Rank Correlation Methods*, Griffin Publishing, London.

DAVID C. HOWELL

Nonparametric Item Response Theory Models

KLAAS SIJTSMA

Volume 3, pp. 1421–1426

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonparametric Item Response Theory Models

Goals of Nonparametric Item Response Theory

Nonparametric item response theory (NIRT) models are used for analyzing the data collected by means of tests and questionnaires consisting of J items. The goals are to construct a scale for the ordering of persons and – depending on the application of this scale – for the items. To attain these goals, test constructors and other researchers primarily want to know whether their items measure the same or different traits, and whether the items are of sufficient quality to distinguish people with relatively low and high standings on these traits. These questions relate to the classical issues of validity (*see* **Validity Theory and Applications**) and reliability, respectively. Other issues of interest are **differential item functioning**, person-fit analysis, and skill identification and cognitive modeling.

NIRT models are most often used in small-scale testing applications. Typical examples are intelligence and personality testing. Most intelligence tests and personality inventories are applied to individuals only once, each individual is administered the same test, and testing often but not necessarily is individual. Another example is the measurement of attitudes, typical of sociological or political science research. Attitude questionnaires typically consist of, say, 5 to 15 items, and the same questionnaire is administered to each respondent in the sample. NIRT is also applied in educational testing, preference measurement in marketing, and health-related quality-of-life measurement in a medical context. See [16] for a list of applications.

NIRT is interesting for at least two reasons. First, because it provides a less demanding framework for test and questionnaire data analysis than parametric item response theory, NIRT is more data-oriented, more exploratory and thus more flexible than parametric IRT; see [7]. Second, because it is based on weaker assumptions than parametric IRT, NIRT can be used as a framework for the theoretical exploration of the possibilities and the boundaries of IRT in general; see, for example, [4] and [6].

Assumptions of Nonparametric IRT Models

The assumptions typical of NIRT models, and often shared with parametric IRT models, are the following:

- *Unidimensionality (UD)*. A unidimensional IRT model contains one **latent variable**, usually denoted by θ , that explains the variation between tested individuals. From a fitting unidimensional IRT model, it is inferred that performance on the test or questionnaire is driven by one ability, achievement, personality trait, or attitude. Multidimensional IRT models exist that assume several latent variables to account for the data.
- *Local independence (LI)*. Let X_j be the random variable for the score on item j ($j = 1, \dots, J$); let x_j be a realization of X_j ; and let $\mathbf{X}$ and $\mathbf{x}$ be the vectors containing J item score variables and J realizations, respectively. Also, let $P(X_j = x_j|\theta)$ be the conditional probability of a score of x_j on item j . Then, a latent variable θ , possibly multidimensional, exists such that the joint conditional probability of J item responses can be written as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{j=1}^J P(X_j = x_j|\theta). \quad (1)$$

An implication of LI is that for any pair of items, say j and k , their conditional covariance equals 0; that is, $\text{Cov}(X_j, X_k|\theta) = 0$.

- *Monotonicity (M)*. For binary item scores, $X_j \in \{0, 1\}$, with score zero for an incorrect answer and score one for a correct answer, we define $P_j(\theta) \equiv P(X_j = 1|\theta)$. This is the item response function (IRF). Assumption M says that the IRF is monotone nondecreasing in θ . For ordered rating scale scores, $X_j \in \{0, \dots, m\}$, a similar monotonicity assumption can be made with respect to response probability, $P(X_j \geq x_j|\theta)$.

Parametric IRT models typically restrict the IRF, $P_j(\theta)$, by means of a parametric function, such as the logistic. A well-known example is the three-parameter logistic IRF. Let γ_j denote the lower asymptote of the logistic IRF for item j , interpreted as the pseudochance probability; let δ_j denote the location of item j on the θ scale, interpreted as the

2 Nonparametric Item Response Theory Models

difficulty; and let $\alpha_j(>0)$ correspond to the steepest slope of the logistic function, which is located at parameter δ_j and interpreted as the discrimination. Then the IRF of the three-parameter logistic model is

$$P_j(\theta) = \gamma_j + (1 - \gamma_j) \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (2)$$

Many other parametric IRT models have been proposed; see [20] for an overview.

NIRT models only impose order restrictions on the IRF, but refrain from a parametric definition. Thus, assumption M may be the only restriction, so that for any two fixed values $\theta_a < \theta_b$, we have that

$$P_j(\theta_a) \leq P_j(\theta_b). \quad (3)$$

The NIRT model based on assumptions UD, LI, and M is the monotone homogeneity model [8]. Assumption M may be further relaxed by assuming that the mean of the J IRFs is increasing, but not each of the individual IRFs [17]. This mean is the test response function, denoted by $T(\theta)$ and defined as

$$T(\theta) = J^{-1} \sum_{j=1}^J P_j(\theta), \text{ increasing in } \theta. \quad (4)$$

Another relaxation of assumptions is that of strict unidimensionality, defined as assumption UD, to essential unidimensionality [17]. Here, the idea is that one dominant trait drives test performance in particular, but that there are also nuisance traits active, whose influence is minor.

In general, one could say that NIRT strives for defining models that are based on relatively weak assumptions while maintaining desirable measurement properties. For example, it has been shown [2] that the assumptions of UD, LI, and M imply that the total score $X_+ = \sum_{j=1}^J X_j$ stochastically orders latent variable θ ; that is, for two values of X_+ , say $x_{+a} < x_{+b}$, and any value t of θ , assumptions UD, LI, and M imply that

$$P(\theta \geq t | X_+ = x_{+a}) \leq P(\theta \geq t | X_+ = x_{+b}). \quad (5)$$

Reference [3] calls (2) a stochastic ordering in the latent trait (SOL). SOL implies that for higher X_+ values the θ s are expected to be higher on average. Thus, (5) defines an ordinal scale for person measurement: if the monotone homogeneity model fits the data, total score X_+ can be used for ordering

persons with respect to latent variable θ , which by itself is not estimated.

The three-parameter logistic model is a special case of the monotone homogeneity model, because it has monotone increasing logistic IRFs and assumes UD and LI. Thus, SOL also holds for this model. The item parameters, γ , δ , and α , and the latent variable θ can be solved from the likelihood of this model. These estimates can be used to calibrate a metric scale that is convenient for equating, item banking, and adaptive testing in large-scale testing [20]. NIRT models are candidates for test construction, in particular, when an ordinal scale for respondents is sufficient for the application envisaged.

Another class of NIRT models is based on stronger assumptions. For example, to have an ordering of items which is the same for all values of θ , with the possible exception of ties for some θ s, it is necessary to assume that the J items have IRFs that do not intersect. This is called an invariant item ordering (IIO, [15]). Formally, J items have an IIO, when they can be ordered and numbered such that

$$P_1(\theta) \leq P_2(\theta) \leq \dots \leq P_J(\theta), \text{ for all } \theta. \quad (6)$$

A set of items that is characterized by an IIO facilitates the interpretation of results from differential item functioning and person-fit analysis, and provides the underpinnings of the use of individual starting and stopping rules in intelligence testing and the hypothesis testing of item orderings that reflect, for example, ordered developmental stages. The NIRT model based on the assumptions of UD, LI, M, and IIO is the double monotonicity model [8].

The generalization of the SOL and IIO properties from dichotomous-item IRT models to polytomous-item IRT models is not straightforward. Within the class of known polytomous IRT models, SOL can only be generalized to the parametric partial credit model [20] but not to any other model. Reference [19] demonstrated that although SOL is not implied by most models, it is a robust property for most tests in most populations, as simulated in a robustness study. For J polytomously scored items, an IIO is defined as

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \dots \leq E(X_J|\theta), \text{ for all } \theta. \quad (7)$$

Thus, the ordering of the mean item scores is the same, except for possible ties, for each value of

θ . IIO can only be generalized to the parametric rating scale model [20] and similarly restrictive IRT models. See [13] and [14] for nonparametric models that imply an IIO.

Because SOL and IIO are not straightforwardly generalized to polytomous-item IRT models, and because these models are relatively complicated, we restrict further attention mostly to dichotomous-item IRT models. More work on the foundations of IRT through studying NIRT has been done, for example, by [1, 3, 4, 6, and 17]. See [8] and [16] for monographs on NIRT.

Evaluating Model-data Fit

Several methods exist for investigating fit of NIRT models to test and questionnaire data. These methods are based mostly on one of two properties of observable variables implied by the NIRT models.

Conditional-association Based Methods

The first observable property is conditional association [4]. Split item score vector $\mathbf{X}$ into two disjoint part vectors, $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Define f_1 and f_2 to be non-decreasing functions in the item scores from $\mathbf{Y}$, and g to be some function of the item scores in $\mathbf{Z}$. Then UD, LI, and M imply conditional association in terms of the covariance, denoted by Cov, as

$$\text{Cov}[f_1(\mathbf{Y}), f_2(\mathbf{Y})|g(\mathbf{Z}) = z] \geq 0, \quad \text{for all } z. \quad (8)$$

Two special cases of (8) constitute the basis of model-data fit methods:

Unconditional inter-item covariances. If function $g(\mathbf{Z})$ selects the whole group, and $f_1(\mathbf{Y}) = X_j$ and $f_2(\mathbf{Y}) = X_k$, then a special case of (8) is

$$\text{Cov}(X_j, X_k) \geq 0, \quad \text{all pairs } j, k; j < k. \quad (9)$$

Negative inter-item covariances give evidence of model-data misfit.

Let $\text{Cov}(X_j, X_k)_{\max}$ be the maximum **covariance** possible given the marginals of the cross table for the bivariate frequencies on these items. Reference [8] defined coefficient H_{jk} as

$$H_{jk} = \frac{\text{Cov}(X_j, X_k)}{\text{Cov}(X_j, X_k)_{\max}}. \quad (10)$$

Equation 9 implies that $0 \leq H_{jk} \leq 1$. Thus, positive values of H_{jk} found in real data support the monotone homogeneity model, while negative values reject the model. Coefficient H_{jk} has been generalized to (1) an item coefficient, H_j , which expresses the degree to which item j belongs with the other $J - 1$ in one scale; and (2) a scalability coefficient, H , which expresses the degree to which persons can be reliably ordered on the θ scale using total score X_+ .

An item selection algorithm has been proposed [8, 16] and implemented in the computer program MSP5 [9], which selects items from a larger set into clusters that contain items having relatively high H_j values with respect to one another – say, $H_j \geq c$, often with $c > 0.3$ (c user-specified) – while unselected items have H_j values smaller than c . Because, for a set of J items, $H \geq \min(H_j)$ [8], item selection produces scales for which $H \geq c$. If $c \geq 0.3$, person ordering is at least weakly reliable [16]. Such scales can be used in practice for person measurement, while each scale identifies another latent variable.

Conditional inter-item covariances. First, define a total score – here, called a *rest score* and denoted R – based on $\mathbf{X}$ as,

$$R_{(-j, -k)} = \sum_{h \neq j, k} X_h. \quad (11)$$

Second, define function $g(\mathbf{Z}) = R_{(-j, -k)}$, and let $f_1(\mathbf{Y}) = X_j$ and $f_2(\mathbf{Y}) = X_k$. Equation 8 implies that,

$$\begin{aligned} \text{Cov}[X_j, X_k | R_{(-j, -k)} = r] &\geq 0, \quad \text{all } j, k; j < k; \\ \text{all } r &= 0, 1, \dots, J - 2. \end{aligned} \quad (12)$$

That is, in the subgroup of respondents that have the same rest score r , the covariance between items j and k must be nonnegative. Equation 12 is the basis of procedures that try to find an item subset structure for the whole test that approximates local independence as good as possible. The optimal solution best represents the latent variable structure of the test data. See the computer programs DETECT and HCA/CCPROX [18] for exploratory item selection, and DIMTEST [17] for confirmatory hypothesis testing with respect to test composition.

Manifest-monotonicity Based Methods

The second observable property is manifest monotonicity [6]. It can be used to investigate assumption

4 Nonparametric Item Response Theory Models

M. To estimate the IRF for item j , $P_j(\theta)$, first a sum score on $J - 1$ items excluding item j ,

$$R_{(-j)} = \sum_{k \neq j} X_k, \quad (13)$$

is used as an estimate of θ , and then the conditional probability $P[X_j = 1 | R_{(-j)} = r]$ is calculated for all values r of $R_{(-j)}$. Given the assumptions of UD, LI, and M, the conditional probability $P[X_j = 1 | R_{(-j)}]$ must be nondecreasing in $R_{(-j)}$; this is manifest monotonicity.

Investigating assumption M. The computer program MSP5 can be used for estimating probabilities, $P[X_j = 1 | R_{(-j)}]$, plotting the discrete response functions for $R_{(-j)} = 0, \dots, J - 1$, and testing violations of manifest monotonicity for significance. The program TestGraf98 [11, 12] estimates continuous response functions using kernel smoothing, and provides many graphics. These response functions include, for example, the option response functions for each of the response options of a multiple-choice item.

Investigating assumption IIO. To investigate whether the items j and k have intersecting IRFs, the conditional probabilities $P[X_j = 1 | R_{(-j, -k)}]$ and $P[X_k = 1 | R_{(-j, -k)}]$ can be compared for each value $R_{(-j, -k)} = r$, and the sign of the difference can be compared with the sign of the difference of the sample item means, $\bar{X}_j$ and $\bar{X}_k$, for the whole group. Opposite signs for some r values indicate intersection of the IRFs and are tested against the null hypothesis that $P[X_j = 1 | R_{(-j, -k)} = r] = P[X_k = 1 | R_{(-j, -k)} = r]$ – meaning that the IRFs coincide locally – in the population. This method and other methods for investigating an IIO have been discussed and compared by [15] and [16]. MSP5 [9] can be used for investigating IIO.

Many of the methods mentioned have been generalized to polytomous items, but research in this area is still going on. Finally, we mention that methods for estimating the reliability of total score X_+ have been developed under the assumptions of UD, LI, M, and IIO, both for dichotomous and polytomous items [16].

Developments, Alternative Models

NIRT developed later than parametric IRT. It is an expanding area, both theoretically and practically.

New developments are in adaptive testing, differential item functioning, person-fit analysis, and cognitive modeling. The analysis of the dimensionality of test and questionnaire data has received much attention, using procedures implemented in the programs DETECT, HCA/CCPROX, DIMTEST, and MSP5. Latent class analysis has been used to formulate NIRT models as discrete, ordered latent class models and to define fit statistics for these models. Modern estimation methods such as **Markov Chain Monte Carlo** have been used to estimate and fit NIRT models. Many other developments are ongoing.

The theory discussed so far was developed for analyzing data generated by means of monotone IRFs, that is, data that conform to the assumption that a higher θ value corresponds with a higher expected item score, both for dichotomous and polytomous items. Some item response data reflect a choice process governed by personal preferences for some but not all items or stimuli, and assumption M is not adequate. For example, a marketing researcher may present subjects with J brands of beer, and ask them to pick any number of brands that they prefer in terms of bitterness; or a political scientist may present a sample of voters with candidates for the presidency and ask them to order them with respect to perceived trustworthiness. The data resulting from such tasks require IRT models with single-peaked IRFs. The maximum of such an IRF identifies the item location or an interval on the scale – degree of bitterness or trustworthiness – at which the maximum probability of picking that stimulus is obtained. See [10] and [5] for the theoretical foundation of NIRT models for single-peaked IRFs and methods for investigating model-data fit.

References

- [1] Ellis, J.L. & Van den Wollenberg, A.L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model, *Psychometrika* **58**, 417–429.
- [2] Grayson, D.A. (1988). Two-group classification in latent trait theory: scores with monotone likelihood ratio, *Psychometrika* **53**, 383–392.
- [3] Hemker, B.T., Sijtsma, K., Molenaar, I.W. & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models, *Psychometrika* **62**, 331–347.
- [4] Holland, P.W. & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models, *The Annals of Statistics* **14**, 1523–1543.

-
- [5] Johnson, M.S. & Junker, B.W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models, *Journal of Educational and Behavioral Statistics* **28**, 195–230.
 - [6] Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models, *The Annals of Statistics* **21**, 1359–1378.
 - [7] Junker, B.W. & Sijtsma, K. (2001). Nonparametric item response theory in action: an overview of the special issue, *Applied Psychological Measurement* **25**, 211–220.
 - [8] Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis*, De Gruyter, Berlin.
 - [9] Molenaar, I.W. & Sijtsma, K. (2000). *MSP5 for Windows. User's Manual*, iecProGAMMA, Groningen.
 - [10] Post, W.J. (1992). *Nonparametric Unfolding Models: A Latent Structure Approach*, DSWO Press, Leiden.
 - [11] Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika* **56**, 611–630.
 - [12] Ramsay, J.O. (2000). *A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data*, Department of Psychology, McGill University, Montreal.
 - [13] Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP), *Psychometrika* **60**, 281–304.
 - [14] Sijtsma, K. & Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models, *Psychometrika* **63**, 183–200.
 - [15] Sijtsma, K. & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering, *British Journal of Mathematical and Statistical Psychology* **49**, 79–105.
 - [16] Sijtsma, K. & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*, Sage Publications, Thousand Oaks.
 - [17] Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* **55**, 293–325.
 - [18] Stout, W.F., Habing, B., Douglas, J., Kim, H., Rousos, L. & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment, *Applied Psychological Measurement* **20**, 331–354.
 - [19] Van der Ark, L.A. (2004). Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models, *Psychometrika* (in press).
 - [20] Van der Linden, W.J. & Hambleton, R.K., (eds) (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

KLAAS SIJTSMA

Nonparametric Regression

JOHN FOX

Volume 3, pp. 1426–1430

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonparametric Regression

Nonparametric regression analysis traces the dependence of a response variable (y) on one or several predictors (x s) without specifying in advance the function that relates the response to the predictors:

$$E(y_i) = f(x_{1i}, \dots, x_{pi}) \quad (1)$$

where $E(y_i)$ is the mean of y for the i th of n observations. It is typically assumed that the conditional variance of y , $\text{Var}(y_i | x_{1i}, \dots, x_{pi})$ is a constant, and that the conditional distribution of y is normal, although these assumptions can be relaxed.

Nonparametric regression is therefore distinguished from linear regression (see **Multiple Linear Regression**), in which the function relating the mean of y to the x s is linear in the parameters,

$$E(y_i) = \alpha + \beta_1 x_{1i} + \dots + \beta_p x_{pi} \quad (2)$$

and from traditional nonlinear regression, in which the function relating the mean of y to the x s, though nonlinear in its parameters, is specified explicitly,

$$E(y_i) = f(x_{1i}, \dots, x_{pi}; \gamma_1, \dots, \gamma_k) \quad (3)$$

In traditional regression analysis, the object is to estimate the parameters of the model – the β s or γ s. In nonparametric regression, the object is to estimate the regression function directly.

There are many specific methods of nonparametric regression. Most, but not all, assume that the regression function is in some sense smooth. Several of the more prominent methods are described in this article. Moreover, just as traditional linear and nonlinear regression can be extended to **generalized linear model** and nonlinear regression models that accommodate nonnormal error distributions, the same is true of nonparametric regression. There is a large literature on nonparametric regression analysis, both in scientific journals and in texts. For more extensive introductions to the subject, see in particular, Bowman and Azzalini [1], Fox [2, 3], Hastie, Tibshirani, and Freedman [4], Hastie and Tibshirani [5], and Simonoff [6].

The simplest use of nonparametric regression is in smoothing scatterplots (see **Scatterplot Smoothers**). Here, there is a numerical response y and a single predictor x , and we seek to clarify visually

the relationship between the two variables in a scatterplot. Figure 1, for example, shows the relationship between female expectation of life at birth and GDP per capita for 154 nations of the world, as reported in 1998 by the United Nations. Two fits to the data are shown, both employing local-linear regression (described below); the solid line represents a fit to all of the data, while the broken line omits four outlying nations, labelled on the graph, which have values of female life expectancy that are unusually low given GDP per capita. It is clear that although there is a positive relationship between expectation of life and GDP, the relationship is highly nonlinear, leveling off substantially at high levels of GDP.

Three common methods of nonparametric regression are *kernel estimation* (see **Kernel Smoothing**), *local-polynomial regression* (which is a generalization of kernel estimation), and *smoothing splines*. *Nearest-neighbor* kernel estimation proceeds as follows (as illustrated for the UN data in Figure 2):

1. Let x_0 denote a focal x -value at which $f(x)$ is to be estimated; in Figure 2(a), the focal value is the 80th ordered x -value in the UN data, $x_{(80)}$. Find the m nearest x -neighbors of x_0 , where $s = m/n$ is called the *span* of the kernel smoother. In the example, the span was set to $s = 0.5$, and thus $m = 0.5 \times 154 = 77$. Let h represent the half-width of a window encompassing the m nearest neighbors of x_0 . The larger the span (and hence the value of h), the smoother the estimated regression function.
2. Define a symmetric unimodal weight function, centered on the focal observation, that goes to zero (or nearly zero) at the boundaries of the neighborhood around the focal value. The specific choice of weight function is not critical: In Figure 2(b), the tricube weight function is used:

$$w_T(x) = \begin{cases} \left[1 - \left(\frac{|x - x_0|}{h} \right)^3 \right]^3 & \text{for } \frac{|x - x_0|}{h} < 1 \\ 0 & \text{for } \frac{|x - x_0|}{h} \geq 1 \end{cases} \quad (4)$$

A Gaussian (normal) density function is another common choice.

3. Using the tricube (or other appropriate) weights, calculate the weighted average of the y -values to

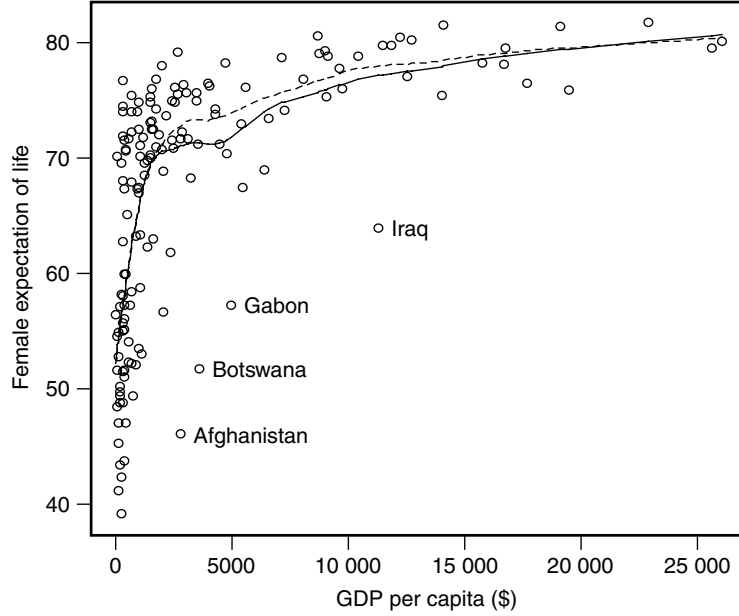


Figure 1 Female expectation of life by GDP per capita, for 154 nations of the world. The solid line is for a local-linear regression with a span of 0.5, while the broken line is for a similar fit deleting the four outlying observations that are labeled on the plot

obtain the fitted value

$$\hat{y}_0 = \hat{f}(x_0) = \frac{\sum W_T(x_i)y_i}{\sum W_T(x_i)} \quad (5)$$

as illustrated in Figure 2(c). Greater weight is thus accorded to observations whose x -values are close to the focal x_0 .

4. Repeat this procedure at a range of x -values spanning the data – for example, at the ordered observations $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. Connecting the fitted values, as in Figure 2(d), produces an estimate of the regression function.

Local-polynomial regression is similar to kernel estimation, but the fitted values are produced by locally weighted regression rather than by locally weighted averaging; that is, $\hat{y}_0$ is obtained in step 3 by the polynomial regression of y on x to minimize the weighted sum of squared residuals

$$\sum W_T(x_i)(y_i - a - b_1x_i - b_2x_i^2 - \dots - b_kx_i^k)^2 \quad (6)$$

Most commonly, the order of the local polynomial is taken as $k = 1$, that is, a local linear fit (as

in Figure 1). Local- polynomial regression tends to be less biased than kernel regression, for example, at the boundaries of data – as is evident in the artificial flattening of the kernel estimator at the right of Figure 2(d). More generally, the bias of the local-polynomial estimator declines and the variance increases with the order of the polynomial, but an odd-ordered local-polynomial estimator has the same asymptotic variance as the preceding even-ordered estimator: Thus, the local-linear estimator (of order 1) is preferred to the kernel estimator (of order 0), and the local-cubic (order 3) estimator to the local-quadratic (order 2).

Smoothing splines are the solution to the *penalized regression problem*: Find $\hat{f}(x)$ to minimize

$$S(h) = \sum [y_i - f(x_i)]^2 + h \int_{x(1)}^{x(n)} [f''(x)]^2 dx \quad (7)$$

Here h is a *roughness penalty*, analogous to the span in nearest-neighbor kernel or local-polynomial regression, and f'' is the second derivative of the

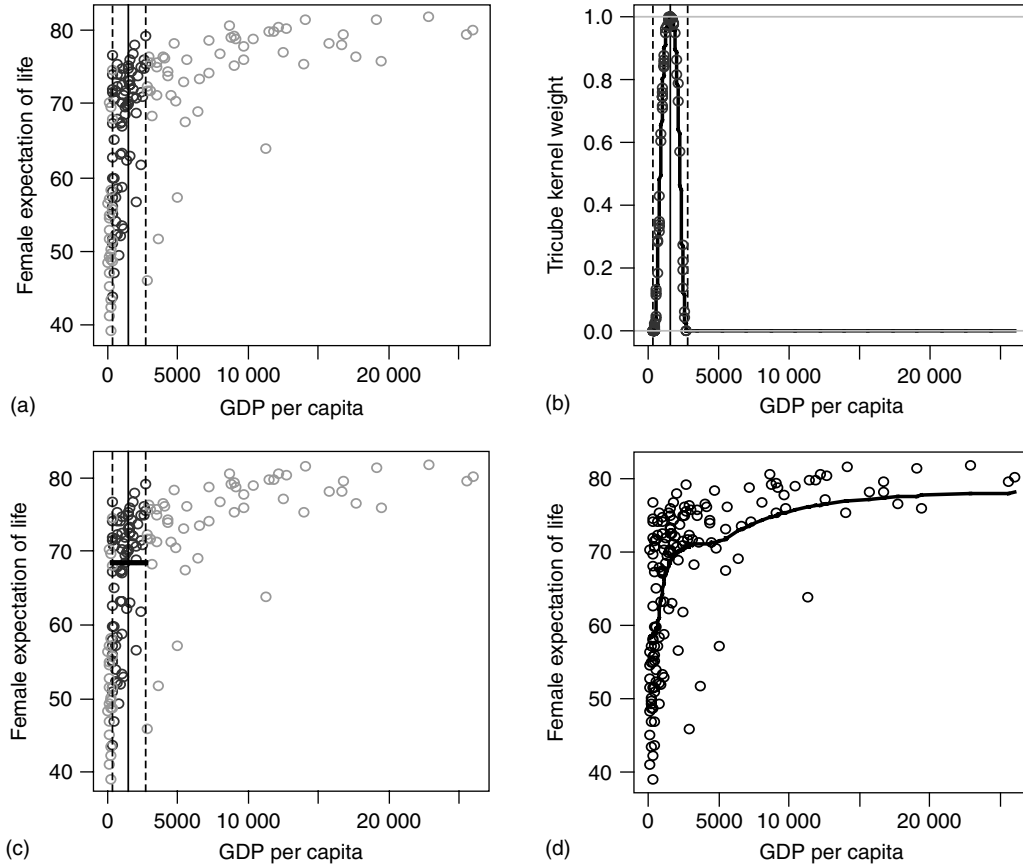


Figure 2 How the kernel estimator works: (a) A neighborhood including the 77 observations closest to $x_{(80)}$, corresponding to a span of 0.5. (b) The tricube weight function defined on this neighborhood; the points show the weights for the observations. (c) The weighted mean of the y -values within the neighborhood, represented as a horizontal line. (d) The nonparametric regression line connecting the fitted values at each observation. (The four outlying points are excluded from the fit)

regression function (taken as a measure of roughness). Without the roughness penalty, nonparametrically minimizing the residual sum of squares would simply interpolate the data. The mathematical basis for smoothing splines is more satisfying than for kernel or local polynomial regression, since an explicit criterion of fit is optimized, but spline and local-polynomial regressions of equivalent smoothness tend to be similar in practice.

Local regression with several predictors proceeds as follows, for example. We want the fit $\hat{y}_0 = \hat{f}(\mathbf{x}_0)$ at the focal point $\mathbf{x}_0 = (x_{10}, \dots, x_{p0})$ in the predictor space. We need the distances $D(\mathbf{x}_i, \mathbf{x}_0)$ between the observations on the predictors and the focal point. If the predictors are on the same scale (as, for example,

when they represent coordinates on a map), then measuring distance is simple; otherwise, some sort of standardization or generalized distance metric will be required. Once distances are defined, weighted polynomial fits in several predictors proceed much as in the bivariate case. Some kinds of spline estimators can also be generalized to higher dimensions.

The generalization of nonparametric regression to several predictors is therefore mathematically straightforward, but it is often problematic in practice. First, multivariate data are afflicted by the so-called *curse of dimensionality*: Multidimensional spaces grow exponentially more sparse with the number of dimensions, requiring very large samples to estimate nonparametric regression models with many

4 Nonparametric Regression

predictors. Second, although slicing the surface can be of some help, it is difficult to visualize a regression surface in more than three dimensions (that is, for more than two predictors).

Additive regression models are an alternative to unconstrained nonparametric regression with several predictors. The additive regression model is

$$E(y_i) = \alpha + f_1(x_{1i}) + \cdots + f_p(x_{pi}) \quad (8)$$

where the f_j are smooth partial-regression functions, typically estimated with smoothing splines or by local regression. This model can be extended in two directions: (1) To incorporate interactions between (or among) specific predictors; for example

$$E(y_i) = \alpha + f_1(x_{1i}) + f_{23}(x_{2i}, x_{3i}) \quad (9)$$

which is not as general as the unconstrained model $E(y_i) = \alpha + f(x_{1i}, x_{2i}, x_{3i})$. (2) To incorporate linear terms, as in the model

$$E(y_i) = \alpha + \beta_1 x_{1i} + f_2(x_{2i}) \quad (10)$$

Such *semiparametric* models are particularly useful for including dummy regressors or other contrasts derived from categorical predictors.

Returning to the UN data, an example of a simple additive regression model appears in Figures 3 and 4. Here female life expectancy is regressed on GDP per capita and the female rate of illiteracy, expressed

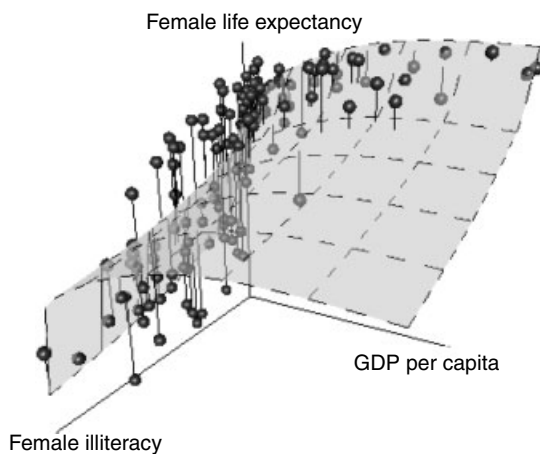


Figure 3 Fitted regression surface for the additive regression of female expectation of life on GDP per capita and female illiteracy. The vertical lines represent residuals

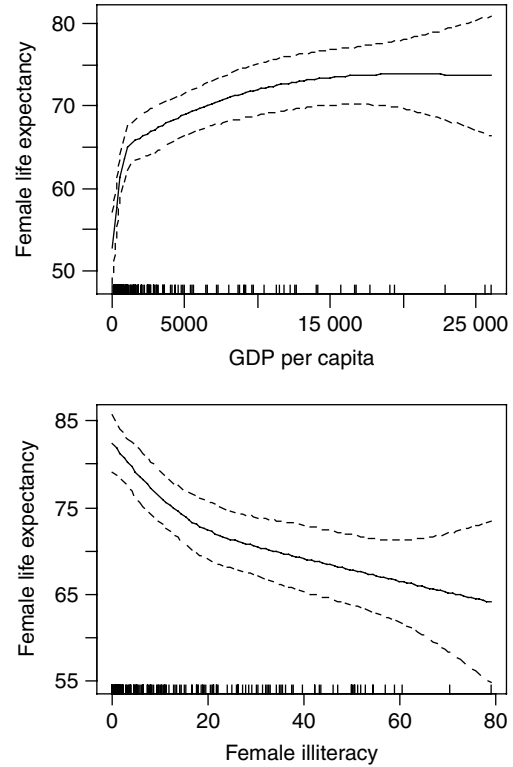


Figure 4 Partial regression functions from the additive regression of female life expectancy on GDP per capita and female illiteracy; each term is fit by a smoothing spline using the equivalent of four degrees of freedom. The “rug plot” at the bottom of each graph shows the distribution of the predictor, and the broken lines give a point-wise 95-percent confidence envelope around the fit

as a percentage. Each term in this additive model is fit as a smoothing spline, using the equivalent of four degrees of freedom. Figure 3 shows the two-dimensional fitted regression surface, while Figure 4 shows the partial-regression functions, which in effect slice the regression surface in the direction of each predictor; because the surface is additive, all slices in a particular direction are parallel, and the two-dimensional surface in three-dimensional space can be summarized by two two-dimensional graphs. The ability to summarize the regression surface with a series of two-dimensional graphs is an even greater advantage when the surface is higher-dimensional.

A central issue in nonparametric regression is the selection of smoothing parameters – such as the span in kernel and local-polynomial regression or the

roughness penalty in smoothing-spline regression (or equivalent degrees of freedom for any of these). In the examples in this article, smoothing parameters were selected by visual trial and error, balancing smoothness against detail. The analogous statistical balance is between variance and bias, and some methods (such as **cross-validation**) attempt to select smoothing parameters to minimize estimated mean-square error (i.e., the sum of squared bias and variance).

References

- [1] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Oxford University Press, Oxford.
- [2] Fox, J. (2000a). *Nonparametric Simple Regression: Smoothing Scatterplots*, Sage Publications, Thousand Oaks.
- [3] Fox, J. (2000b). *Multiple and Generalized Nonparametric Regression*, Sage Publications, Thousand Oaks.
- [4] Hastie, T., Tibshirani, R. & Freedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.
- [5] Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall, London.
- [6] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

(See also **Generalized Additive Model**)

JOHN FOX

Nonrandom Samples

HOWARD WAINER

Volume 3, pp. 1430–1433

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonrandom Samples

Many of the questions that the tools of modern statistics are enlisted to help answer are causal. To what extent does exercise cause a decrease in coronary heart disease? How much of an effect does diet have on cholesterol in the blood? How much of a child's performance in school is due to home environment? What proportion of a product's sales can be traced to a new advertisement campaign? How much of the change in national test scores can be attributed to modifications in educational policy?

Some of these questions lend themselves to the possibility of experimental study. One could randomly assign a large group of individuals to a program of exercise, command another to a more sedentary life, and then compare the coronary health of the two groups years later (see **Clinical Trials and Intervention Studies**). This is possible but impractical. Similar *gedanken* experiments could be imagined on diet, but the likelihood of actually accomplishing any long-term controlled experiment on humans is extremely small. Trying to control the home life of children borders on the impossible.

Modern epistemology [3] points out that the **randomization** associated with a true experimental design is the only surefire way to measure the effects of possible causal agents. The randomization allows us to control for any 'missing third factor' because, under its influence, any unknown factor will be evenly distributed (in the limit) in both the treatment and control conditions. Yet, as we have just illustrated, many of modern causal questions do not lend themselves easily to the kind of randomized assignment that is the hallmark of tightly controlled experiments. Instead, we must rely on observational studies – studies in which our data are not randomly drawn.

In an **observational study**, we must examine intact groups – perhaps one group of individuals who have followed a program of regular exercise and another who have not – and compare their cardiovascular health. But it is almost certain that these two groups differ on some other variable as well – perhaps diet or age or weight or

social class. These must be controlled for statistically. Without random assignment, we can never be sure we have controlled for everything, but this is the best that we can do. In fact, thanks to the ingenious efforts of many statisticians (e.g., [1, 2, 8, 9, 10, 11]) in many circumstances, there are enough tools available for us to do very well indeed.

Let us now consider three separate circumstances involving inferences from nonrandomly gathered data. In the first two, the obvious inferences are wrong, and in the third, we show an ingenious solution that allows what appears to be a correct inference. In each case, we are overly terse, but provide references for more detailed explanation.

Example 1. The most dangerous profession In 1835, the Swiss physician H. C. Lombard [5] published the results of a study on the longevity of various professions. His data were very extensive, consisting of death certificates gathered over more than a half century in Geneva. Each certificate contained the name of the deceased, his profession, and age at death. Lombard used these data to calculate the mean longevity associated with each profession. Lombard's methodology was not original with him, but, instead, was merely an extension of a study carried out by R. R. Madden, Esq. that was published two years earlier [6]. Lombard found that the average age of death for the various professions mostly ranged from the early 50s to the mid 60s. These were somewhat younger than those found by Madden, but this was expected since Lombard was dealing with ordinary people rather than the 'geniuses' in Madden's (the positive correlation between fame and longevity was well known even then). But Lombard's study yielded one surprise; the most dangerous profession – the one with the shortest longevity – was that of 'student' with an average age of death of only 20.7! Lombard recognized the reason for this anomaly, but apparently did not connect it to his other results (for more details see [12]).

Example 2. The twentieth century is a dangerous time In 1997, to revisit Lombard's methodology, Samuel Palmer and Linda Steinberg (reported in [13]) gathered 204 birth and death dates from the Princeton (NJ) Cemetery. This cemetery opened in the mid 1700s, and has people buried in it born in

2 Nonrandom Samples

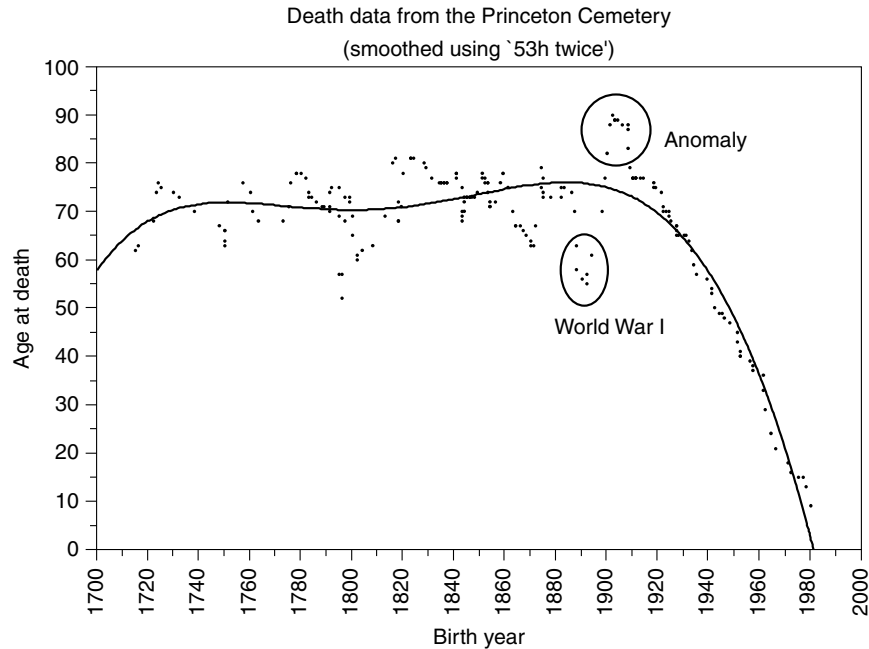


Figure 1 The longevities of 204 people buried in Princeton Cemetery shown as a function of the year of their birth

the early part of that century. Those interred include Grover Cleveland, John Von Neumann, Kurt Gödel, and **John Tukey**.

When age-at-death was plotted as a function of birth year (after suitable smoothing to make the picture coherent), we see the result shown as Figure 1. The age of death stays relatively constant until 1920, when the longevity of the people in the cemetery begins to decline rapidly. The average age of death decreases from around 70 years of age in the 1900s to as low as 10 in the 1980s. It becomes obvious immediately that there must be a reason for the anomaly in the data (what we might call the 'Lombard Surprise'), but what? Was it a war or a plague that caused the rapid decline? Has a neonatal section been added to the cemetery? Was it opened to poor people only after 1920? Obviously, the reason for the decline is nonrandom sampling. People cannot be buried in the cemetery if they are not already dead. Relatively few people born in the 1980s are buried in the cemetery and, thus, no one born in the 1980s that we found in Princeton Cemetery could have been older than 17.

Examples of situations in which this anomaly arises abound. Four of these are:

1. In 100 autopsies, a significant relationship was found between age-at-death and the length of the lifeline on the palm [7]. Actually, what they discovered was that wrinkles and old age go together.
2. In 90% of all deaths resulting from barroom brawls, the victim was the one who instigated the fight. One questions the wit of the remaining 10% who did not point at the body on the floor when the police asked, 'Who started this?'
3. In March of 1991, the *New York Times* reported the results of data gathered by the American Society of Podiatry, which stated that 88% of all women wear shoes at least one size too small. Who would be most likely to participate in such a poll?
4. In testimony before a February 1992 meeting of a committee of the Hawaii State Senate, then considering a law requiring all motorcyclists to wear a helmet, one witness declared that, despite having been in several accidents during his 20 years of motorcycle riding, a helmet would not have prevented any of the injuries he received. Who was unable to testify? Why?

Example 3. Bullet holes and a model for missing data Abraham Wald, in some work he did during World War II [14], was trying to determine where to add extra armor to planes on the basis of the pattern of bullet holes in returning aircraft. His conclusion was to determine carefully where returning planes had been shot and *put extra armor every place else!*

Wald made his discovery by drawing an outline of a plane (crudely shown in Figure 2), and then putting a mark on it where a returning aircraft had been shot. Soon, the entire plane had been covered with marks *except* for a few key areas. It was at this point that he interposed a model for the missing data, the planes that did not return. He assumed that planes had been hit more or less uniformly, and, hence, those aircraft hit in the unmarked places had been unable to return, and, thus, were the areas that required more armor.

Wald's key insight was his model for the non-response. From his observation that planes hit in certain areas were still able to return to base, Wald inferred that the planes that did not return must have been hit somewhere else. Note that if he used a different model analogous to 'those lying within Princeton Cemetery have the same longevity as those without' (i.e., that the planes that returned were hit about the same as those that did not return), he would have arrived at exactly the opposite (and wrong) conclusion.

To test Wald's model requires heroic efforts. Planes that did not return must be found and the patterns of bullet holes in them must be recorded. In short, to test the validity of Wald's model for missing data requires that we sample from the unselected population – we must try to get a random sample,

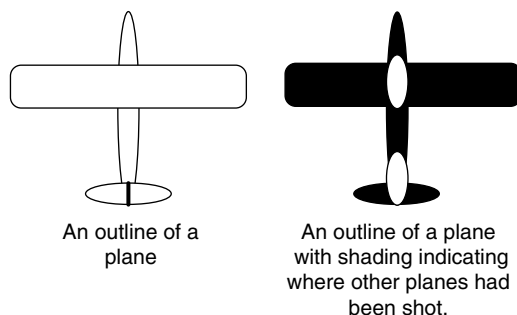


Figure 2 A schematic representation of Abraham Wald's ingenious scheme to investigate where to armor aircraft

even if it is a small one. This strategy remains the basis for the only empirical solution to making inferences from nonrandom samples.

Nonresponse in surveys (*see Nonresponse in Sample Surveys*) exhibits many of the same problems seen in observational studies. Indeed, in a mathematical sense, they are formally identical. Yet, in practice, there are important distinctions that can usefully be drawn between them [4]. This essay was aimed at providing a flavor of the depth of the problems associated with making correct inferences from nonrandomly gathered data, and an illustration of the character of one solution – Abraham Wald's – which effectively models the process that generated the data in the hope of uncovering its secrets.

References

- [1] Cochran, W.G. (1982). *Contributions to Statistics*, Wiley, New York.
- [2] Cochran, W.G. (1983). *Planning and Analysis of Observational Studies*, Wiley, New York.
- [3] Holland, P.W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–960.
- [4] Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
- [5] Lombard, H.C. (1835). De l'Influence des professions sur la Durée de la Vie, *Annales d'Hygiène Publique et de Médecine Légale* **14**, 88–131.
- [6] Madden, R.R. (1833). *The Infirmities of Genius, Illustrated by Referring the Anomalies in Literary Character to the Habits and Constitutional Peculiarities of Men of Genius*, Saunders and Otley, London.
- [7] Newrick, P.G., Affie, E. & Corral, R.J.M. (1990). Relationship between longevity and lifeline: a manual study of 100 patients, *Journal of the Royal Society of Medicine* **83**, 499–501.
- [8] Rosenbaum, P.R. (1995). *Observational Studies*, Springer-Verlag, New York.
- [9] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.
- [10] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse*, Wiley, New York.
- [11] Wainer, H. (1986). *Drawing Inferences from Self-Selected Samples*, Springer-Verlag, New York. Reprinted in 2000 by Lawrence Erlbaum Associates: Hillsdale, NJ.
- [12] Wainer, H. (1999). The most dangerous profession: A note on nonsampling error, *Psychological Methods* **4**(3), 250–256.
- [13] Wainer, H., Palmer, S.J. & Bradlow, E.T. (1998). A selection of selection anomalies, *Chance* **11**(2), 3–7.

4 Nonrandom Samples

- [14] Wald, A. (1980). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*, CRC 432, July 1980. (These are reprints of work done by Wald while a member of Columbia's Statistics Research Group during the period 1942–1945. Copies can be obtained from the

Document Center, Center for Naval Analyses, 2000 N. Beauregard St., Alexandria, VA 22311.)

HOWARD WAINER

Nonresponse in Sample Surveys

PAUL S. LEVY AND STANLEY LEMESHOW

Volume 3, pp. 1433–1436

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonresponse in Sample Surveys

Consider a sample survey of n units from a population containing N units. During the data collection phase of the survey, attempts are made to collect information from each of the n sampled units. Generally, for a variety of reasons, there is a subset of units from which it is not possible to collect the information. For example, if the units are patient hospital medical records, the records may be missing. If the unit is a household, and the survey uses telephone interviewing, the household respondent may rarely be at home or may be unwilling to participate in the survey. These are just two of a countless number of scenarios in which information cannot be obtained from a sampled unit (*see Missing Data*). The nonresponse from a sampled unit is considered to be one of the major threats to the validity of information obtained by the use of sample surveys [6–8].

Effects of Nonresponse

The nonresponse of sample units is responsible for a smaller sample size, and, hence, larger variances or standard errors in the resulting estimates than were planned for in the original study design. More importantly, the nonresponse results in bias that is insidious because it cannot be estimated directly from the survey data. For any variable measured in the survey, this bias is equal to the product of the expected proportion of nonrespondents in the population and the expected difference between the level of the variable among respondents and nonrespondents [9, 12]. Thus, if 30% of households in a population are potential nonrespondents, and if 40% of the potential nonrespondents are below the poverty level as opposed to 10% of the potential respondents, then the expected estimated proportion of households below the poverty level from the survey will be 10%, which is a gross underestimate of the true level which is 19% ($0.30 \times 40\% + 0.70 \times 10\%$).

Types of Response Rates and Issues in Reporting Them

Because of the potential biases due to nonresponse, it is very important that response rates be reported in

the findings of a sample survey. There are many types of response rates that can be reported, each leading to a different interpretation concerning potential biases due to nonresponse. To illustrate, let us suppose that a firm has given special training on the use of a new software package to 1000 employees and, a year later, selects a random sample of 200 for interviews on their satisfaction with the software and the training. Of the 200, they are able to successfully interview 120, which is a response rate of 60% of those given the training. However, of the 80 that were not interviewed, 50 were no longer employed at the firm; 10 refused to be interviewed; and 20 were either ill or on vacation during the period in which interviews occurred. Thus, if one considers only those available during the period of the survey, the response rate would be 120/130 (92.3%). If one considers only those still employed during the survey, the response rate would be 120/150 (80%). Each of these three response rates would lead to a different interpretation with respect to the quality of the data and with respect to interpretations concerning the findings of the survey.

Guidelines frequently cited and used by survey professionals concerning the use of different types of response rates are given in articles by the Council of American Survey Research Organizations (CASRO) [3] and the American Association for Public Opinion Research (AAPOR) [1]. The material in these articles can be downloaded from their respective websites: <http://www.casro.org/resprates.cfm> and <http://www.aapor.org>. The sampling text by Lohr also discusses the various types of response rates commonly used as well as guidelines on what constitutes an acceptable response rate [12, pp. 281–282].

Methods of Improving Response Rates

While 100% response rates are generally unattainable, there are a variety of methods commonly used to improve response rates. Most of these methods can be found in recent textbooks on sampling or survey methodology [9, 12], or in more specialized publications dealing with response issues [4, 6–8, 10, 11]. We now present a brief description of several widely used methods.

Endorsements. Response rates can be increased if the survey material contains endorsements by an

2 Nonresponse in Sample Surveys

agency or organization whose sphere of interest is strongly associated with the subject matter of the survey. This works best if the name of the endorsing organization or individual signing the endorsement letter is well known to the potential respondents. This is especially effective in surveys of business, trade, or professional establishments.

Incentives. Monetary or other gifts for participation in the survey have been shown to be effective in obtaining the participation of selected units [6]. Generally, the value of the incentive should be proportional to the extent of the burden to the unit. With the increased attention given to protection of human subjects, there is a tendency for Institutional Review Boards to frown on excessive gifts, fearing that the incentives would be a form of coercion.

Refusal Conversion Attempts. Attempts to convert initial refusals often involve the use of interviewers of a different age/gender/ethnicity than that used initially (especially in face-to-face surveys). Changes in the interviewing script with more flexibility permitted for the interviewer to *ad lib* (especially in telephone surveys), and attempts to interview at a different time from that used initially are also widely used. Although the success of refusal conversion can vary considerably depending on the subject matter and interviewing mode of the survey, success rates are generally in the 15 to 25% range.

Subsampling Nonrespondents (Double Sampling). If a sample of n units from a population containing N units results in n_1 respondents and $n_2 = n - n_1$ nonrespondents, a random subsample of n^* units is taken from the n_2 units that were nonrespondents at the first phase of sampling. Intensive attempts are made to obtain data from this sample of original nonrespondents. If k of the n^* subsampled original nonrespondents are successfully interviewed, then by weighting these units by the factor n_2/k , we can obtain estimates that partially compensate for the nonresponses. The accuracy of the estimates obtained by this method depends, to a large extent, on obtaining a high response rate from the sample of original nonrespondents. Detailed descriptions of this method are given in [9].

Advance Letters in Telephone Sampling. With the switch in emphasis from the Mitofsky–Waksberg

[13] method to list assisted methods [2] of random digit dialing (RDD) in telephone surveys, it is possible to obtain addresses for a relatively high proportion of the telephone numbers obtained by RDD. By sending advance letters to the households corresponding to these numbers informing them of the nature and importance of the survey, and requesting their support, it has been shown that the response rate can be increased [11].

Multimode Options. In many surveys, where one mode of data collection (e.g., face-to-face personal interview) has not succeeded in obtaining an interview, it is possible that offering the potential target unit an alternative mode (e.g., web or mail) will result in a response. Link et al. [10] has performed an experiment in which this technique has resulted in increased response rates.

Differential Effects on Subpopulations. Although the above methods may increase overall response rates, there may be differences among subpopulations with respect to their level of effectiveness, and this may increase the diversity among various population groups with respect to response rates. How this would effect biases in the overall survey findings has become a recent concern as suggested in the recent paper by Eyerman et al. [4].

Statistical Methods of Adjusting for Nonresponse

In spite of efforts such as those mentioned above to collect data from all units selected in the sample, there will often be some units from whom survey data have not been collected. There are several methods used to adjust for this in the subsequent statistical estimation process. The ‘classic’ method is to construct classes based on characteristics that are thought to be related to the responses to the items in the survey and that are known about the nonrespondent units (often these are ecological or demographic). If a given class has n units in the original sample, but only n^* respondent units, then the initial sampling weight is multiplied by n/n^* for every unit in the particular class. This, of course, does not consider the possibility that nonrespondents, even within the same class, might differ from respondents with respect to their levels of variables in the survey. More refined methods of

adjustment (e.g., *raking*, *propensity scores*) are now used fairly commonly and are discussed in more recent literature [5].

References

- [1] American Association for Public Opinion Research (2004). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for RDD Telephone Surveys and In-Person Household Surveys*, AAPOR, Ann Arbor.
- [2] Casady, R.J. & Lepkowski, J.M. (1993). Stratified telephone sampling designs, *Survey Methodology* **19**(1), 103–113.
- [3] Council of American Survey Research Organizations. (1982). *On the Definition of Response Rates*, A special report of the CASRO task force on completion rates. CASRO, Port Jefferson.
- [4] Eyerman, J., Link, M., Mokdad, A. & Morton, J. (2004). Assessing the impact of methodological enhancements on different subpopulations in a BRFSS experiment., in *Proceedings of the American Statistical Association, Survey Methodology Section*, American Statistical Association (CD-ROM), Alexandria.
- [5] Folsom, R.E. & Singh, A.C. (2000). A generalized exponential model for sampling weight calibration for a unified approach to nonresponse, poststratification, and extreme weight adjustments, in *Proceedings of the American Statistical Association, Section on Survey Research Methods* 598–603.
- [6] Groves, R.M. (1989). *Survey Errors and Survey Costs*, Wiley, New York.
- [7] Groves, R.M. & Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*, Wiley, New York.
- [8] Groves, R.M., Dillman, D.A., Eltinge, J.L. & Little, R.J.A. (2002). *Survey Nonresponse*, Wiley, New York.
- [9] Levy, P.S. & Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications*, Wiley, New York.
- [10] Link, M.W. & Mokdad, A. (2004). Are web and mail modes feasible options for the behavioral risk factor surveillance system? in *Proceedings of the 8th Conference on Health Survey Research Method*, Centers for Disease Control and Prevention, Atlanta.
- [11] Link, M.W., Mokdad, A., Town, M., Roe, D. & Weiner, J. (2004). Improving response rates for the BRFSS: use of lead letters and answering machine messages, in *Proceedings of the American Statistical Association, Survey Methodology Section*, American Statistical Association (CD-ROM), Alexandria.
- [12] Lohr, S. (1999). *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove.
- [13] Waksberg, J. (1978). Sampling methods for random digit dialing, *Journal of the American Statistical Association* **73**, 40–46.

(See also **Survey Sampling Procedures**)

PAUL S. LEVY AND STANLEY LEMESHOW

Nonshared Environment

HILARY TOWERS AND JENAE M. NEIDERHISER

Volume 3, pp. 1436–1439

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nonshared Environment

Nonshared environment refers to the repeated finding that across a very wide range of individual characteristics and behaviors, children raised in the same home are often very *different* from one another [6]. Most importantly, these behavioral differences are the product of some aspect of the children's environment – and not to differences in their genes. Nonshared environmental influences may include different peer groups or friendships, differential parenting, differential experience of trauma, different work environments, and different romantic partners or spouses. The difficulty, however, has been in finding evidence of the substantial and systematic impact of any of these variables on sibling differences. Nonshared environment remains a 'hot topic' in the field of quantitative genetic research because it has been shown to be important to a wide array of behaviors across the entire life-span, and yet its *systematic* causes remain elusive [12].

How are we able to discern the extent to which nonshared environmental factors are operating on a behavior? Just as genetic influences on behaviors are inferred from comparisons of different sibling types, so can nonshared environmental outcomes be detected through similar comparisons. For example, monozygotic (MZ, or identical) twins are 100% genetically identical. This means that when members of an MZ twin pair differ, these differences cannot be due to genetic differences – they must be the result of differential effects of factors within each twin's respective *environment*. This unique opportunity to separate the effects of genes and environment on behavior has made identical twins very popular among behavioral geneticists (see **Twin Designs**).

Prior to describing the promising results of one recent study of nonshared environmental associations, we wish to point out an important distinction among such studies. While it is clear that within-family differences are substantial for most behaviors, until recently the *sources* of these differences have remained 'anonymous'. That is, quantitative genetic analyses of individual behaviors at one time point (i.e., univariate, cross-sectional analyses) can only tell us siblings are different because of factors in their environments. For example, if we find substantial differences in levels of depression *within MZ twin pairs*, nonshared environmental influence

on depression is indicated. We still do not know, however, what factors are causing, or contributing to, these behavioral differences. It is only recently that researchers have begun to assess the causes of within-family differences by examining nonshared environmental *associations* between constructs thought to be environmental (e.g., peer relations) and adjustment (e.g., depression).

In the simplest form of these bivariate analyses, MZ twin differences on an environmental factor are correlated with MZ twin differences on an adjustment factor. This correlation between MZ twin differences on two constructs allows the 'pure' nonshared environmental association for this association to be estimated. The MZ twin difference method is ideal in its relative simplicity and its stringency, since reduced sample sizes and the general tendency for MZ twins to be *more* similar rather than less similar to one another than DZ twins decrease the likelihood of finding significant associations. Yet, there have been very few quantitative genetic analyses that have used an MZ differences approach. Of those that have, there has been an almost exclusive reliance on self-reports of both the environment, often retrospective family environment, and current psychological adjustment (e.g. [1, 2]). While this does not diminish the importance of the findings, the use of within-reporter difference correlations (i.e., correlations between two measures reported by the same informant) is subject to the same criticism that applies to standard, two-variable correlations: the correlations are more likely to represent rater bias of some sort than are correlations between ratings from two different reporters (see **Rater Bias Models**).

One recent MZ difference analysis of nonshared environmental influences on the socio-emotional development of preschool children provides evidence for at least moderate, and in some cases substantial, nonshared environmental associations using parent, interviewer, and observer ratings [3]. In this study, across-rater MZ difference correlations between interviewer reports of parental harshness, for example, and parent reports of the children's problem behaviors, emotionality, and prosocial behavior were .36, .38, and -.27, respectively. Cross-rater MZ difference correlations were also significant for: (a) parental positivity and child positivity, (b) parental positivity and child problem behavior (a negative association), (c) parental negative control and child prosocial behavior (a negative association), and (d) parental

2 Nonshared Environment

harsh discipline and child responsiveness (a negative association). These findings can be interpreted as evidence that the *differential treatment* of very young siblings living in the same family by their parents *matters* – and it matters *apart from any genetically influenced traits of the child*. For example, when a parent is harsh toward one child, and less harsh to the other, the child receiving the harsh discipline is more likely to have behavior problems, be more emotionally labile, and be less prosocial than their sibling.

One alternative explanation of the findings reported by Deater–Deckard and colleagues [3] is that differences in the children’s behaviors elicit differences in parenting, or that the relationship is reciprocal. Unfortunately, it is not possible to directly test for this using cross-sectional data. A similar set of preliminary analyses of parent–child data from an adult twin sample, the Twin Moms project [7], found several substantial cross-rater MZ difference correlations between the differential parenting patterns of adult twin women and characteristics of their adolescent children [10]. While it is also not possible to confirm directionality in this cross-sectional sample, such findings in an *adult-based* sample imply that differences in the children are contributing to differences in (or nonshared environmental variation in) the parenting behavior of the mothers, since the mothers’ genes and environment are the focus of assessment.

Regardless of the direction of effect, the associations reported by Deater–Deckard and colleagues [3] are remarkable for at least two reasons. First, within-family differences have been shown to increase over time – and to be less easily detected in early childhood, when parents are more likely to treat their children similarly [5, 9]. Finding associations between differential parenting, a family environment variable, and childhood behaviors using a sample of three and a half-year-old children is notable, and has implications for clinical intervention efforts directed toward promoting more ‘equitable’ parenting of siblings. Second, the fact that these associations exist across three different rater ‘perspectives’ lends credence to the validity of the associations – they cannot be attributed entirely to the bias of a single reporter. Third, unlike univariate analyses of nonshared environment, analyses of multivariable associations minimize the amount of measurement error in the estimate of nonshared environment.

In light of this apparent success, one may ask, why all the excitement over nonshared environment? Again, the issue is an inability to isolate *systematic* sources of within-family differences in behavior, which are so prevalent. In other words – where single studies have succeeded in specifying sources of sibling differences, there has been a general failure to replicate these findings across samples. A recent **meta-analysis** of the effect produced by candidate sources of nonshared environment on behavior found the average effect size of studies *using genetically informed designs* to be roughly 2% [12]. Revelations of this kind have stirred discussion about the definition of nonshared environment itself [4, 8]. So far, however, the study of nonshared environmental associations has been limited almost exclusively to analyses of children and adolescents.

As adult siblings often go their separate ways – rent their own place, get a job, get married, and so on – it comes as no surprise that nonshared environmental variation may increase with age [5]. It seems a viable possibility, then, that primary sources of nonshared environmental influence on behavior are to be found in adulthood. Indeed, results from the Twin Moms project [7] indicate that a portion of the nonshared environmental variation in women’s marital satisfaction covaries with differences in the women’s spousal characteristics, current family environment, and stressful life events [11]. In other words, the extent to which women are satisfied or dissatisfied with their marriages has been found to vary as a function of spouse characteristics, family climate, and stressors. Importantly, these relationships exist *independently of any genetic influence* that may be operating on her feelings about her marriage. These findings confirm the potential to isolate nonshared environmental relationships among adult behaviors and relationships, and suggest an important area of exploration for future behavioral genetic work.

Finally, studies of nonshared environmental associations are important because they provide information that complements our understanding of genetic influences on behavior and relationships. Our success in quantifying the relative impact of genetic *and* environmental factors (and their mutual interplay) on human behavior has implications for therapeutic intervention. For example, the purpose of many types of therapy is to produce changes in individuals’ social environments. How much more effective these therapies may be when based on knowledge that

a particular psychosocial intervention has the potential to alter the mechanisms of genetic expression for a particular trait. For instance, intervention research has demonstrated the potential for altering behavioral responses of mothers toward their infants with difficult temperaments [13]. The intent is not only to change mothers' behaviors or thought processes, but to eliminate the *evocative association* between the infants' genetically influenced, irritable behaviors and their mothers' corresponding expressions of hostility. By identifying specific sources of nonshared environmental influences, we will be better able to focus our attention on the environmental factors that matter in eliciting behavior, as in the intervention study cited above, or that are prime candidates for examination in studies looking for **genotype x environment interaction**.

References

- [1] Baker, L.A. & Daniels, D. (1990). Nonshared environmental influences and personality differences in adult twins, *Journal of Personality and Social Psychology* **58**(1), 103–110.
- [2] Bouchard, T.J. & McGue, M. (1990). Genetic and rearing environmental influences on adult personality: an analysis of adopted twins reared apart, *Journal of Personality* **58**(1), 263–292.
- [3] Deater-Deckard, K., Pike, A., Petrill, S.A., Cutting, A.L., Hughes, C. & O'Connor, T.G. (2001). Nonshared environmental processes in socio-emotional development: an observational study of identical twin differences in the preschool period, *Developmental Science* **4**(2), F1–F6.
- [4] Maccoby, E.E. (2000). Parenting and its effects on children: on reading and misreading behavior genetics, *Annual Review of Psychology* **51**, 1–27.
- [5] McCartney, K., Harris, M.J. & Bernieri, F. (1990). Growing up and growing apart: a developmental meta-analysis of twin studies, *Psychological Bulletin* **107**(2), 226–237.
- [6] Reiss, D., Plomin, R., Hetherington, E.M., Howe, G.W., Rovine, M., Tryon, A. & Hagan, M.S. (1994). The separate worlds of teenage siblings: an introduction to the study of the nonshared environment and adolescent development, in *Separate Social Worlds of Siblings: The Impact of Nonshared Environment on Development*, E.M. Hetherington, D. Reiss & R. Plomin, eds, Lawrence Erlbaum, Hillsdale.
- [7] Reiss, D., Pedersen, N., Cederblad, M., Hansson, K., Lichtenstein, P., Neiderhiser, J. & Elthammer, O. (2001). Genetic probes of three theories of maternal adjustment: universal questions posed to a Swedish sample, *Family Process* **40**(3), 247–259.
- [8] Rowe, D.C., Woulbroun, J. & Gulley, B.L. (1994). Peers and friends as nonshared environmental influences, in *Separate Social Worlds of Siblings: The Impact of Nonshared Environment on Development*, E.M. Hetherington, D. Reiss & R. Plomin eds, Lawrence Erlbaum, Hillsdale.
- [9] Scarr, S. & Weinberg, R.A. (1983). The Minnesota adoption studies: genetic differences and malleability, *Child Development* **54**, 260–267.
- [10] Towers, H., Spotts, E., Neiderhiser, J., Hansson, K., Lichtenstein, P., Cederblad, M., Pedersen, N.L., Elthammer, O. & Reiss, D. (2001). Nonshared environmental associations between mothers and their children, in *Paper Presented at the Biennial Meeting of the Society for Research in Child Development*, Minneapolis.
- [11] Towers, H. (2003). Contributions of current family factors and life events to women's adjustment: nonshared environmental pathways, *Dissertation-Abstracts-International: Section-B: The Sciences and Engineering* **63**(12-B), 6123.
- [12] Turkheimer, E. & Waldron, M. (2000). Nonshared environment: a theoretical, methodological, and quantitative review, *Psychological Bulletin* **126**(1), 78–108.
- [13] van den Boom, D.C. (1994). The influence of temperament and mothering on attachment and exploration: an experimental manipulation of sensitive responsiveness among lower-class mothers with irritable infants, *Child Development* **65**(5), 1457–1477.

HILARY TOWERS AND JENAE M. NEIDERHISER

Normal Scores and Expected Order Statistics

CLIFFORD E. LUNNEBORG

Volume 3, pp. 1439–1441

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Normal Scores and Expected Order Statistics

The use of normal scores as a replacement for ranks has been advocated as a way of increasing **power** in rank-sum tests (see **Power Analysis for Categorical Methods**). The rationale is that power is compromised by the presence of tied sums in the permutation distribution (see **Permutation Based Inference**) of rank-based test statistics, for example, $1 + 6 = 2 + 5 = 3 + 4 = 7$. The likelihood of such tied sums is reduced when the ranks (integers) are replaced by scores drawn from a continuous distribution.

There are three normal score transformations in the nonparametric inference literature. For each, the raw scores are first transformed to ranks.

van der Waerden scores [7] are the best known of the normal score transformations. The ranks are replaced with quantiles of the standard normal random variable. In particular, the k th of n ranked score is transformed to the $q(k) = [k/(n+1)]$ quantile of the $N(0, 1)$ random variable. For example, for $k = 3$, $n = 10$, $q(k) = 3/11$. The corresponding van der Waerden score is that z score below which fall $3/11$ of the distribution of the standard normal. The van der Waerden scores can be obtained, of course, from tables of the standard normal. Most statistics packages, however, have normal distribution functions. For example, in the statistical language **R** (<http://www.R-project.org>) the *qnorm* function returns normal quantiles:

We can get the set of n van der Waerden scores this way:

```
> n <- 8
> qnorm((1 : n)/(n + 1))
[1] -1.22064035 -0.76470967
-0.43072730 -0.13971030 0.13971030
0.43072730
[7] 0.76470967 1.22064035
```

Expected normal order statistics were proposed by Fisher and Yates [3]. The k th smallest raw score among n is replaced by the expectation of the k th smallest score in a random sample of n observations from the standard normal distribution. Special

tables [3] or [6] often are used for the transformation but the function *normOrder* in the **R** package *SuppDists* makes the transformation somewhat more accessible.

```
> normOrder(8)
[1] -1.42363084 -0.85222503
-0.47281401 -0.15255065 0.15255065
0.47281401
[7] 0.85222503 1.42363084
```

Random normal scores can be used to extend further the support for the permutation distribution. Both van der Waerden and expected normal order statistics are symmetric about the expected value of the normal random variable. Thus, there are a number of equivalent sums involving pairs of the normal scores. By using a set of ordered randomly chosen normal scores to code the ranks, these sum equivalencies are eliminated. The rationale for this approach is further developed in [1]. In **R**, the function *rnorm* returns sampled normal scores. Thus, the command

```
> sort(rnorm(8))
[1] -1.287122544 -1.026961817
-0.149881236 -0.052743551 0.373083915
[6] 0.408619154 1.254177021 1.736467480
```

provides such a sorted set of random normal observations. A characteristic of this approach, of course, is that a second call to the function will produce a different set of random normal scores:

```
> sort(rnorm(8))
[1] -2.00484017 -0.88854417
-0.82976365 -0.75127395 -0.61653373
-0.50975136
[7] -0.14575532 0.57991888
```

An analysis that is carried out using the first set of scores will differ, perhaps substantially, from an analysis that is carried out using the second set. Is this a drawback? Perhaps, though it reminds us that whenever we replace the original observations, with ranks or with other scores, the outcome of our subsequent analysis will be dependent on our choice of recoding.

Example

The dependence of the analytic outcome on the coding of scores is illustrated by an example adopted from Loftus, G. R. & Loftus, E. F. (1988) *Essence of statistics*. New York: Knopf. Eight student volunteers watched a filmed simulated automobile accident. Then, the students read what they were told was a newspaper account of the accident, an account that contained misinformation about the accident. Four randomly selected students were told that the article was taken from the *New York Daily News*, the other four that it appeared in the *New York Times*. Finally, students were tested on their memory for the details of the accident. The substantive (alternative) hypothesis is that recall of the accident would be more influenced by the *Times* account than by one attributed to the *Daily News*. In consequence, the *Times* group was expected to make greater errors of recall than would the *Daily News* students. The reported number of errors were (4, 8, 10, 5) for the *Daily News* group and (9, 13, 15, 7) for the *Times* group.

Permutation tests (see **Permutation Based Inference**) are used here to test the null hypothesis of equal errors for the two attributions against the alternative. The test statistic used in each instance is the sum of the error score codes for the *Times* students. The P value for each test is the proportion of the $8!/(4!4!) = 70$ permutations of the error score codes for which the *Times* sum is as large or larger than for the observed randomization of cases. Table 1 shows the variability in the P value over the different codings of the error scores.

Table 1 P values for different codings of responses

Error score coding	P value
Raw scores	$5/70 = 0.07143$
Ranks	$7/70 = 0.10000$
van der Waerden scores	$6/70 = 0.08571$
Expected normal order statistics	$8/70 = 0.11430$
Random normals, set 1	$5/70 = 0.07143$
Random normals, set 2	$7/70 = 0.10000$

The raw scores permutation test is the **Pitman test** and the rank-based permutation test is the **Wilcoxon Mann-Whitney test**. The normal codings of the ranked errors for the remaining four permutation tests are the ones reported earlier. All six permutation tests were carried out using the **R** function `perm.test` in the `exactRankTests` package.

Summary

The indeterminacy of the random normal scores and the relative difficulty of determining expected normal order statistics have made van der Waerden scores the most common choice for normal scores. Asymptotic inference techniques for normal scores tests are outlined in [2]. More appropriate to small sample studies, however, are inferences based on the exact permutation distribution of the test statistic or a Monte Carlo approximation to that distribution. Such normal scores test are available, for example, in StatXact (<http://www.cytel.com>) and the general ideas are described in [4] and [5].

References

- [1] Bell, C.B. & Doksum, K.A. (1965). Some new distribution-free statistics, *The Annals of Mathematical Statistics* **36**, 203–214.
- [2] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [3] Fisher, R.A. & Yates, F. (1957). *Statistical Tables for Biological, Agricultural and Medical Research*, 5th Edition, Oliver & Boyd, Edinburgh.
- [4] Good, P. (2000). *Permutation Tests*, 2nd Edition, Springer, New York.
- [5] Lunneborg, C.E. (2000). *Data Analysis by Resampling*, Duxbury, Pacific Grove.
- [6] Owen, D.B. (1962). *Handbook of Statistical Tables*, Addison-Wesley, Reading.
- [7] van der Waerden, B.L. (1952). Order tests for the two-sample problem and their power, *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* **55**, 453–458.

CLIFFORD E. LUNNEBORG

Nuisance Variables

C. MITCHELL DAYTON

Volume 3, pp. 1441–1442

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Nuisance Variables

Nuisance variables are associated with variation in an outcome (dependent variable) that is extraneous to the effects of independent variables that are of primary interest to the researcher. In experimental comparisons among randomly formed treatment groups, the impact of nuisance variables is to increase experimental error and, thereby, decrease the likelihood that true differences among groups will be detected. Procedures that control nuisance variables, and thereby reduce experimental error, can be expected to increase the **power** for detecting group differences. In addition to randomization, there are three fundamental approaches to controlling for the effects of nuisance variables. First, cases may be selected to be similar for one or more nuisance variables (e.g., only 6-year-old girls with no diagnosed learning difficulties are included in a study). Second, statistical adjustments may be made using stratification (e.g., cases are stratified by sex and grade in school). Note that the most complete **stratification** occurs when cases are matched one-to-one prior to application of experimental treatments. Stratification, or **matching**, on more than a small number of nuisance variables is often not practical. Third, statistical adjustment can be made using regression procedures (e.g., academic ability is used as a covariate in **analysis of covariance**, ANCOVA). In the social sciences, **randomization** is considered a *sine qua non* for experimental research. Thus, even when selection, stratification, and/or covariance adjustment is used, randomization is still necessary for controlling unrecognized nuisance variables.

Although randomization to experimental treatment groups guarantees their equivalence *in expectation*, the techniques of selection, stratification, and/or covariance adjustment are still desirable in order to increase statistical power. For example, if cases are stratified by grade level in school, then differences among grades and the interaction between grades and experimental treatments are no longer confounded with experimental error. In practice, substantial increases in statistical power can be achieved by controlling nuisance variances. Similarly, when

ANCOVA is used, the variance due to experimental error is reduced in proportion to the squared correlation between the dependent variable and the covariate. Thus, the greatest benefit from statistical adjustment occurs when the nuisance variables are strongly related to the dependent variable. In effect, controlling nuisance variables can reduce the sample size required to achieve a desired level of statistical power when making experimental comparisons.

While controlling nuisance variables may enhance statistical power in nonexperimental studies, the major impetus for this control is that, in the absence of randomization, comparison groups cannot be assumed to be equivalent in expectation. Thus, in nonexperimental studies, the techniques of matching, stratification, and/or ANCOVA are utilized in an effort to control preexisting differences among comparison groups. Of course, complete control of nuisance variables is not possible without randomization. Thus, nonexperimental studies are always subject to threats to **internal validity** from unrecognized nuisance variables. For example, in a **case-control study**, controls may be selected to be similar to the cases and matching, stratification, and/or ANCOVA may be used, but there is no assurance that all relevant nuisance variables have been taken into account. Uncontrolled nuisance variables may be confounded with the effects of treatment in such a way that comparisons are biased.

For studies involving the prediction of an outcome rather than comparisons among groups (e.g., **multiple regression** and **logistic regression** analysis), the same general concepts that apply to nonexperimental studies are relevant. In regression analysis, the independent contribution of a specific predictor, say *X*, can be assessed in the context of other predictors in the model. These other predictors may include dummy-coded stratification variables and/or variables acting as covariates (see **Dummy Variables**). Then, a significance test for the regression coefficient corresponding to *X* will be adjusted for these other predictors. However, as in the case of any nonexperimental study, this assessment may be biased owing to unrecognized nuisance variables.

C. MITCHELL DAYTON

Number of Clusters

DAVID WISHART

Volume 3, pp. 1442–1446

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Number of Clusters

An issue that frequently occurs in cluster analysis is how many clusters are present, or which partition is ‘best’ (see **Cluster Analysis: Overview**)? This article discusses stopping rules that have been proposed for determining the best number of clusters, and concludes with a pragmatic approach to the problem for social science investigations.

In **hierarchical classification** the investigator is presented with a complete range of clusters, from one (the entire data set), to n , the number of cases (where each cluster has one member). In the early stages of agglomerative clustering, each cluster comprises either a single case or a small homogeneous grouping of cases that are very similar. At the later stages, clusters are combined to form larger – more heterogeneous – groups of cases, and at the penultimate stage there are only two clusters that constitute, in terms of the clustering method, the best two groups. Clusters formed at the later stages therefore exhibit increasingly less homogeneity or greater heterogeneity. Divisive clustering methods work in the opposite direction – at the first stage dividing the sample into the two most homogeneous subgroups, then further subdividing the subgroups, and so on until a stopping rule can be applied or each cluster contains a single case and cannot be further subdivided. The issue of which partition is ‘best’ therefore resolves in some sense as to how much heterogeneity can be tolerated in the latest cluster formed by agglomeration or split by division.

Numerous procedures have been proposed for determining the best number of clusters in a data set. In a review of 30 such rules, Milligan and Cooper [8] present a Monte Carlo evaluation of the performance of the rules when applied to the analysis of artificial data sets containing 2, 3, 4, or 5 distinct clusters by four hierarchical clustering methods. The procedures described below were found by Milligan and Cooper to outperform the others covered in their review.

Mojena [9] approached the question of when to stop a hierarchical agglomerative process by evaluating the distribution of clustering criterion values that form the clusters at each level. These values normally constitute a series that is increasing or decreasing with each fusion. The approach is to search for an unusually large jump in an increasing series, corresponding to a disproportionate loss of homogeneity

caused by the forced fusion of two disparate clusters, despite their being the most similar of all possible fusion choices.

For example, Ward’s method obtains a series of monotonically increasing fusion values $\alpha = \{\alpha_1, \dots, \alpha_{n-1}\}$ each of which is the smallest increase in the total within-cluster error sum of squares resulting from the fusion of two clusters. Such a series is shown in Table 1. It is argued that the ‘best’ partition is the one that precedes a significant jump in α , implying that the union of two clusters results in a large loss of homogeneity and should not therefore be combined. From visual inspection of Table 1 the four-cluster partition is a likely candidate for the best partition, because the increase in the error sums of squares, E , resulting from the next fusion to form 3 clusters (4.74) is more than three times the corresponding value (1.5) at which 4 clusters were formed. There is, therefore, a large loss of homogeneity at 3 clusters.

Mojena’s Upper Tail Rule

This rule is proposed by Mojena [9]. It looks for the first value α_{j+1} that lies in the upper tail of the distribution of fusion values α , selecting the partition corresponding to the first stage j satisfying:

$$\alpha_{j+1} > \mu_\alpha + c\sigma_\alpha \quad (1)$$

where μ_α is the mean and σ_α the standard deviation of the distribution of fusion values $\alpha = \{\alpha_1, \dots, \alpha_{n-1}\}$, and c is a chosen constant, sometimes referred to as the ‘critical score’. The rule states that $n - j$ clusters are best for the first α_{j+1} that satisfies the rule. If no value for α_{j+1} satisfies the rule then one cluster is best. When α_{j+1} satisfies the rule, this means that α_{j+1} exceeds the mean μ_α of α by at least c standard deviations σ_α . It is the first big jump in α , suggesting that partition j is reasonably homogeneous whereas partition $j + 1$ is not. Partition j is therefore selected as the first best partition.

Using Ward’s method, Mojena and Wishart [10] found values of $c \leq 2.5$ performed best, whereas Milligan and Cooper [8] suggest $c = 1.25$. The reason for the discrepancy between these findings is not clear, but it may lie in the choice of proximity measure: for Ward’s method it is necessary to use a proximity matrix of *squared* Euclidean distances, whereas Milligan and Cooper computed Euclidean

2 Number of Clusters

Table 1 Fusion values in the last 10 stages of hierarchical clustering by Ward’s method. Each value is the increase in error sum of squares, E , resulting from the next fusion, thus 28.821 is the increase in E caused by joining the last 2 clusters

Number of clusters	10	9	8	7	6	5	4	3	2
Fusion values α	0.457	0.457	0.903	1.172	1.228	1.500	4.740	7.303	28.821

distances. Having evaluated a range of values of c for different artificial data sets, Milligan and Cooper concluded that the best critical score c also varied with the numbers of clusters. As the number of clusters is not known *a priori*, this finding is not generally useful.

The series in Table 1 has a mean of 2 and standard deviation of 5.96, so rule (1) checks for the first fusion value that exceeds 9.45 for a critical score of $c = 1.25$, or higher values for higher critical score c . This is not very helpful, as the only partition for which α_{j+1} exceeds 9.45 is at 2 clusters. An alternative approach is to look for the first fusion value that represents a significant departure from the mean μ_α of α , using the t test statistic:

$$\frac{\alpha_{j+1} - \mu_\alpha}{\sigma_\alpha / \sqrt{n - 1}} \quad (2)$$

which has a t distribution with $n - 2$ degrees of freedom under normality. The first partition for which (2) is significant at the 5% probability level indicates a large loss of homogeneity, and hence the previous partition j is taken as best. This has the advantages that the test takes into account the length of the α series, which determines the degrees of freedom, and eliminates the need to specify a value for the critical score c . It is implemented by Wishart [11] in his ‘Best Cut’ Clustan procedure. For the data in Table 1, which were drawn from a fusion sequence of 24 values on 25 entities, the t statistics that are significant at the 5% level with 23 degrees of freedom are 22.05, 4.36, and 2.25 at the 2, 3, and 4 cluster partitions, respectively. The t test therefore confirms the earlier intuitive conclusion for Table 1 that 4 clusters is the best level of disaggregation for these data.

It should be noted that Mojena’s upper tail rule is only defined for a hierarchical classification: it cannot be used to compare two partitions obtained by optimization because it requires a series of fusion values. Mojena [9] also evaluated a moving average quality control rule, and Mojena and Wishart [10] evaluated a double exponential smoothing rule. However, both

of these performed poorly in tests although the latter rule proved to be most accurate in predicting the next value in the α series at each stage. They concluded that the upper tail rule performed best using Ward’s method with artificial data sets.

Caliński and Harabasz’ Variance Ratio

Caliński and Harabasz [4] propose a variance ratio test statistic:

$$C(k) = \frac{\text{trace}(\mathbf{B})/(k - 1)}{\text{trace}(\mathbf{W})/(n - k)} \quad (3)$$

where $\mathbf{W}$ is the within-groups dispersion matrix and $\mathbf{B}$ is the between-groups dispersion matrix for a partition of n cases into k clusters. Since $\text{trace}(\mathbf{W})$ is the within-groups sum of squares, and $\text{trace}(\mathbf{B})$ is the between-groups sum of squares, we have $\text{trace}(\mathbf{W}) + \text{trace}(\mathbf{B}) = \text{trace}(\mathbf{T})$, the total sum of squares, which is a constant. It follows that $C(k)$ is a weighted function of the within-groups error sum of squares $\text{trace}(\mathbf{W})$. Caliński and Harabasz suggest that the partition with maximum $C(k)$ is indicated as the best; if $C(k)$ increases monotonically with k then a best partition of the data may not exist.

This rule has the advantage that no critical score needs to be chosen. It can be used with both hierarchical classifications and partitioning methods, because it is separately evaluated for each partition of k clusters. Milligan and Cooper [8] found the variance ratio criterion to be the top performer in their tests. Everitt et al. [6] used $C(k)$ to evaluate different partitions of garden flower data, for which 4 clusters were selected.

Beale’s F test

Beale [3] proposes a test statistic on the basis of the sum of squared distances between the cases and the cluster means. It is an F test, defined for two

partitions into k_1 and k_2 clusters ($k_2 > k_1$):

$$F(k_1, k_2) = \frac{(S_1^2 - S_2^2)/S_2^2}{[(n - k_1)/(n - k_2)](k_2/k_1)^{2/n} - 1} \quad (4)$$

where S_1^2 and S_2^2 are the sum of squared distances for clusters 1 and 2. The statistic is compared with an F distribution with $(k_2 - k_1)$ and $(n - k_2)$ degrees of freedom. If the F -ratio is significant for any k_2 then Beale concludes that the partition k_1 is not entirely adequate.

Beale's F test was proposed for the comparison of two k -means partitions generally, but it can also be used with hierarchical methods for all $(k_j, j = n, \dots, 2)$ in a dendrogram (see **k -means Analysis**). In hierarchical agglomerative clustering, for example, the rule stops at the first partition j for which $F(k_{j+1}, k_j)$ is significant with 1 and $(n - k_j)$ degrees of freedom. This rule has the advantage that no critical score needs to be chosen.

Duda and Hart's Error Ratio

Duda and Hart [5] propose a local criterion that evaluates the division of a cluster m into two subclusters, without reference to the remainder of the data set:

$$\frac{\text{Je}(2)}{\text{Je}(1)} \quad (5)$$

where $\text{Je}(1)$ is the sum of the squared distances of the cases from the mean of cluster m , and $\text{Je}(2)$ is the sum of the squared distances of the cases from the means of the two subclusters following division of cluster m . The null hypothesis that cluster m is homogeneous, and hence the cluster should be subdivided, is rejected if:

$$L(m) = \{1 - \text{Je}(2)/\text{Je}(1) - 2/(\pi p)\} \times \{n_m p / (2[1 - 8/(\pi^2 p)])\}^{1/2} \quad (6)$$

exceeds the critical value from a standard normal distribution, where p is the number of variables and n_m is the number of cases in cluster m . The rule can be applied to hierarchical agglomerative clustering, when the fusion of two clusters to form cluster m is being considered. Although it is local to the consideration of cluster m alone, the rule can be applied globally to any hierarchical agglomerative

or divisive procedures by considering all the test statistics $\{L(m), m = 1, \dots, k\}$ in a series of fusions or divisions, and stopping when the test statistic exceeds its critical value.

Milligan and Cooper [8] conclude that the error ratio performed best with a critical value of 3.3 in their tests.

Hubert and Levin's C Index

For a given partition into k clusters, Hubert and Levin [7] compute the sum D_k of the within-cluster dissimilarities for a series or set of partitions of the same data. The minimum $D_{\min}$ and maximum $D_{\max}$ of D_k are found, and their C -index is a standardized D_k :

$$C = \frac{D_k - D_{\min}}{D_{\max} - D_{\min}} \quad (7)$$

The minimum value of C across all partitions evaluated is taken to be the best.

Baker and Hubert's Gamma Index

Given the input dissimilarities d_{ij} and output ultrametric distances h_{ij} corresponding to a hierarchical classification, Baker and Hubert [2] define a Gamma index:

$$\gamma = \frac{S_+ - S_-}{S_+ + S_-} \quad (8)$$

where S_+ denotes the number of subscript combinations (i, j, k, l) for which $g_{ijkl} \equiv (d_{ij} - d_{kl})(h_{ij} - h_{kl})$ is positive, and S_- is the number that are negative. It is a measure of the concordance of the classification: S_+ is the number of consistent comparisons of within-cluster and between-cluster dissimilarities, and S_- is the number of inconsistent comparisons. When $\gamma = 1$, the correspondence is perfect. It is evaluated at all partitions and the maximum value for γ indicates the best partition (see **Measures of Association**).

It should be noted that γ is only defined for a classification obtained by the transformation of a proximity matrix into a dendrogram. It cannot, therefore, be used with optimization methods.

Wishart's Bootstrap Validation

Wishart [11] compares a dendrogram obtained for a given data set with a series of dendrograms generated

at random from the same data set or proximity matrix (see **Proximity Measures**). The method searches for partitions that manifest the greatest departure from randomness. In statistical terms, it seeks to reject the null hypothesis that the structure displayed by a partition is random.

The data are rearranged at random within columns (variables) and a dendrogram is obtained from the randomized data. This randomization is then repeated in a series of r random trials, where each trial generates a different dendrogram for the data arranged in a different random order. The mean μ_j and standard deviation σ_j are computed for the fusion values obtained at each partition j in r trials, and a confidence interval for the resulting family of dendrograms is obtained. The dendrogram for the given data, before randomization, is then compared with the distribution of dendrograms obtained with the randomized data, looking for significant departures from randomness. The null hypothesis that the given data are random and exhibit no structure is evaluated by a t test on the fusion value α_j at each partition j compared with the distribution of randomized dendrograms:

$$\frac{\alpha_j - \mu_j}{\sigma_j / \sqrt{r - 1}} \quad (9)$$

which is compared with the t distribution with $r - 2$ degrees of freedom. The best partition is taken to be the one that exhibits the greatest departure from randomness, if any, as indicated by the largest value for the t statistic (9) that is significant at the 5% probability level. This rule can be used with hierarchical agglomerative and partitioning clustering methods.

Summary

A classification is a summary of the data. As such, it is only of value if the clusters make sense and the resulting typology can be usefully applied. They can make sense at different levels of disaggregation – biological organisms are conventionally classified as families, genera and species, and can be further subdivided as subspecies or hybrids. All of these make sense, either in the context of a broad-brush phylogeny or more narrowly in species-specific terms. The same can apply in the behavioral sciences, where people can be classified in an infinite variety of ways. We should never forget that every person is a unique individual. However, it is the stuff of social science

to look for key structure and common features in an otherwise infinitely heterogeneous population. The analysis does not necessarily need to reveal a satisfactory typology for all of the data – it may be enough, for example, to find a cluster of individuals who are sufficiently distinctive to warrant further study.

This article has presented various tests for the number of clusters in a data set, to satisfy an expressed need to justify the choice of one partition from several alternatives. Some editors of applied journals regrettably demand proof of ‘significance’ before accepting cluster analysis results for publication. In noting that there is a preoccupation with finding an ‘optimal’ number of groups, and a single ‘best’ classification, Anderberg [1] observes that the possibility of several alternative classifications reflecting different aspects of the data is rarely entertained.

A more pragmatic approach is to select the partition into the largest number of clusters that makes sense as a workable summary of the data and can be adequately described. This constitutes a first baseline of greatest disaggregation. Higher-order partitions may then be reviewed with reference to this baseline, where more simplification is indicated. Clusters of particular interest can be drawn out for special attention. If appropriate, a hierarchy of the relationships between the baseline clusters can be presented, and it may be helpful to order this hierarchy optimally so that higher-order clusters are grouped in a meaningful way. A ‘best’ partition may then be drawn out as a level of special interest in the hierarchy, if indeed it does constitute a sensible best summary. It may be helpful to recall the parallel of a biological classification – at the family, genus, and species levels of detail. In practice, the presentation of a classification is no different from writing intelligently and purposefully on any subject, where the degree of technical detail or broad summary depends crucially on what the author aims to communicate to his intended audience. In the final analysis, the crucial test is whether or not the presentation is communicated successfully to the reader and survives peer review and the passage of time.

References

- [1] Anderberg, M.R. (1973). *Cluster Analysis for Applications*, Academic Press, London.
- [2] Baker, F.B. & Hubert, L.J. (1975). Measuring the power of hierarchical cluster analysis, *Journal American Statistical Association* **70**, 31–38.

-
- [3] Beale, E.M.L. (1969). Euclidean cluster analysis, *Proceedings ISA* **43**, 92–94.
 - [4] Caliński, T. & Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics* **3**, 1–27.
 - [5] Duda, R.O. & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*, Wiley, New York.
 - [6] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th Edition, Edward Arnold, London.
 - [7] Hubert, L.J. & Levin, J.R. (1976). A general statistical framework for assessing categorical clustering in free recall, *Psychological Bulletin* **83**, 1072–1080.
 - [8] Milligan, G.W. & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika* **50**, 159–179.
 - [9] Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation, *Computer Journal* **20**, 359–363.
 - [10] Mojena, R. & Wishart, D. (1980). Stopping rules for Ward's clustering method, in *COMPSTAT 1980: Proceedings in Computational Statistics*, M.M. Barritt & D. Wishart, eds, Physica-Verlag, Wien, pp. 426–432.
 - [11] Wishart, D. (2004). *ClustanGraphics Primer: A Guide to Cluster Analysis*, Clustan, Edinburgh.

DAVID WISHART

Number of Matches and Magnitude of Correlation

MICHAEL J. ROVINE

Volume 3, pp. 1446–1448

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Number of Matches and Magnitude of Correlation

The first expression of a correlation coefficient for categorical variables is often attributed to **G. Udney Yule**. For two dichotomous variables, x and y , the **2 × 2 contingency table** (see Table 1).

Yule's Q was defined as

$$Q = \frac{ad - bc}{ad + bc}. \quad (1)$$

The coefficient functions as an index of the number of observations that occur in the main diagonal of the 2 × 2 table. According to Yule [9], this coefficient has the defining characteristics of a proper correlation coefficient: namely, its maximum values are 1 and -1, and 0 represents no association between the two variables. Coming out of the tradition of Pearson bivariate normal notion of the correlation coefficient, Yule interpreted the coefficient as the size of the association between two random variables much as he would interpret a **Pearson product moment correlation coefficient**. Pearson [2] emphasized this definition as he described a set of alternative versions of Q ($Q_1, Q_3 - Q_5$ renaming Yule's Q, Q_2). Pearson's coefficients were more easily seen as the association of a pair of dichotomized normal variables.

Although these were the first coefficients that were described using the word, 'correlation', they were preceded by a set of association coefficients (see **Measures of Association**) developed to estimate the association between the prediction of an event and the occurrence of an event. In an 1884 science paper, the American logician, C. S. Peirce, defined a coefficient of 'the science of the method' that he used to assess the accuracy of tornado prediction [5]. His coefficient, i , was defined as

$$i = \frac{ad - bc}{(a + c)(b + d)} \quad (2)$$

like Yule's Q and is interesting. The numerator in Peirce's i is the determinant of a 2 × 2 matrix. This numerator is common to most association coefficients for dichotomous variables. The denominator is a scaling factor that limits the maximum absolute value of the coefficient to 1. For Peirce's i , the denominator is the product of the marginals. It is interesting to compare Peirce's i to the φ -coefficient, the computational form of Pearson's product-moment coefficient:

$$\varphi = \frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}}. \quad (3)$$

For equal marginal values (i.e., $(a + b) = (a + c)$ and $(b + d) = (c + d)$), the two formulas are identical. Where i is an asymmetric coefficient, φ is a symmetric coefficient.

As an association coefficient, the correlation coefficient had a number of different interpretations [3]. The work of **Spearman** [7] on the use of the correlation coefficient to assess reliability led to a number of different forms and additional interpretations of the coefficient. Spearman and Thorndike [8] presented papers that brought the correlation coefficient into the area of measurement error. Thorndike suggested the use of a median-ratio coefficient as having a more direct relation to each pair of variable values rather than considering the table as a whole. **Cohen** [1] suggested an association coefficient based on the idea of inter-rater agreement (see **Rater Agreement – Kappa**). His κ coefficient defined agreement above and beyond chance. For the 2 × 2 table,

$$\kappa = \frac{(a + d) - (a_e + d_e)}{N - (a_e + d_e)} \quad (4)$$

where a_e and d_e are the cell values expected by chance.

An indicator of effect size (see **Effect Size Measures**) more directly interpreted in terms of each pair of variable values has been presented by Rosenthal and Rubin [4]. Their **Binomial Effect Size Display (BESD)** was used to relate a treatment to an outcome. In a 2 × 2 table, the BESD indicates the increase in success rate of the treatment over the outcome. An example appears in Table 2. The correlation coefficient, $r = 0.60$ compares the proportions of those in the treatment group with good outcomes to those

Table 1 A 2 × 2 contingency table

	x		Total
	1	0	
Y	1	a b	$a + b$
	0	c d	$c + d$
Total	$a + c$	$b + d$	N

2 Number of Matches and Magnitude of Correlation

in the control group. It, thus, represents the increase in proportion of positive outcomes expected as one moves from the control group to the treatment group. Rosenthal and Rubin [4] extended the BESD to a continuous outcome variable.

Rovine and von Eye [6] approached the correlation coefficient in a similar fashion, considering the correlation coefficient in the case of two continuous variables. They considered the proportion of matches in a bivariate distribution. They began with 100 observations of two uncorrelated variables, x and y , each drawn from a normal distribution. They created one match by replacing the first observation of y with the first observation of x ; two matches by replacing the first two observations of y with the first two observations of x ; and so on. In this manner, they created 100 new data sets with 1 to 100 matches. They calculated the correlation for each of these data sets. The correlation approximately equaled the proportion of matches in the data set. They extended this definition by defining a match for two standardized variables as the instance in which the values of the y falls within a specific interval, $[-A, A]$ of the value of x . In this case, matches do not have to be exact. They can be targeted so that a match can represent values on two variables falling in the same general location in each distribution as defined by an interval around the value of the variable, x . For matches in which the values for variable x is a , the interval would be $a \pm A$. If the value b on variable y fell within that range, a match would occur. For k matches in n observations where a match is defined as described above, the correlation coefficient one would observe is

$$r \approx \frac{\frac{k}{n}}{\left(1 + \frac{k}{n} \frac{A^2}{3}\right)^{1/2}}. \quad (5)$$

According to this equation, increasing the interval in which a match occurs decreases the size of

the correlation one could expect to observe. If the interval, A is 0, then a match represents an exact match, and the correlation represents the proportion of matches. As the interval that defines a match increases, it would allow unrelated values of y to more likely qualify as a match. As a result, the observed correlation would tend to be smaller.

This result yields a way of interpreting the correlation coefficient that relates back to the notion of a contingency table. For categorical variables, the observed frequencies in a contingency can be used to help interpret the meaning of a correlation coefficient. For continuous variables, the idea of a match allows the investigator to consider the relationship between two continuous variables in a similar fashion. The correlation as a proportion of matches represents another way of interpreting what a correlation coefficient of a particular size can mean.

References

- [1] Cohen, J.J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1), 37–46.
- [2] Pearson, K. (1901). On the correlation of characters not quantitatively measurable, *Philosophical Transactions of the Royal Society of London, Series A* **195**, 1–47.
- [3] Rodgers, J.L. & Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient, *American Statistician* **42**(1), 59–66.
- [4] Rosenthal, R. & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect, *Journal of Educational Psychology* **74**(2), 166–169.
- [5] Rovine, M.J. & Anderson, D.A. (2004). Peirce and Bowditch: an American contribution to correlation and regression, *The American Statistician* **59**(3), 232–236.
- [6] Rovine, M.J. & von Eye, A. (1997). A 14th way to look at a correlation coefficient: correlation as a proportion of matches, *The American Statistician* **51**(1), 42–46.
- [7] Spearman, C. (1907). Demonstration of formulae for true measurement of correlation, *American Journal of Psychology* **XVIII**, 160–169.
- [8] Thorndike, E.L. (1907). Empirical studies in the theory of measurement, New York.
- [9] Yule, G.U. (1900). On the association of attributes in statistics: with illustrations from the material from the childhood society, and so on, *Philosophical Transactions of the Royal Society of London, Series A* **194**, 257–319.

Further Reading

Peirce, C.S. (1884). The numerical measure of the success of predictions, *Science* **4**, 453–454.

MICHAEL J. ROVINE

Table 2 A BESD coefficient

	Outcome		Proportion
	Good	Bad	
Treatment y	40	10	0.80
Control	10	40	0.20

$$\text{BESD} = 0.80 - 0.20 = 0.60.$$

Number Needed to Treat

EMMANUEL LESAFFRE

Volume 3, pp. 1448–1450

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Number Needed to Treat

The Number Needed to Treat and Related Measures

Suppose that in a randomized controlled **clinical trial** two treatments are administered and that the probabilities of showing a positive response are π_1 (new treatment) and π_2 (control treatment), respectively. The absolute risk reduction (AR) is defined as $\pi_1 - \pi_2$ and expresses in absolute terms the benefit (or the harm) of the new over the control treatment. For example, when $\pi_1 = 0.10$ and $\pi_2 = 0.05$, then in 100 patients there are on average five more responders in the new treatment group. Other measures to indicate the benefit of the new treatment are the **relative risk (RR)**, the relative risk reduction (RRR) and the **odds-ratio (OR)**. On the other hand, Laupacis et al.[3] suggested the number needed to treat (NNT) as a measure to express the benefit of the new treatment. The NNT is defined as the *inverse of the absolute risk reduction*, that is, $NNT = 1/AR$. In our example, NNT is equal to 20 and is interpreted that 20 patients need to be treated with the new treatment to ‘save’ one extra patient if these patients were treated with the control treatment.

In practice, the true response rates π_1 and π_2 are estimated by $p_j = r_j/n_j$ ($j = 1, 2$), respectively where n_j ($j = 1, 2$) are the number of patients in the new and control treatment, respectively with r_j ($j = 1, 2$) responders. Replacing the true response rates by their estimates in the expressions above yields an estimated value for AR, NNT, RR, and OR, denoted by $\widehat{ar}$, $\widehat{NNT}$, $\widehat{rr}$ and $\widehat{or}$, respectively. However, in the literature the notation NNT is used most often for the estimated number needed to treat.

A Clinical Example

In a double-blind, placebo-controlled study, 23 epileptic patients were randomly assigned to topiramate 400 mg/day treatment (as adjunctive therapy) and 24 patients to control treatment [4]. It is customary to measure the effect of an antiepileptic drug by the percentage of patients who show at least 50% reduction in seizure rate at the end of the treatment period. Two patients on placebo showed at least 50% reduction in seizure rate, compared

to eight patients on topiramate ($p = 0.036$, Fisher’s Exact test). The proportion of patients with at least a 50% reduction is $p_2 = 2/24 = 0.083$ for placebo and $p_1 = 8/23 = 0.35$ for topiramate. The absolute risk reduction $ar = p_1 - p_2 = 0.26$ expresses the absolute gain due to the administration of the active (add-on) medication on epileptic patients. The odds-ratio $(0.35/0.65)/(0.083/0.917) = 5.87$ is more popular in epidemiology and has attractive statistical properties [2], but is difficult to interpret by clinicians. With the relative risk = 4.17, we estimate that the probability of being a 50% responder is about 4 times as high under topiramate as under placebo. The estimated NNT equals $1/0.26 = 3.78$, which is interpreted that about four patients must receive topiramate if we are to achieve one additional good outcome compared to placebo.

Controversies on the Number Needed to Treat

Proponents of the NNT (often, but not exclusively, clinical researchers) argue that this measure is much better understood by clinicians than the odds-ratio. Further, they argue that relative measures like or and rr cannot express the benefit of the new over the control treatment in daily life. Initially, some false claims were made such as that the NNT can give a clearer picture of the difference between treatment results, see for example [2] for an illustration.

Antagonists of the NNT state that it is not clear whether clinicians indeed understand this measure better than the other measures. Further, they point out that clinicians neglect the statistical uncertainty with which NNT is estimated in practice. The statistical properties of $\widehat{NNT}$ have been investigated by Lesaffre and Pledger [4]. They indicate the following problems: (a) when $ar = 0$, $\widehat{NNT}$ becomes infinite; (b) the distribution of $\widehat{NNT}$ is complicated because its behavior around $ar = 0$; (c) the moments of $\widehat{NNT}$ do not exist; and (d) simple calculations with $\widehat{NNT}$ like addition can give nonsensical results. To take the statistical uncertainty of estimating NNT into account, one could calculate the $100(1 - \alpha)\%$ confidence interval (CI) for NNT on the basis of $\widehat{NNT}$. For the above example, the 95% C.I. is equal to [2.01, 24.6]. However, when the $100(1 - \alpha)\%$ C.I. for AR, $[ar_{un}, ar_{up}]$, includes 0 the $100(1 - \alpha)\%$ C.I.

for NNT splits up into two disjoint intervals $] - \infty, 1/ar_{up}] \cup [1/ar_{un}, \infty[$.

There is also the risk that $\widehat{NNT}$ is only reported when $H_0 : AR = 0$ is rejected, or when $\widehat{NNT}$ is positive. Altman [1] suggested using the term number needed to treat for harm (NNTH) when $\widehat{NNT} < 0$, implying that it would represent the (average) number of patients needed to treat with the new treatment to harm one extra patient than when these patients were treated with the control treatment. The NNTH has been used for reporting adverse events in a clinical trial. When $\widehat{NNT} > 0$, Altman called it the *number needed to treat for benefit* (NNTB).

The Use of the Number Needed to Treat in Evidence-based Medicine

Lesaffre and Pledger [4] showed that the number needed to treat deduced from a **meta-analysis** should definitely not be calculated on the NNT-scale, but on the AR-scale if that scale shows homogeneity. However, there is often heterogeneity on the AR-scale and hence also for the NNT, because they both depend on the baseline risk. This also implies that one needs to be very careful when extrapolating the estimated NNT to new patient populations. On the other hand, there is more homogeneity on the OR- and the RR-scale. Suppose that the meta-analysis has been done on the relative risk scale then, because $NNT = 1/(1 - RR)\pi_2$, one can estimate the NNT for a patient with baseline risk $k\pi_2$ by $k\widehat{NNT}$, where $\widehat{NNT}$ is obtained from the meta-analysis. This result is the basis for developing nomograms for individual patients with risks different from the risks observed in the meta-analysis.

Furthermore, the literature clearly shows that NNT has penetrated in all therapeutic areas and even more importantly is a key tool for inference in Evidence-based Medicine and even so for the authorities.

Conclusion

Despite the controversy between clinicians and statisticians and the statistical deficiencies of NNT, the number needed to treat is becoming increasingly popular in Evidence-based Medicine and is well appreciated by the authorities. In a certain sense, this popularity is surprising since $100AR$ is at least as easy to interpret; namely, it is the average number of extra patients that benefit from the new treatment (over the control treatment) when 100 patients were treated. Moreover, ar has better statistical properties than $\widehat{NNT}$.

References

- [1] Altman, D.G. (1998). Confidence intervals for the number needed to treat, *British Medical Journal* **317**, 1309–1312.
- [2] Hutton, J. (2000). Number needed to treat: properties and problems, *Journal of the Royal Statistical Society, Series A* **163**, 403–419.
- [3] Laupacis, A., Sackett, D.L. & Roberts, R.S. (1988). An assessment of clinically useful measures of the consequences of treatment, *New England Journal of Medicine* **318**, 1728–1733.
- [4] Lesaffre, E. & Pledger, G. (1999). A note on the number needed to treat, *Controlled Clinical Trials* **20**, 439–447.

Further Reading

Chatellier, G., Zapletal, E., Lemaitre, D., Menard, J. & Degoulet, P. (1996). The number of needed to treat: a clinically useful nomogram in its proper context, *British Medical Journal* **312**, 426–429.

EMMANUEL LESAFFRE

Observational Study

PAUL R. ROSENBAUM

Volume 3, pp. 1451–1462

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Observational Study

Observational Studies Defined

In the ideal, the effects caused by treatments are investigated in experiments that randomly assign subjects to treatment or control, thereby ensuring that comparable groups are compared under competing treatments [1, 5, 15, 23]. In such an experiment, comparable groups prior to treatment ensure that differences in outcomes after treatment reflect effects of the treatment (*see Clinical Trials and Intervention Studies*). Random assignment uses chance to form comparable groups; it does not use measured characteristics describing the subjects before treatment. As a consequence, random assignment tends to make the groups comparable both in terms of measured characteristics and characteristics that were not or could not be measured. It is the unmeasured characteristics that present the largest difficulties when **randomization** is not used. More precisely, random assignment ensures that the only differences between treated and control groups prior to treatment are due to chance – the flip of a coin in assigning one subject to treatment, another to control – so if a common statistical test rejects the hypothesis that the difference is due to chance, then a treatment effect is demonstrated [18, 22].

Experiments with human subjects are often ethical and feasible when (a) all of the competing treatments under study are either harmless or intended and expected to benefit the recipients, (b) the best treatment is not known, and in light of this, subjects consent to be randomized, and (c) the investigator can control the assignment and delivery of treatments. Experiments cannot ethically be used to study treatments that are harmful or unwanted, and experiments are not practical when subjects refuse to cede control of treatment assignment to the experimenter. When experiments are not ethical or not feasible, the effects of treatments are examined in an observational study. Cochran [12] defined an *observational study* as an empiric comparison of treated and control groups in which:

the objective is to elucidate cause-and-effect relationships [... in which it] is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects

it is desired to discover, or to assign subjects at random to different procedures.

When subjects are not assigned to treatment or control at random, when subjects select their own treatments or their environments inflict treatments upon them, differing outcomes may reflect these initial differences rather than effects of the treatments [12, 59]. Pretreatment differences or *selection biases* are of two kinds: those that have been accurately measured, called *overt biases*, and those that have not be measured but are suspected to exist, called *hidden biases*. Removing overt biases and addressing uncertainty about hidden biases are central issues in observational studies. Overt biases are removed by adjustments, such as **matching**, **stratification** or covariance adjustment (*see Analysis of Covariance*), which are discussed in the Section titled ‘Adjusting for Biases Visible in Observed Covariates’. Hidden biases are addressed partly in the design of an observational study, discussed in the Section titled ‘Design of Observational Studies’ and ‘Elaborate Theories’, and partly in the analysis of the study, discussed in the Section titled ‘Appraising Sensitivity to Hidden Bias’ and ‘Elaborate Theories’ (*see Quasi-experimental Designs; Internal Validity; and External Validity*).

Examples of Observational Studies

Several examples of observational studies are described. Later sections refer to these examples.

Long-term Psychological Effects of the Death of a Close Relative

In an attempt to estimate the long-term psychological effects of bereavement, Lehman, Wortman, and Williams [30] collected data following the sudden death of a spouse or a child in a car crash. They matched 80 bereaved spouses and parents to 80 controls drawn from 7581 individuals who came to renew a drivers license. Specifically, they matched for gender, age, family income before the crash, education level, number, and ages of children. Contrasting their findings with the views of Bowlby and Freud, they concluded:

Contrary to what some early writers have suggested about the duration of the major symptoms

2 Observational Study

of bereavement . . . both spouses and parents in our study showed clear evidence of depression and lack of resolution at the time of the interview, which was 5 to 7 years after the loss occurred.

Effects on Criminal Violence of Laws Limiting Access to Handguns

Do laws that ban purchases of handguns by convicted felons reduce criminal violence? It would be difficult, perhaps impossible, to study this in a randomized experiment, and yet an observational study faces substantial difficulties as well. One could not reasonably estimate the effects of such a law by comparing the rate of criminal violence among convicted felons barred from handgun purchases to the rate among all other individuals permitted to purchase handguns. After all, convicted felons may be more prone to criminal violence and may have greater access to illegally purchased guns than typical purchasers of handguns without felony convictions. For instance, in his ethnographic account of violent criminals, Athens (1997, p. 68) depicts their sporadic violent behavior as consistent with stable patterns of thought and interaction, and writes: “. . . the self-images of violent criminals are always congruent with their violent criminal actions.”

Wright, Wintemute, and Rivara [68] compared two groups of individuals in California: (a) individuals who attempted to purchase a handgun but whose purchase was denied because of a prior felony conviction, and (b) individuals whose purchase was approved because their prior felony arrest had not resulted in a conviction. The comparison looked forward in time from the attempt to purchase a handgun, recording arrest charges for new offenses in the subsequent three years. Presumably, group (b) is a mixture of some individuals who did not commit the felony for which they were arrested and others who did. If this presumption were correct, group (a) would be more similar to group (b) than to typical purchasers of handguns, but substantial biases may remain.

Effects on Children of Occupational Exposures to Lead

Morton, Saah, Silberg, Owens, Roberts, and Saah [39] asked whether children were harmed by lead brought home in the clothes and hair of parents

who were exposed to lead at work. They matched 33 children whose parents worked in a battery factory to 33 unexposed control children of the same age and neighborhood, and used Wilcoxon's signed rank test to compare the level of lead found in the children's blood, finding elevated levels of lead in exposed children.

In addition, they compared exposed children whose parents had varied levels of exposure to lead at the factory, finding that parents who had higher exposures on the job in turn had children with more lead in their blood. Finally, they compared exposed children whose parents had varied hygiene upon leaving the factory at the end of the day, finding that poor hygiene of the parent predicted higher levels of lead in the blood of the child.

Design of Observational Studies

Observational studies are sometimes referred to as *natural experiments* [36, 56] or as *quasi-experiments* [61] (see **Quasi-experimental Designs**). These differences in terminology reflect certain differences in emphasis, but a shared theme is that the early stages of planning or designing an observational study attempt to reproduce, as nearly as possible, some of the strengths of an experiment [47].

A *treatment* is a program, policy, or intervention which, in principle, may be applied to or withheld from any subject under study. A variable measured prior to treatment is not affected by the treatment and is called a *covariate*. A variable measured after treatment may have been affected by the treatment and is called an *outcome*. An analysis that does not carefully distinguish covariates and outcomes can introduce biases into the analysis where none existed previously [43]. The *effect caused by a treatment* is a comparison of the outcome a subject exhibited under the treatment the subject actually received with the potential but unobserved outcome the subject would have exhibited under the alternative treatment [40, 59]. Causal effects so defined are sometimes said to be *counterfactual* (see **Counterfactual Reasoning**), in the specific sense that they contrast what did happen to a subject under one treatment with what would have happened under the other treatment. Causal effects cannot be calculated for individuals, because each individual is observed under treatment or under control, but not both. However, in a randomized experiment, the treated-minus-control difference

in mean outcomes is an unbiased and consistent estimate of the average effect of the treatment on the subjects in the experiment.

In planning an observational study, one attempts to identify circumstances in which some or all of the following elements are available [47].

- *Key covariates and outcomes are available for treated and control groups.* The most basic elements of an observational study are treated and control groups, with important covariates measured before treatment, and outcomes measured after treatment. If data are carefully collected over time as events occur, as in a *longitudinal study* (see **Longitudinal Data Analysis**), then the temporal order of events is typically clear, and the distinction between covariates and outcomes is clear as well. In contrast, if data are collected from subjects at a single time, as in a *cross-sectional study* (see **Cross-sectional Design**) based on a single survey interview, then the distinction between covariates and outcomes depends critically on the subjects' recall, and may not be sharp for some variables; this is a weakness of cross-sectional studies. Age and sex are covariates whenever they are measured, but current recall of past diseases, experiences, moods, habits, and so forth can easily be affected by subsequent events.
- *Haphazard treatment assignment rather than self-selection.* When randomization is not used, treated and control groups are often formed by deliberate choices reflecting either the personal preferences of the subjects themselves or else the view of some provider of treatment that certain subjects would benefit from treatment. Deliberate selection of this sort can lead to substantial biases in observational studies. For instance, Campbell and Boruch [10] discuss the substantial systematic biases in many observational studies of compensatory programs intended to offset some disadvantage, such as the US Head Start Program for preschool children. Campbell and Boruch note that the typical study compares disadvantaged subjects eligible for the program to controls who were not eligible because they were not sufficiently disadvantaged. When randomization is not possible, one should try to identify circumstances in which an ostensibly irrelevant event, rather than deliberate choice, assigned subjects to treatment or control. For instance, in the United States, class sizes in government run schools are largely determined by the degree of wealth in the local region, but in Israel, a rule proposed by Maimonides in the 12th century still requires that a class of 41 must be divided into two separate classes. In Israel, what separates a class of size 40 from classes half as large is the enrollment of one more student. Angrist and Lavy [2] exploited Maimonides rule in their study of the effects of class size on academic achievement in Israel. Similarly, Oreopoulos [41] studies the economic effects of living in a poor neighborhood by exploiting the policy of Toronto's public housing program of assigning people to housing in quite different neighborhoods simply based on their position in a waiting list. Lehman et al. [30], in their study of bereavement in the Section titled 'Long-term Psychological Effects of the Death of a Close Relative', limited the study to car crashes for which the driver was not responsible, on the grounds that car crashes for which the driver was responsible were relatively less haphazard events, perhaps reflecting forms of addiction or psychopathology. Random assignment is a fact, but haphazard assignment is a judgment, perhaps a mistaken one; however, haphazard assignments are preferred to assignments known to be severely biased.
- *Special populations offering reduced self-selection.* Restriction to certain subpopulations may diminish, albeit not eliminate, biases due to self-selection. In their study of the effects of adolescent abortion, Zabin, Hirsch, and Emerson [69] used as controls young women who visited a clinic for a pregnancy test, but whose test result came back negative, thereby ensuring that the controls were also sexually active. The use in the Section titled 'Effects on Criminal Violence of Laws Limiting Access to Handguns' of controls who had felony arrests without convictions may also reduce hidden bias.
- *Biases of known direction.* In some settings, the direction of unobserved biases is quite clear even if their magnitude is not, and in certain special circumstances, a treatment effect that overcomes a bias working against it may yield a relatively unambiguous conclusion. For

instance, in the Section titled ‘Effects on Criminal Violence of Laws Limiting Access to Handguns’, one expects that the group of convicted felons denied handguns contains fewer innocent individuals than does the arrested-but-not-convicted group who were permitted to purchase handguns. Nonetheless, Wright et al. [68] found fewer subsequent arrests for gun and violent offenses among the convicted felons, suggesting that the denial of handguns may have had an effect large enough to overcome a bias working in the opposite direction. Similarly, it is often claimed that payments from disability insurance provided by US Social Security deter recipients from returning to work by providing a financial disincentive. Bound [6] examined this claim by comparing disability recipients to rejected applicants, where the rejection was based on an administrative judgment that the injury or disability was not sufficiently severe. Here, too, the direction of bias seems clear: rejected applicants should be healthier. However, Bound found that even among the rejected applicants, relatively few returned to work, suggesting that even fewer of the recipients would return to work without insurance. Some general theory about studies that exploit biases of known direction is given in Section 6.5 of [49].

- *An abrupt start to intense treatments.* In an experiment, the treated and control conditions are markedly distinct, and these conditions become active at a specific known time. Lehman et al.’s [30] study of the psychological effects of bereavement in the Section titled ‘Long-term Psychological Effects of the Death of a Close Relative’ resembles an experiment in this sense. The study concerned the effects of the sudden loss of a spouse or a child in a car crash. In contrast, the loss of a distant relative or the gradual loss of a parent to chronic disease might possibly have effects that are smaller, more ambiguous, more difficult to discern. In a general discussion of studies of stress and depression, Kessler [28] makes this point clearly:

“... a major problem in interpret[ation] ... is that both chronic role-related stresses and the chronic depression by definition have occurred for so long that deciding unambiguously which came first is difficult ... The researcher, however, may focus on stresses that can

be assumed to have occurred randomly with respect to other risk factors of depression and to be inescapable, in which case matched comparison can be used to make causal inferences about long-term stress effects. A good example is the matched comparison of the parents of children having cancer, diabetes, or some other serious childhood physical disorder with the parents of healthy children. Disorders of this sort are quite common and occur, in most cases, for reasons that are unrelated to other risk factors for parental psychiatric disorder. The small amount of research shows that these childhood physical disorders have significant psychiatric effects on the family.” (p. 197)

- *Additional structural features in quasi-experiments intended to provide information about hidden biases.* The term quasi-experiment is often used to suggest a design in which certain structural features are added in an effort to provide information about hidden biases; see Section titled ‘Elaborate Theories’ for detailed discussion. In the Section titled ‘Effects on Children of Occupational Exposures to Lead’, for instance, data were collected for control children whose parents were not exposed to lead, together with data about the level of lead exposure and the hygiene of parents exposed to lead. An actual effect of lead should produce a quite specific pattern of associations: more lead in the blood of exposed children, more lead when the level of exposure is higher, more lead when the hygiene is poor. In general terms, Cook et al. [14] write that:

“... the warrant for causal inferences from quasi-experiments rests [on] structural elements of design other than random assignment—pretests, comparison groups, the way treatments are scheduled across groups ... —[which] provide the best way of ruling out threats to internal validity ... [C]onclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.” (pp. 570-1)

Randomization will produce treated and control groups that were comparable prior to treatment, and it will do this mechanically, with no understanding of the context in which the study is being conducted. When randomization is not used, an understanding of the context becomes much more important.

Context is important whether one is trying to identify what covariates to measure, or to locate settings that afford haphazard treatment assignments or subpopulations with reduced selection biases, or to determine the direction of hidden biases. Ethnographic and other qualitative studies (e.g., [3, 21]) may provide familiarity with context needed in planning an observational study, and moreover qualitative methods may be integrated with quantitative studies [55].

Because even the most carefully designed observational study will have weaknesses and ambiguities, a single observational study is often not decisive, and replication is often necessary. In replicating an observational study, one should seek to replicate the actual treatment effects, if any, without replicating any biases that may have affected the original study. Some strategies for doing this are discussed in [48].

Adjusting for Biases Visible in Observed Covariates

Matched Sampling

Selecting from a Reservoir of Potential Controls.

Among methods of adjustment for overt biases, the most direct and intuitive is matching, which compares each treated individual to one or more controls who appear comparable in terms of observed covariates. Matched sampling is most common when a small treated group is available together with a large reservoir of potential controls [57].

The structure of the study of bereavement by Lehman et al. [30] in the Section titled ‘Long-term Psychological Effects of the Death of a Close Relative’ is typical. There were 80 bereaved spouses and parents and 7581 potential controls, from whom 80 matched controls were selected. Routine administrative records were used to identify and match bereaved and control subjects, but additional information was needed from matched subjects for research purposes, namely, psychiatric outcomes. It is neither practical nor important to obtain psychiatric outcomes for all 7581 potential controls, and instead, matching selected 80 controls who appear comparable to treated subjects.

Most commonly, as in both Lehman et al.’s [30] study of bereavement in the Section titled ‘Long-term Psychological Effects of the Death of a Close

Relative’ and Morton et al.’s [39] study of lead exposure in the Section titled ‘Effects on Children of Occupational Exposures to Lead’, each treated subject is matched to exactly one control, but other matching structures may yield either greater bias reduction or estimates with smaller standard errors or both. In particular, if the reservoir of potential controls is large, and if obtaining data from controls is not prohibitively expensive, then the standard errors of estimated treatment effects can be substantially reduced by matching each treated subject to several controls [62]. When several controls are used, substantially greater bias reduction is possible if the number of controls is not constant, instead varying from one treated subject to another [37].

Multivariate Matching Using Propensity Scores.

In matching, the first impulse is to try to match each treated subject to a control who appears nearly the same in terms of observed covariates; however, this is quickly seen to be impractical when there are many covariates. For instance, with 20 binary covariates, there are 2^{20} or about a million types of individuals, so even with thousands of potential controls, it will often be difficult to find a control who matches a treated subject on all 20 covariates.

Randomization produces covariate balance, not perfect matches. Perfect matches are not needed to balance observed covariates. Multivariate matching methods attempt to produce matched pairs or sets that balance observed covariates, so that, in aggregate, the distributions of observed covariates are similar in treated and control groups. Of course, unlike randomization, matching cannot be expected to balance unobserved covariates.

The **propensity score** is the conditional probability (see **Probability: An Introduction**) of receiving the treatment rather than the control given the *observed* covariates [52]. Typically, the propensity score is unknown and must be estimated, for instance, using **logistic regression** [19] of the binary category, treatment/control on the observed covariates. The propensity score is defined in terms of the *observed* covariates, even when there are concerns about hidden biases due to *unobserved* covariates, so estimating the propensity score is straightforward because the needed data are available. For nontechnical surveys of methods using propensity scores, see [7, 27], and see [33] for discussion of propensity scores for doses of treatment.

Matching on one variable, the propensity score, tends to balance all of the observed covariates, even though matched individuals will typically differ on many observed covariates. As an alternative, matching on the propensity score and one or two other key covariates will also tend to balance all of the observed covariates. If it suffices to adjust for the observed covariates – that is, if there is no hidden bias due to unobserved covariates – then it also suffices to adjust for the propensity score alone. These results are Theorems 1 through 4 of [52]. A study of the psychological effects of prenatal exposures to barbiturates balanced 20 observed covariates by matching on an estimated propensity score and sex [54].

One can and should check to confirm that the propensity score has done its job. That is, one should check that, after matching, the distributions of observed covariates are similar in treated and control groups; see [53, 54] for examples of this simple process. Because theory says that a correctly estimated propensity score should balance observed covariates, this check on covariate balance is also a check on the model used to estimate the propensity score. If some covariates are not balanced, consider adding to the logit model interactions or quadratics involving these covariates; then check covariate balance with the new propensity score.

Bergstralh, Kosanke, and Jacobsen [4] provide SAS software for an optimal matching algorithm.

Stratification

Stratification is an alternative to matching in which subjects are grouped rather than paired. Cochran [13] showed that five strata formed from a single continuous covariate can remove about 90% of the bias in that covariate. Strata that balance many covariates at once can often be formed by forming five strata at the quintiles of an estimated propensity score. A study of coronary bypass surgery balanced 74 covariates using five strata formed from an estimated propensity score [53].

The optimal stratification – that is, the stratification that makes treated and control subjects as similar as possible within strata – is a type of matching called *full matching* in which a treated subject can be matched to several controls or a control can be matched to several treated subjects [45]. An optimal full matching, hence also an optimal stratification, can be determined using network optimization.

Model-based Adjustments

Unlike matched sampling and stratification, which compare treated subjects directly to actual controls who appear comparable in terms of observed covariates, model-based adjustments, such as covariance adjustment, use data on treated and control subjects without regard to their comparability, relying on a model, such as a linear regression model, to predict how subjects would have responded under treatments they did not receive. In a case study from labor economics, Dehejia and Wahba [20] compared the performance of model-based adjustments and matching, and Rubin [58, 60] compared performance using simulation. Rubin found that model-based adjustments yielded smaller standard errors than matching when the model is precisely correct, but model-based adjustments were less robust than matching when the model is wrong. Indeed, he found that if the model is substantially incorrect, model-based adjustments may not only fail to remove overt biases, they may even increase them, whereas matching and stratification are fairly consistent at reducing overt biases. Rubin found that the combined use of matching and model-based adjustments was both robust and efficient, and he recommended this strategy in practice.

Appraising Sensitivity to Hidden Bias

With care, matching, stratification, model-based adjustments and combinations of these techniques may often be used to remove overt biases accurately recorded in the data at hand, that is, biases visible in imbalances in observed covariates. However, when observational studies are subjected to critical evaluation, a common concern is that the adjustments failed to control for some covariate that was not measured. In other words, the concern is that treated and control subjects were not comparable prior to treatment with respect to this unobserved covariate, and had this covariate been measured and controlled by adjustments, then the conclusions about treatment effects would have been different. This is not a concern in randomized experiments, because randomization balances both observed and unobserved covariates. In an observational study, a **sensitivity analysis** (see **Sensitivity Analysis in Observational Studies**) asks how such hidden biases of various magnitudes might alter the conclusions of

the study. Observational studies vary greatly in their sensitivity to hidden bias.

Cornfield et al. [17] conducted the first formal sensitivity analysis in a discussion of the effects of cigarette smoking on health. The objection had been raised that smoking might not cause lung cancer, but rather that there might be a genetic predisposition both to smoke and to develop lung cancer, and that this, not an effect of smoking, was responsible for the association between smoking and lung cancer. Cornfield et al. [17] wrote:

... if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than among nonsmokers. (p. 40)

Though straightforward to compute, their sensitivity analysis is an important step beyond the familiar fact that association does not imply causation. A sensitivity analysis is a specific statement about the *magnitude* of hidden bias that would need to be present to explain the associations actually observed. Weak associations in small studies can be explained away by very small biases, but only a very large bias can explain a strong association in a large study.

A simple, general method of sensitivity analysis introduces a single sensitivity parameter Γ that measures the degree of departure from random assignment of treatments. Two subjects with the same observed covariates may differ in their odds of receiving the treatment by at most a factor of Γ . In an experiment, random assignment of treatments ensures that $\Gamma = 1$, so no sensitivity analysis is needed. In an observational study with $\Gamma = 2$, if two subjects were matched exactly for observed covariates, then one might be twice as likely as the other to receive the treatment because they differ in terms of a covariate not observed. Of course, in an observational study, Γ is unknown. A sensitivity analysis tries out several values of Γ to see how the conclusions might change. Would small departures from random assignment alter the conclusions? Or, as in the studies of smoking and lung cancer, would only very large departures from random assignment alter the conclusions? For each value of Γ , it is possible to place bounds on a statistical inference – perhaps for $\Gamma = 3$, the true P value is unknown, but must be between

0.0001 and 0.041. Analogous bounds may be computed for point estimates and confidence intervals. How large must Γ be before the conclusions of the study are qualitatively altered? If for $\Gamma = 9$, the P value for testing no effect is between 0.00001 and 0.02, then the results are highly insensitive to bias – only an enormous departure from random assignment of treatments could explain away the observed association between treatment and outcome. However, if for $\Gamma = 1.1$, the P value for testing no effect is between 0.01 and 0.3, then the study is extremely sensitive to hidden bias – a tiny bias could explain away the observed association. This method of sensitivity analysis is discussed in detail with many examples in Section 4 of [49] and the references given there.

For instance, Morton et al.'s [39] study in the Section titled 'Effects on Children of Occupational Exposures to Lead' used Wilcoxon's **signed-ranks test** to compare blood lead levels of 33 exposed and 33 matched control children. The pattern of matched pair differences they observed would yield a P value less than 10^{-5} in a randomized experiment. For $\Gamma = 3$, the range of possible P values is from about 10^{-15} to 0.014, so a bias of this magnitude could not explain the higher lead levels among exposed children. In words, if Morton et al. [39] had failed to control by matching a variable strongly related to blood lead levels and three times more common among exposed children, this would not have been likely to produce a difference in lead levels as large as the one they observed. The upper bound on the P value is just about 0.05 when $\Gamma = 4.35$, so the study is quite insensitive to hidden bias, but not as insensitive as the studies of heavy smoking and lung cancer. For $\Gamma = 5$ and $\Gamma = 6$, the upper bounds on the P value are 0.07 and 0.12, respectively, so biases of this magnitude could explain the observed association. Sensitivity analyses for point estimates and confidence intervals for this example are in Section 4.3.4 and Section 4.3.5 of [49].

Several other methods of sensitivity analysis are discussed in [16, 24, 26], and [31].

Elaborate Theories

Elaborate Theories and Pattern Specificity

What *can* be observed to provide evidence about hidden biases, that is, biases due to covariates that

were *not* observed? Cochran [12] summarizes the view of **Sir Ronald Fisher**, the inventor of the randomized experiment:

About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: “Make your theories elaborate.” The reply puzzled me at first, since by Occam’s razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.

Similarly, Cook & Shadish [15] (1994, p. 565): “Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations. (p. 95)” Some patterns of response are scientifically plausible as treatment effects, but others are not [25], [65]. “[W]ith more pattern specificity,” writes Trochim [63], “it is generally less likely that plausible alternative explanations for the observed effect pattern will be forthcoming. (p. 580)”

Example of Reduced Sensitivity to Hidden Bias Due to Pattern Specificity

Morton et al.’s [39] study of lead exposures in the Section titled ‘Effects on Children of Occupational Exposures to Lead’ provides an illustration. Their elaborate theory predicted: (a) higher lead levels in the blood of exposed children than in matched control children, (b) higher lead levels in exposed children whose parents had higher lead exposure on the job, and (c) higher lead levels in exposed children whose parents practiced poorer hygiene upon leaving the factory. Since each of these predictions was consistent with observed data, to attribute the observed associations to hidden bias rather than an actual effect of lead exposure, one would need to postulate biases that could produce all three associations.

In a formal sense, elaborate theories play two roles: (a) they can aid in detecting hidden biases [49], and (b) they can make a study less sensitive to hidden bias [50, 51]. In Section titled ‘Effects on Children of Occupational Exposures to Lead’, if exposed children had lower lead levels than controls,

or if higher exposure predicted lower lead levels, or if poor hygiene predicted lower lead levels, then this would be difficult to explain as an effect caused by lead exposure, and would likely be understood as a consequence of some unmeasured bias, some way children who appeared comparable were in fact not comparable. Indeed, the pattern specific comparison is less sensitive to hidden bias. In detail, suppose that the exposure levels are assigned numerical scores, 1 for a child whose father had either low exposure or good hygiene, 2 for a father with high exposure and poor hygiene, and 1.5 for the several intermediate situations. The sensitivity analysis discussed in the Section titled ‘Appraising Sensitivity to Hidden Bias’ used the signed rank test to compare lead levels of the 33 exposed children and their 33 matched controls, and it became sensitive to hidden bias at $\Gamma = 4.35$, because the upper bound on the P value for testing no effect had just reached 0.05. Instead, using the dose-signed-rank statistic [46, 50] to incorporate the predicted pattern, the comparison becomes sensitive at $\Gamma = 4.75$; that is, again, the upper bound on the P value for testing no effect is 0.05. In other words, some biases that would explain away the higher lead levels of exposure children are not large enough to explain away the pattern of associations predicted by the elaborate theory. To explain the entire pattern, noticeably larger biases would need to be present.

A reduction in sensitivity to hidden bias can occur when a correct elaborate theory is strongly confirmed by the data, but an increase in sensitivity can occur if the pattern is contradicted [50]. It is possible to contrast competing design strategies in terms of their ‘design sensitivity,’ that is, their ability to reduce sensitivity to hidden bias [51].

Common Forms of Pattern Specificity

There are several common forms of pattern specificity or elaborate theories [44, 49, 61].

- *Two control groups.* Campbell [8] advocated selecting two control groups to systematically vary an unobserved covariate, that is, to select two different groups not exposed to the treatment, but known to differ in terms of certain unobserved covariates. For instance, Card and Krueger [11] examined the common claim among economists that increases in the

minimum wage cause many minimum wage earners to loose their jobs. They did this by looking at changes in employment at fast food restaurants – Burger Kings, Wendy’s, KFCs, and Roy Rogers’ – when New Jersey increased its minimum wage by about 20%, from \$4.25 to \$5.05 per hour, on 1 April 1992, comparing employment before and after the increase. In certain analyses, they compared New Jersey restaurants initially paying \$4.25 to two control groups: (a) restaurants in the same chains across the Delaware River in Pennsylvania where the minimum wage had not increased, and (b) restaurants in the same chains in affluent sections of New Jersey where the starting wage was at least \$5.00 before 1 April 1992. An actual effect of raising the minimum wage should have negligible effects on both control groups. In contrast, one anticipates differences between the two control groups if, say, Pennsylvania Burger Kings were poor controls for New Jersey Burger Kings, or if employment changes in relatively affluent sections of New Jersey are very different from those in less affluent sections. Card and Krueger found similar changes in employment in the two control groups, and similar results in their comparisons of the treated group with either pair control group. An algorithm for optimal pair matching with two control groups is illustrated with Card and Krueger’s study in [32].

- *Coherence among several outcomes and/or several doses.* Hill [25] emphasized the importance of a coherent pattern of associations and of dose-response relationships, and Weiss [66] further developed these ideas. Campbell [9] wrote: “... great inferential strength is added when each theoretical parameter is exemplified in two or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned (p. 33).” Webb [64] speaks of triangulation. The lead example in the Section titled ‘Example of Reduced Sensitivity to Hidden Bias Due to Pattern Specificity’ provides one illustration and Reynolds and West [42] provide another. Related statistical theory is in [46, 51] and the references given there.

- *Unaffected outcomes; ineffective treatments.* An elaborate theory may predict that certain outcomes should not be affected by the treatment or certain treatments should not affect the outcome; see Section 6 of [49] and [67]. For instance, in a *case-crossover study* [34], Mittleman et al. [38] asked whether bouts of anger might cause myocardial infarctions or heart attacks, finding a moderately strong and highly significant association. Although there are reasons to think that bouts of anger might cause heart attacks, there are also reasons to doubt that bouts of curiosity cause heart attacks. Mittleman et al. found curiosity was not associated with myocardial infarction, writing: “the specificity observed for anger ... as opposed to curiosity ... argue against recall bias.” McKilip [35] suggests that an unaffected or ‘control’ outcome might sometimes serve in place of a control group, and Legorreta et al. [29] illustrate this possibility in a study of changes in the demand for a type of surgery following a technological change that reduced cost and increased safety.

Summary

In the design of an observational study, an attempt is made to reconstruct some of the structure and strengths of an experiment. Analytical adjustments, such as matching, are used to control for overt biases, that is, pretreatment differences between treated and control groups that are visible in observed covariates. Analytical adjustments may fail because of hidden biases, that is, important covariates that were not measured and therefore not controlled by adjustments. Sensitivity analysis indicates the magnitude of hidden bias that would need to be present to alter the qualitative conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias; therefore, it is important to know whether a particular study is sensitive to small biases or insensitive to quite large biases. Hidden biases may leave visible traces in observed data, and a variety of tactics involving pattern specificity are aimed at distinguishing actual treatment effects from hidden biases. Pattern specificity may aid in detecting hidden bias or in reducing sensitivity to hidden bias.

Acknowledgment

This work was supported by grant SES-0345113 from the US National Science Foundation.

References

- [1] Angrist, J.D. (2003). Randomized trials and quasi-experiments in education research. *NBER Reporter*, Summer, 11–14.
- [2] Angrist, J.D. & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement, *Quarterly Journal of Economics* **114**, 533–575.
- [3] Athens, L. (1997). *Violent Criminal Acts and Actors Revisited*, University of Illinois Press, Urbana.
- [4] Bergstralh, E.J., Kosanke, J.L. & Jacobsen, S.L. (1996). Software for optimal matching in observational studies, *Epidemiology* **7**, 331–332. <http://www.mayo.edu/hsr/sasmac.html>
- [5] Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation*, Sage Publications, Thousand Oaks.
- [6] Bound, J. (1989). The health and earnings of rejected disability insurance applicants, *American Economic Review* **79**, 482–503.
- [7] Braitman, L.E. & Rosenbaum, P.R. (2002). Rare outcomes, common treatments: analytic strategies using propensity scores, *Annals of Internal Medicine* **137**, 693–695.
- [8] Campbell, D.T. (1969). Prospective: artifact and Control, in *Artifact in Behavioral Research*, R. Rosenthal & R. Rosnow, eds, Academic Press, New York.
- [9] Campbell, D.T. (1988). *Methodology and Epistemology for Social Science: Selected Papers*, University of Chicago Press, Chicago, pp. 315–333.
- [10] Campbell, D.T. & Boruch, R.F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects, in *Evaluation and Experiment*, C.A. Bennett & A.A. Lumsdaine, eds, Academic Press, New York, pp. 195–296.
- [11] Card, D. & Krueger, A. (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review* **84**, 772–793.
- [12] Cochran, W.G. (1965). The planning of observational studies of human populations (with Discussion), *Journal of the Royal Statistical Society. Series A* **128**, 134–155.
- [13] Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics* **24**, 295–313.
- [14] Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi-experimentation, in *Handbook of Industrial and Organizational Psychology*, M. Dunnette & L. Hough, eds, Consulting Psychologists Press, Palo Alto, pp. 491–576.
- [15] Cook, T.D. & Shadish, W.R. (1994). Social experiments: Some developments over the past fifteen years, *Annual Review of Psychology* **45**, 545–580.
- [16] Copas, J.B. & Li, H.G. (1997). Inference for non-random samples (with discussion), *Journal of the Royal Statistical Society. Series B* **59**, 55–96.
- [17] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.
- [18] Cox, D.R. & Read, N. (2000). *The Theory of the Design of Experiments*, Chapman & Hall/CRC, New York.
- [19] Cox, D.R. & Snell, E.J. (1989). *Analysis of Binary Data*, Chapman & Hall, London.
- [20] Dehejia, R.H. & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**, 1053–1062.
- [21] Estroff, S.E. (1985). *Making it Crazy: An Ethnography of Psychiatric Clients in an American Community*, University of California Press, Berkeley.
- [22] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [23] Friedman, L.M., DeMets, D.L. & Furberg, C.D. (1998). *Fundamentals of Clinical Trials*, Springer-Verlag, New York.
- [24] Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.
- [25] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.
- [26] Imbens, G.W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* **93**, 126–132.
- [27] Joffe, M.M. & Rosenbaum, P.R. (1999). Propensity scores, *American Journal of Epidemiology* **150**, 327–333.
- [28] Kessler, R.C. (1997). The effects of stressful life events on depression, *Annual Review of Psychology* **48**, 191–214.
- [29] Legorreta, A.P., Silber, J.H., Costantino, G.N., Kobylinski, R.W. & Zatz, S.L. (1993). Increased cholecystectomy rate after the introduction of laparoscopic cholecystectomy, *Journal of the American Medical Association* **270**, 1429–1432.
- [30] Lehman, D., Wortman, C. & Williams, A. (1987). Long-term effects of losing a spouse or a child in a motor vehicle crash, *Journal of Personality and Social Psychology* **52**, 218–231.
- [31] Lin, D.Y., Psaty, B.M. & Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics* **54**, 948–963.
- [32] Lu, B. & Rosenbaum, P.R. (2004). Optimal pair matching with two control groups, *Journal of Computational and Graphical Statistics* **13**, 422–434.

- [33] Lu, B., Zanutto, E., Hornik, R. & Rosenbaum, P.R. (2001). Matching with doses in an observational study of a media campaign against drug abuse, *Journal of the American Statistical Association* **96**, 1245–1253.
- [34] Maclure, M. & Mittleman, M.A. (2000). Should we use a case-crossover design? *Annual Review of Public Health* **21**, 193–221.
- [35] McKillip, J. (1992). Research without control groups: a control construct design, in *Methodological Issues in Applied Social Psychology*, F.B. Bryant, J. Edwards & R.S. Tindale, eds, Plenum Press, New York, pp. 159–175.
- [36] Meyer, B.D. (1995). Natural and quasi-experiments in economics, *Journal of Business and Economic Statistics* **13**, 151–161.
- [37] Ming, K. & Rosenbaum, P.R. (2000). Substantial gains in bias reduction from matching with a variable number of controls, *Biometrics* **56**, 118–124.
- [38] Mittleman, M.A., Maclure, M., Sherwood, J.B., Mulry, R.P., Tofler, G.H., Jacobs, S.C., Friedman, R., Benson, H. & Muller, J.E. (1995). Triggering of acute myocardial infarction onset by episodes of anger, *Circulation* **92**, 1720–1725.
- [39] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555.
- [40] Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, Section 9 (In Polish), *Roczniki Nauk Rolniczych*, **Tom X**, 1–51, Reprinted in English with Discussion in *Statistical Science* 1990., **5**, 463–480.
- [41] Oreopoulos, P. (2003). The long-run consequences of living in a poor neighborhood, *Quarterly Journal of Economics* **118**, 1533–1575.
- [42] Reynolds, K.D. & West, S.G. (1987). A multiplist strategy for strengthening nonequivalent control group designs, *Evaluation Review* **11**, 691–714.
- [43] Rosenbaum, P.R. (1984a). The consequences of adjustment for a concomitant variable that has been affected by the treatment, *Journal of the Royal Statistical Society. Series A* **147**, 656–666.
- [44] Rosenbaum, P.R. (1984b). From association to causation in observational studies, *Journal of the American Statistical Association* **79**, 41–48.
- [45] Rosenbaum, P.R. (1991). A characterization of optimal designs for observational studies, *Journal of the Royal Statistical Society. Series B* **53**, 597–610.
- [46] Rosenbaum, P.R. (1997). Signed rank statistics for coherent predictions, *Biometrics* **53**, 556–566.
- [47] Rosenbaum, P.R. (1999). Choice as an alternative to control in observational studies (with Discussion), *Statistical Science* **14**, 259–304.
- [48] Rosenbaum, P.R. (2001). Replicating effects and biases, *American Statistician* **55**, 223–227.
- [49] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, Springer-Verlag, New York.
- [50] Rosenbaum, P.R. (2003). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1–10.
- [51] Rosenbaum, P.R. (2004). Design sensitivity in observational studies, *Biometrika* **91**, 153–164.
- [52] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.
- [53] Rosenbaum, P. & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**, 516–524.
- [54] Rosenbaum, P. & Rubin, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *American Statistician* **39**, 33–38.
- [55] Rosenbaum, P.R. & Silber, J.H. (2001). Matching and thick description in an observational study of mortality after surgery, *Biostatistics* **2**, 217–232.
- [56] Rosenzweig, M.R. & Wolpin, K.I. (2000). Natural “natural experiments” in economics, *Journal of Economic Literature* **38**, 827–874.
- [57] Rubin, D. (1973a). Matching to remove bias in observational studies, *Biometrics* **29**, 159–183. Correction: 1974., **30**, 728.
- [58] Rubin, D.B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies, *Biometrics* **29**, 185–203.
- [59] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.
- [60] Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association* **74**, 318–328.
- [61] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.
- [62] Smith, H.L. (1997). Matching with multiple controls to estimate treatment effects in observational studies, *Sociological Methodology* **27**, 325–353.
- [63] Trochim, W.M.K. (1985). Pattern matching, validity and conceptualization in program evaluation, *Evaluation Review* **9**, 575–604.
- [64] Webb, E.J. (1966). Unconventionality, triangulation and inference, *Proceedings of the Invitational Conference on Testing Problems*, Educational Testing Service, Princeton, pp. 34–43.
- [65] Weed, D.L. & Hursting, S.D. (1998). Biologic plausibility in causal inference: current method and practice, *American Journal of Epidemiology* **147**, 415–425.
- [66] Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of ‘dose-response’, *American Journal of Epidemiology* **113**, 487–90.
- [67] Weiss, N.S. (2002). Can the “specificity” of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology* **13**, 6–8.

12 Observational Study

- [68] Wright, M.A., Wintemute, G.J. & Rivara, F.P. (1999). Effectiveness of denial of handgun purchase to persons believed to be at high risk for firearm violence, *American Journal of Public Health* **89**, 88–90.
- [69] Zabin, L.S., Hirsch, M.B. & Emerson, M.R. (1989). When urban adolescents choose abortion: effects on education, psychological status, and subsequent pregnancy, *Family Planning Perspectives* **21**, 248–255.

PAUL R. ROSENBAUM

Odds and Odds Ratios

TAMÁS RUDAS

Volume 3, pp. 1462–1467

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Odds and Odds Ratios

Odds

Odds are related to possible observations (events) associated with an experiment or **observational study**. The value of the odds tells how many times one event is more likely to occur than another one. The value of the odds is obtained as the ratio of the probabilities of the two events. These probabilities may be the true or estimated probabilities, in which the estimated probabilities may be taken from some data available or may be based on a statistical model. For example, in a study on unemployment it was found that 40% of the respondents were out of the labor force, 50% were employed, and 10% were unemployed. Then, the value of the odds of being unemployed versus employed in the data is $0.1/0.5 = 0.2$. Notice that the same odds could be obtained by using the respective frequencies instead of the probabilities.

Odds are only defined for events that are disjoint, that is, cannot occur at the same time. For example, the ratio of the probabilities of being unemployed and of not working (that is, being out of the labor force or being unemployed), $0.1/(0.4 + 0.1) = 0.2$ is not the value of odds, rather that of a conditional probability (see **Probability: An Introduction**). However, odds may be used to specify a distribution just as well as probabilities. Saying that the value of odds of being unemployed versus employed is 0.2 and that of being employed versus being out of the labor force is 1.25 completely specifies the distribution of the respondents in the three categories.

Of particular interest is the value of the odds of an event versus all other possibilities, that is, versus its complement, sometimes called the *odds of the event*. The odds of being out of the labor force in the above example is $0.4/0.6 = 2/3$. If p is the probability of an event, the value of its odds is $p/(1 - p)$. The (natural) logarithm of this quantity, $\log(p/(1 - p))$, is the logit of the event. In **logistic regression**, the effects of certain explanatory variables on a binary response are investigated by approximating the logit of the variable with a function of the explanatory variables (see [1]).

The odds are relatively simple parameters of a distribution, and their statistical analysis is usually confined to constructing confidence intervals for the

true value of the odds and to testing hypotheses that the value of the odds is equal to a specified value. To obtain an asymptotic **confidence interval** for the true value of the odds, one may use the asymptotic variance of the estimated value of the odds. All these statistical tasks can be transformed to involve the logarithm of the odds. If the estimated probabilities of the events are p and q , then the logarithm of the odds is estimated as $\log(p/q)$, and its asymptotic (approximate) variance is $s^2 = 1/np + 1/nq$, where n is the sample size. Therefore, an approximate 95% confidence interval for the true value of the logarithm of the odds is $(t - 2s, t + 2s)$. The hypothesis that the logarithm of the odds is equal to a specified value, say, u (against the alternative that it is not equal to u), is rejected at the 95% level, if u lies outside of this confidence interval. The value $u = 0$ means that the two events have the same probability, that is, $p = q$.

In a two-way **contingency table**, the odds related to events associated with one variable may be computed within the different categories of the other variable. If, for example, the employment variable is cross-classified by gender, as shown in Table 1, one obtains a 2×3 table. Here $P(1, 2)/P(1, 1)$ is the value of the *conditional odds* of being unemployed versus employed for men. The value of the same conditional odds for women is $P(2, 2)/P(2, 1)$. The same odds, computed for everyone, that is, $P(+, 2)/P(+, 1)$, are called the *marginal odds* here.

Odds Ratios for 2×2 Contingency Tables

Ratios of conditional odds are called *odds ratios*. In a 2×2 table (like the first two columns of Table 1 that cross classify the employment status of those in the labor force with gender), the value of the odds ratio is $(P(1, 1)/P(1, 2))/(P(2, 1)/P(2, 2))$. This is the ratio of the conditional odds of the column variable, conditioned on the categories of the row variable. If the ratio of the conditional odds

Table 1 Cross-classification of sex and employment status

	Employed	Unemployed	Out of the labor force	Total
Men	$P(1, 1)$	$P(1, 2)$	$P(1, 3)$	$P(1, +)$
Women	$P(2, 1)$	$P(2, 2)$	$P(2, 3)$	$P(2, +)$
Total	$P(+, 1)$	$P(+, 2)$	$P(+, 3)$	$P(+, +)$

of the row variable, conditioned on the categories of the column variable is considered, one obtains $(P(1, 1)/P(2, 1))/(P(1, 2)/P(2, 2))$, and this has the same value as the previous one. The common value, $\alpha = (P(1, 1)P(2, 2))/(P(1, 2)P(2, 1))$, the odds ratio, is also called the *cross product ratio*. When the order of the categories of either one of the variables is changed, the odds ratio changes to its reciprocal. When the order of the categories of both variables is changed, α remains unchanged.

When the two variables in the table are independent, that is, when the rows and the columns are proportional, the conditional odds of either one of the variables, given the categories of the other one, are equal to each other, and the value of the odds ratio is 1. Therefore, a test of the hypothesis that $\log \alpha = 0$ is a test of independence. A test of this hypothesis, and a test of any hypothesis assuming a fixed value of the odds ratio, may be based on the asymptotic (approximate) variance of the estimated odds ratio, $s^2 = (1/n)(1/P(1, 1) + 1/P(1, 2) + 1/P(2, 1) + 1/P(2, 2))$ (see [3]).

The more the conditional odds of one variable, given the different categories of the other variable differ, the further away from 1 is the value of the odds ratio. Therefore, the odds ratio is a measure of how different the conditional odds are from each other and it may be used as a measure of the strength of association between the two variables forming the table.

The Odds Ratio as a Measure of Association

Because changing the order of the categories of one variable obviously does not affect the strength of association (*see Measures of Association*) but changes the value of the odds ratio to its reciprocal, the values α and $1/\alpha$ refer to the same strength of association. With ordinal variables or with categorical variables having the same categories, the categories of the two variables have a fixed order with respect to each other (both in the same order) and in such cases α and $1/\alpha$ mean different directions of association (positive and negative) of the same strength. In cases when the order of the categories of one variable does not determine the order of the categories of the other variable, the association does not have a direction. For further discussion of the direction of association, see [6].

An interpretation of the odds ratio, as a measure of the strength of association between two categorical variables, is that it measures how helpful it is to know in which category of one of the variables an observation is in order to guess in which category of the other variable it is. If in a 2×2 table $P(1, 1)/P(1, 2) = P(2, 1)/P(2, 2)$, then knowing that the observation is in the first or the second row does not help in guessing in which column it is. In this case, $\alpha = 1$. If, however, say, $P(1, 1)/P(1, 2)$ is bigger than $P(2, 1)/P(2, 2)$, then a larger fraction of those in the first row will be in the first column and only a smaller fraction of those in the second row. In this case, α has a large value. If any of the entries in a 2×2 table is equal to 0, the value of the odds ratio cannot be computed. One may think that if $P(1, 1) = 0$, α may be computed and its value is 0, but by changing the order of categories one would have to divide by zero in the computation of α , which is impossible.

When working with odds ratios as a measure of the strength of association (or with any measure of association, for that matter), it should be kept in mind that having a large observed value of the odds ratio and having a statistically significant one are very different things. Tables 2 and 3 show two different cross classifications of 40 observations. In Table 2, $\alpha = 1.53$, $\log \alpha = 0.427$ and its approximate standard deviation is 0.230, therefore α is not significantly different from 1 at the 95% level. On the other hand, in Table 3 $\alpha = 1.31$, $\log \alpha = 0.270$ and its approximate standard deviation is 0.102. Therefore, in Table 3 the observed (estimated) odds ratio is significantly different from 1, at the 95% level. The larger observed value did not prove significant, while

Table 2 Cross-classification of 40 observations with $\alpha = 1.53$

23	1
15	1

Table 3 Another cross-classification of 40 observations with $\alpha = 1.31$

11	12
7	10

the smaller one did, at the same level and using the same sample size. While some users may find this fact counterintuitive, it is a reflection of the different distributions seen in Tables 2 and 3. The observed distribution is much more uniform in Table 3 than in Table 2 and therefore one may consider it a more reliable estimate of the underlying population distribution than the one in Table 2. Consequently, the estimate of the odds ratio derived from Table 3 is more reliable than the estimate derived from Table 2 and thus one has a smaller variance of the estimate in Table 3 than in Table 2.

Variation Independence of the Odds Ratio and the Marginal Distributions

One particularly attractive property of the odds ratio as a measure of association is its variation independence from the marginal distributions of the table. This means that any pair of marginal distributions may be combined with any value of the odds ratio, and it implies that the odds ratio measures an aspect of association that is not influenced by the marginal distributions. In fact, one can say that the odds ratio represents the information that is contained in the joint distribution in addition to the marginal distributions (see [6]).

Variation independence is also important for the interpretation of the value of the odds ratio as a measure of the strength of association: it makes the values of the odds ratio comparable across tables with different marginal distributions. The difficulties associated with measures that are not variationally independent from the marginals are illustrated using the data in Tables 4 and 5. One possible measure of association is $\beta = P(1, 1)/P(+, 1) - P(2, 1)/P(+, 2)$, that is, the difference between the conditional probabilities of the first row category (e.g., success) in the two categories of the column variables (e.g., treatment and control). The value of β is 0.1 for both sets of data. In spite of this, one would not think association has the same strength (e.g., being in treatment or control has a same effect on success) in the two tables because, given the marginals, for Table 4, the maximum value of β is 0.3 and for Table 5 the maximum value is 0.9. In Table 4, β is one third of its possible maximum, while in Table 5 it is one ninth of its maximum value. The measure β is not variationally independent from the marginals and therefore lacks calibration.

Table 4 A 2×2 table with $\beta = 0.1$

10	5
40	45

Table 5 Another 2×2 table with $\beta = 0.1$

25	20
25	30

$I \times J$ Tables

The concept of odds ratio as a measure of the strength of association may be extended beyond 2×2 tables. For $I \times J$ tables, one possibility is to consider local odds ratios. These are obtained by selecting two adjacent rows and two adjacent columns of the table. Their intersection is a local 2×2 subtable of the contingency table and the odds ratio computed here is a local odds ratio. If the two variables forming the table are independent, all the local odds ratios are equal to 1. The number of different possible local odds ratios is $(I - 1)(J - 1)$ and they may be naturally arranged into an $(I - 1) \times (J - 1)$ table. In the (k, l) position of this table, one has the odds ratio $(P(k, l)P(k + 1, l + 1))/(P(k, l + 1)P(k + 1, l))$. Under independence, the table of local odds ratios contains 1s in all positions. For another generalization of the concept of odds ratio to $I \times J$ tables, see [6].

In practice, the model of independence often proves too restrictive. In fact, the model of independence specifies, out of the total number of parameters, $IJ - 1$, the values of $(I - 1)(J - 1)$ (the local odds ratios) and leaves only $I + J - 2$ parameters (the marginal distributions of the two variables) to estimate. In a 6×8 table, for example, this means fixing 35 parameters and leaving 12 free. Therefore, models for two-way tables that are less restrictive than the model of independence and still postulate some kind of a simple structure are of interest. A class of such models may be obtained by restricting the structure of the $(I - 1) \times (J - 1)$ table of local odds ratios. Such models are called $I \times J$ association models (see [4]).

The uniform association model (U-model) assumes that the local odds ratios have a common (unspecified) value. If that common value happens to be 1,

4 Odds and Odds Ratios

Table 6 A distribution with all local odds ratios being equal to 3

10	20	20	30
30	180	540	2430
40	720	6480	87480

one has independence. Under the U-model, the table of local odds ratios is uniformly distributed, but the original table may show actually quite strong association. For example, Table 6 illustrates a distribution with common local odds ratios (all equal to 3). The row association model (R-model) assumes that the structure of the table of local odds ratios is such that there is a row effect in the sense that the local odds ratios are constant within the rows but are different in different rows. The column association model (C-model) is defined similarly. There are two models available that assume the presence of both row and column effects on the local odds ratio. The $R + C$ -model assumes that these effects are additive, the RC-model assumes that they are multiplicative on a logarithmic scale (see [4]). All these models are in between assuming the very restrictive model of independence and not assuming any structure at all.

Conditional and Marginal Association Measured by the Odds Ratio

Table 7 presents a (hypothetical) cross-classification of respondents according to two binary variables, sex (male or female) and income (high or low). The odds ratio is equal to 1.67, indicating that the value of the

Table 7 Hypothetical cross-classification according to sex and income

	High income	Low income
Men	300	200
Women	180	200

odds of men to have a high income versus a low one is greater than that of women. Table 8 presents a breakdown of the data according to the educational level of the respondent, also with two categories (high and low). Table 7 is a two-way marginal of Table 8 and the odds ratio computed from that table is the marginal odds ratio of gender and income. The odds ratios computed from the two conditional tables in Table 8, one obtained by conditioning on educational level being low and another one conditioned on educational level being high are conditional odds ratios of gender and income, conditioned on educational level. The conditional odds ratios measure the strength of association among sex and income, separately for people with low and high levels of education. For a more detailed discussion of marginal and conditional aspects of association, see [2].

Many of us would tend to expect that if for all people considered together, men have a certain advantage (odds ratio greater than 1), and then if the respondents are divided into two groups (according to educational level in the example), men also should have an advantage (odds ratio greater than 1) in at least one of the groups. Table 8 illustrates that this is not necessarily true. The conditional odds ratios computed from Table 8 are 0.75 for people with low income and 0.55 for people with high income. That is, in both groups women have higher odds of having a high salary (versus a low one) than men do. A short (but imprecise) way to refer to this phenomenon is to say that the direction of the marginal association is opposite to the direction of the conditional associations. This is called *Simpson's paradox* (see **Paradoxes**). Notice that the reversal of the direction of the marginal association upon conditioning is not merely a mathematical possibility, it does occur in practice (see, for example, [5]).

What is wrong with the expectation that if the marginal odds of men for higher income are higher than those of women, the same should be true for men and women with low and also for men and women with high levels of education (or at least in one of

Table 8 Breakdown of the data in Table 7 according to educational level

Low educational level			High educational level		
	High income	Low income		High income	Low income
Men	80	160	Men	220	40
Women	130	195	women	50	5

these categories)? The higher odds of men would imply higher odds for men in both subcategories if the higher value of the odds was somehow an inherent property of men and not something that is a consequence of other factors. If there was a causal relationship between sex and income, that would have to manifest itself not only for all people together but also among people with low and high levels of educations. However, with the present data, the higher value of the odds of men for higher incomes is a consequence of different compositions of men and women according to their educational levels. The data show advantage for women in both categories of educational level, and men appear to have advantage when educational level is disregarded only because a larger fraction of them has high level of education than that of women, and higher education leads to higher income.

The paradox (our wrong expectation) occurs because of the routine assumption of causal relationships in cases when there is only an association. When there is unobserved heterogeneity (like the different compositions of the male and female respondents according to educational level), the conditional odds ratios may have directions different from the marginal odds ratio. When the data are collected with the goal of establishing a causal relationship, random assignment into the different categories of the treatment variable is applied to reduce unobserved heterogeneity. Obviously, such a random assignment is not always possible, for example, one cannot assign people randomly into the categories of being a man or a woman.

Higher-order Odds Ratios

In Table 8, the conditional odds ratios of sex and income, given educational level, are different. If these two conditional odds ratios were the same, one would say that educational level has no effect on the strength of association between sex and income. Further, the more different the two conditional odds ratios are, the larger is the effect of education on the strength of association between sex and income. The ratio of the two conditional odds ratios, $\text{COR}(G, I|E = \text{low}) / \text{COR}(G, I|E = \text{high}) = 0.75/0.55 = 1.36$ is a measure of the strength of this effect. Similarly, the conditional odds ratios between educational level and income may be computed for men and women.

The ratio of these two quantities is a measure of how strong the effect of sex is on the strength of relationship between educational level and income. This is $\text{COR}(E, I|G = \text{men}) / \text{COR}(E, I|G = \text{women}) = 0.091/0.067 = 1.36$. Although it has little substantive meaning in the example, one could also define a measure of the strength of effect of income on the strength of association between sex and educational level, as $\text{COR}(G, E|I = \text{high}) / \text{COR}(G, E|I = \text{low})$, and its value would also be 1.36.

It is not by pure coincidence that the three measures in the previous paragraph have the same value: this is true for all data sets, indicating that the common value does not measure a property related to any grouping of the variables. Rather, it measures association among all three variables and is called *second-order interaction*. The second-order interaction is

$$\frac{P(1, 1, 1)P(1, 2, 2)P(2, 1, 2)P(2, 2, 1)}{P(2, 2, 2)P(2, 1, 1)P(1, 2, 1)P(1, 1, 2)}$$

where $P(i, j, k)$ is the probability of an observation in the i th category of the first, j th category of the second and k th category of the third variable. Values of the second-order odds ratio show how strong is the effect of any of the variables on the strength of association between the other two variables. When the second-order odds ratio is equal to 1, no variable has an effect on the strength of association between the other two variables.

Similarly, in a four-way table, conditional second-order odds ratios may be defined and, as the common value of their ratios, one obtains third-order odds ratios. Similarly, higher-order odds ratios may be defined for tables of arbitrary dimension (see [6]).

Odds Ratios and Log-linear Models

Log-linear models (see [1] and [3]) are the most widely used tools to analyze cross-classified data. Log-linear models may be given a simple interpretation using odds ratios. A log-linear model is usually defined by listing subsets of variables that are the maximal allowed interactions (see **Interaction Effects**). Subsets of these variables are called *interactions*. Such a specification is equivalent to saying that for any subset of the variables that is not an interaction, the conditional odds ratios of all involved variables, given all possible categories of the variable not in the subset, are equal to 1. That is, these

greater subsets have no (conditional) association or interaction. For example, with variables X, Y, Z , the log-linear model with maximal interactions XY, YZ means that $\text{COR}(X, Z|Y = j) = 1$, for all categories j of Y and that the second-order odds ratio of X, Y, Z is equal to 1. The first of these conditions is the conditional independence of X and Z , given Y . See [6] and [7] for the details and advantages of this interpretation.

The interpretation of log-linear models based on conditional odds ratio also shows that the association structures that may be modeled for discrete data using log-linear models are substantially more complex than those available for continuous data under the assumption of multivariate normality. A multivariate normal distribution (see **Catalogue of Probability Density Functions**) is entirely specified by its expected values and covariance matrix. The assumption of normality implies that the bivariate marginals completely specify the distribution, and the association structure among several variables cannot be more complex than the pairwise associations. With log-linear models for categorical data, one may model complex association structures involving several variables by allowing higher-order interactions in the model.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley.
- [2] Bergsma, W. & Rudas, T. (2003). On conditional and marginal association, *Annales de la Faculte des Sciences de Toulouse* **11**(6), 443–454.
- [3] Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press.
- [4] Clogg, C.C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Variables*, Sage.
- [5] Radelet, M. (1981). Racial characteristics and the imposition of the death penalty, *American Sociological Review* **46**, 918–927.
- [6] Rudas, T. (1998). *Odds Ratios in the Analysis of Contingency Tables*, Sage.
- [7] Rudas, T. (2002). Canonical representation of log-linear models, *Communications in Statistics (Theory and Methods)* **31**, 2311–2323.

(See also **Relative Risk; Risk Perception**)

TAMÁS RUDAS

One Way Designs: Nonparametric and Resampling Approaches

ANDREW F. HAYES

Volume 3, pp. 1468–1474

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

One Way Designs: Nonparametric and Resampling Approaches

Analysis of variance (ANOVA) is arguably one of the most widely used statistical methods in behavioral science. This no doubt is attributable to the fact that behavioral science research so often involves some kind of comparison between groups that either exist naturally (such as men versus women) or are created artificially by the researcher (such as in an experiment). In a one-way design, one that includes only a single grouping factor with $k \geq 2$ levels or groups, analysis of variance is used initially to test the null hypothesis (H_0) that the k population means on some outcome variable Y are equal against the alternative hypothesis that at least two of the k population means are different:

$$\begin{aligned} H_0: \mu_1 &= \mu_2 = \cdots = \mu_k. \\ H_a: &\text{at least two } \mu \text{ are different} \end{aligned} \quad (1)$$

The test statistic in ANOVA is F , defined as the ratio of an estimate of variability in Y between groups, s_B^2 , to an estimate of the within-group variance in Y , s_E^2 (i.e., $F = s_B^2/s_E^2$). As ANOVA is typically implemented, within-group variance is estimated by deriving a pooled variance estimate, assuming equality of within-group variance in Y across the k groups. Under a true H_0 , and when the assumptions of analysis of variance are met, the P value for the obtained F ratio, F_{obtained} , can be derived in reference to the F distribution on $df_{\text{num}} = k - 1$ and $df_{\text{den}} = N - k$, where N is the total sample size. If $p \leq \alpha$, H_0 is rejected in favor of H_a , where α is the level of significance chosen for the test.

As with all null hypothesis testing procedures, the decision about H_0 is based on the P value, and, thus, it is important that you can trust the accuracy of the P value. Poor estimation of the P value can produce a decision error. When errors in estimation of the P value occur, this is typically attributable to poor correspondence between the actual sampling distribution of the obtained F ratio, under H_0 , and the theoretical sampling distribution of that ratio. In the case of one-way ANOVA, the P value can be trusted if the distribution of *all possible values* of F that you

could have observed, conditioned on H_0 being true, does, in fact, follow the F distribution. Unfortunately, there is no way of knowing whether F_{obtained} does follow the F distribution because you only have a single value of F_{obtained} describing the results of your study. However, the accuracy of the P value can be inferred indirectly by determining whether the assumptions that gave rise to the use of F as the appropriate sampling distribution are met. Those assumptions are typically described as normality of Y in each population, equality of within-group variance on Y in the k populations, and independence of the observations. Of course, the first two of these assumptions can never be proven true, so there is no way of knowing for certain whether F is the appropriate sampling distribution for computing the P value for F_{obtained} .

With the advent of high-speed, low-cost computing technology, it is no longer necessary to rely on a theoretically derived but assumption-constrained sampling distribution in order to generate the P value for F_{obtained} (see **Permutation Based Inference**). It is possible, through the use of a resampling method, to generate a P value for the hypothesis test by generating the null reference distribution empirically. The primary advantages of resampling methods are (a) the assumptions required to trust the accuracy of the P value are usually weaker and/or fewer in number and (b) the test statistic used to quantify the obtained result is not constrained to be one with a theoretical sampling distribution. This latter advantage is especially exciting, because the use of resampling allows you to choose any test statistic that is sensitive to the hypothesis of interest, not just the standard F ratio as used in one-way ANOVA.

I focus here on the two most common and widely studied resampling methods: **bootstrapping** and **randomization tests**. Bootstrapping and randomization tests differ somewhat in assumptions, purpose, and computation, but they are conceptually similar to all resampling methods in that the P value is generated by comparing the obtained result to some null reference distribution generated empirically through resampling or ‘reshuffling’ of the existing data set in some way.

Define θ as some test statistic used to quantify the obtained difference between the groups. In analysis of variance, $\theta = F$, defined as above. But θ could be any test statistic that you invent that suitably describes the obtained result. For example, θ could

2 One Way Designs

Table 1 A sample of 15 units distributed across three groups measured on variable Y

Group 1 ($n_1 = 4$)	Group 2 ($n_1 = 6$)	Group 3 ($n_3 = 5$)
6	4	2
5	8	4
8	7	5
10	6	2
	9	6
	8	
$\bar{Y}_1 = 7.25$ $s_1 = 2.22$	$\bar{Y}_2 = 7.00$ $s_2 = 1.79$	$\bar{Y}_3 = 3.80$ $s_3 = 1.79$

be a quantification of group differences using the group medians, trimmed means, or anything you feel best quantifies the obtained result. Of course, θ is a random variable, but in any study, you observe only a single instantiation of θ . A different sample of N total units drawn from a population and then distributed in the same way among k groups would produce a different value of θ . Similarly, a different random reassignment of N available units in an experiment into k groups would also produce a different value of θ .

Below, I describe how bootstrapping and randomization tests can be used to generate the desired P value to test a null hypothesis. Toward the end, I discuss some issues involved in the selection of a test statistic, and how this influences what null hypothesis is actually tested.

For the sake of illustrating bootstrapping and randomization tests, I will refer to a hypothetical data set. The design and nature of the study will differ for different examples I present. For each study, however, $N = 15$ units are distributed among $k = 3$ groups of size $n_1 = 4$, $n_2 = 6$, and $n_3 = 5$. Each unit is measured on an outcome variable Y , and the individual values of Y and the sample mean and standard deviation of Y in each group are those presented in Table 1.

Bootstrapping the Sampling Distribution

Bootstrapping [6] is typically used in observational studies, where N units (e.g., people) are randomly sampled from k populations, with the k populations distinguished from each other on some (typically, but not necessarily) qualitative dimension.

For example, a political scientist might randomly poll N people from a population by randomly dialing phone numbers, assessing the political knowledge of the respondents in some fashion, and asking them to self-identify as either a Democrat (group 1), a Republican (group 2), or an Independent (group 3). Imagine that a sample of 15 people yielded the data in Table 1, where the outcome variable is the number of 12 questions about current political events that the respondent correctly answered. Using these data, $MS_{\text{between}} = 18.23$, $MS_{\text{within}} = 3.63$, and so $F_{\text{obtained}} = 5.02$. From the F distribution, the P value for the test of the null hypothesis that $\mu_{\text{Democrats}} = \mu_{\text{Republicans}} = \mu_{\text{Independents}}$ is 0.026. So, if the null hypothesis is true, the probability of obtaining an F ratio as large or larger than 5.02 is 0.026. Using $\alpha = 0.05$, the null hypothesis is rejected. It appears that affiliates of different political groups differ in their political knowledge on average.

To generate the P value via bootstrapping, a test statistic θ must first be generated describing the obtained result. To illustrate that any test statistic can be employed, even one without a convenient sampling distribution, I will use $\theta = \sum n_j |(\bar{Y}_j - \bar{Y})|$. So, the obtained result is quantified as the weighted sum of the absolute mean deviations from the grand mean. In the data in Table 1, $\theta_{\text{obtained}} = 22$. The traditional F ratio or any alternative test statistic could be used and the procedure described below would be unchanged. The P value is defined, as always, as the proportion of possible values of θ that are as large or larger than θ_{obtained} in the distribution of possible values of θ , assuming that the null hypothesis is true.

The distribution of possible values of θ is generated by resampling from the $N = 15$ units *with replacement*. There are two methods of bootstrap resampling discussed in the bootstrapping literature. They differ in the assumptions made about the populations sampled, and, as a result, they can produce different results.

The first approach to bootstrapping assumes that the sample of N units represents a set of k independent samples from k distinct populations. No assumption is made that the distribution of Y has any particular form, or that the distribution is identical in each of the k populations. The usual null hypothesis is that these k populations have a common mean.

To accommodate this null hypothesis, the original data must first be transformed before resampling

can occur. This transformation is necessary because the P value is defined as the probability of θ_{obtained} conditioned on the null hypothesis being true, but the data in their raw form may be derived from a world where the null hypothesis is false.

The transformation recommended by Efron and Tibshirani [6, p. 224] is

$$Y'_{ij} = Y_{ij} - \bar{Y}_j + \bar{Y} \quad (2)$$

where Y_{ij} is the value of Y for case i in group j , $\bar{Y}_j$ is the mean Y for group j , and $\bar{Y}$ is the mean of all N values of Y (in Table 1, $\bar{Y} = 6$). Once this transformation is applied, then the group means are all the same, $\bar{Y}'_1 = \bar{Y}'_2 = \dots = \bar{Y}'_j = \bar{Y}$. The result of applying this transformation to the data in Table 1 can be found in Table 2.

To generate the bootstrapped sampling distribution of θ , for each of the k groups, a sample of n_j scores is randomly taken *with replacement* from the transformed scores for group j , yielding a new sample of n_j values. This is repeated for each of the k groups, resulting in a total sample size of N . In the resulting 'resampled' data set, θ^* is computed, where θ^* is the same test statistic used to describe the obtained results in the untransformed data. So $\theta^* = \sum n_j |(\bar{Y}'_j - \bar{Y}')|$. This procedure is repeated, resampling with replacement from the transformed data set (Table 2) a total of b times. The *bootstrapped null sampling distribution of θ* is defined as the b values of θ^* . The P value for θ_{obtained} is calculated as the proportion of values of θ^* in this distribution that are equal to or larger than θ_{obtained} ; that is, $p = \#(\theta^* \geq \theta_{\text{obtained}})/b$.

Because of the random nature of bootstrapping, p is a random variable. A good estimate of the P

value can be derived with values of b as small as 1000, but given the speed of today's computers, there is no reason not to do as many as 10 000 bootstrap resamples. The larger the value of b , the more precise the estimate of p .

The second approach to generating a bootstrap hypothesis test is more nearly like classical ANOVA. We assume that, under the null hypothesis, each of the k samples was drawn from a single common population. We do not assume that this population is normally distributed, only that it is well represented by the sample of N observations. Bootstrap samples for each group can then be taken from this pool of N scores, again by randomly sampling with replacement the requisite number of times. As with the first approach, we build up a bootstrap null sampling distribution by repeating these random samplings a large number, b , of times, computing our group comparison statistic each time.

The choice between these two approaches is dictated largely by group sizes. In the first approach, we assume that each of the k sampled populations is well represented by the sample drawn from that population. This assumption makes greater sense the larger the group size. For small group sizes, we may doubt the relevance of the assumption. If group sizes are smaller than, say, 10 or 12, it is preferable to make the additional assumption of a commonness of the sampled populations.

There is a variety of statistical programs on the market that can conduct this test, as well as some available for free on the Internet. No doubt, as bootstrapping becomes more common, more and more software packages will include bootstrapping routines. Even without a statistical package capable of conducting this test, it is not difficult to program in whatever language is convenient. For example, I wrote a GAUSS program [2] to generate the bootstrapped sampling distribution of θ or the data in Table 1 using the first approach. Setting b to 10 000, 5 of the 10 000 values of θ^* were equal to or greater than 22, yielding an estimated P value of $5/10\,000 = 0.0005$.

The Randomization Test

The bootstrapping approach to generating the P value is based on the idea that each Y value observed in the sample represents only one of perhaps many units in

Table 2 Data in Table 1 after transformation using Equation (2)

Group 1 ($n_1 = 4$)	Group 2 ($n_1 = 6$)	Group 3 ($n_3 = 5$)
4.75	3.00	4.20
3.75	7.00	6.20
6.75	6.00	7.20
8.75	5.00	4.20
	8.00	8.20
	7.00	
$\bar{Y}'_1 = 6.00$ $s_1 = 2.22$	$\bar{Y}'_2 = 6.00$ $s_2 = 1.79$	$\bar{Y}'_3 = 6.00$ $s_3 = 1.79$

the population with that same value of Y . No assumption is made about the form of the distribution of Y , as is traditional in ANOVA, where the assumption is that Y follows a normal distribution. So, the bootstrapped sampling distribution of θ is conditioned only on the values of Y that are known to be possible when randomly sampling from the k populations. The randomization test goes even farther by doing a form of sampling *without* replacement, so each of the N values of Y is used only once. Thus, no assumption is made that each unit in the sample is representative of other units with the same Y score in some larger population.

A randomization test [5] would typically be used when the k groups are generated through the random assignment of N units into k groups, as in an experiment. Although it is ideal to make sure that the random assignment mechanism produces k groups of roughly equal size, this is not mathematically necessary. For example, suppose that rather than randomly sampling 15 people from some population, the 15 people in Table 1 were 15 people that were conveniently available to the political scientist, and each was randomly assigned into one of three experimental conditions. Participants assigned to group 1 were given a copy of the *New York Times* and asked to read it for 30 min. Participants assigned to group 2 were also given a copy of the *New York Times*, but all advertisements in the paper had been blackened out so they couldn't be read. For participants assigned to group 3, the advertisements were highlighted with a bright yellow marker, making them stand out on the page. After the 30-minute reading period, the participants were given a 12-item multiple-choice test covering the news of the previous day. Of interest is whether the salience of the advertisements distracted people from the content of the newspaper, and, thereby, affected how much they learned about the previous day's events. This would typically be assessed by testing whether the three treatment population means are the same.

The randomization test begins with the assumption that, under the null hypothesis, each of the N values of Y observed was in a sense 'preordained.' That is, the value of Y any sampled unit contributed to the data set is the value of Y that unit would have contributed regardless of which condition in the experiment the unit was assigned to. This is the essence of the claim that the manipulation had no effect on the outcome. Under this assumption,

the obtained test statistic θ_{obtained} is only one of many possible values of θ that could have been observed merely through the random assignment of the N units into k groups. A simple combinations formula tells us that there are $N!/(n_1!n_2!\dots n_k!)$ possible ways of randomly assigning N cases (and their corresponding Y values) into k groups of size $n_1, n_2, \dots, n_k$. Therefore, under the assumption that the salience of the advertisements had no effect on learning, there are $15!/(4!6!5!) = 630\,630$ possible values of θ that could be generated with these data, and, thus 630 630 ways this study could have come out merely through the random assignment of sampled units into groups.

In a randomization test, the P value for the obtained result, θ_{obtained} , is generated by computing every value of θ possible given the k group sample sizes and the N values of Y available. The set of $N!/(n_1!n_2!\dots n_k!)$ values of θ possible is called the *randomization distribution of θ* . Using this randomization distribution, the P value for θ_{obtained} is defined as the proportion of values of θ in the randomization distribution that are at least as large as θ_{obtained} , that is, $p = \#(\theta \geq \theta_{\text{obtained}})/[N!/(n_1!n_2!\dots n_k!)]$. Because θ_{obtained} is itself in the randomization distribution of θ , the smallest possible P value is therefore $1/[N!/(n_1!n_2!\dots n_k!)]$.

Using the data in Table 1, and just for the sake of illustration defining θ as the usual ANOVA F ratio, it has already been determined that $\theta_{\text{obtained}} = 5.02$. Using one of the several software packages available that conduct randomization tests or writing special code in any programming language, it can be derived that there are 19 592 values of θ in the randomization distribution of θ that are at least 5.02. So the P value for the obtained result is $19\,592/630\,630 = 0.031$.

In bootstrapping, the number of values of θ in the bootstrapped sampling distribution is b , a number completely under control of the data analyst and generally small enough to make the computation of the distribution manageable. In a randomization test, the number of values of θ in the randomization distribution is $[N!/(n_1!n_2!\dots n_k!)]$. So the size of the randomization distribution is governed mostly by the total sample size N and the number of groups k . The size of the randomization distribution can easily explode to something unmanageable, such that it would be computationally impractical if not entirely infeasible to generate every possible value of θ . For example, merely doubling the sample

size and the size of each group in this study yields $[30!/(8!12!10!)]$ or about 3.8×10^{12} possible values of θ in the randomization distribution that must be generated just to derive a single P value. The generation of that many statistics would simply take too long even with today's (and probably tomorrow's) relatively fast desktop computers.

Fortunately, there is an easy way around this seemingly insurmountable problem. Rather than generating the full randomization distribution of θ , an *approximate randomization distribution* of θ can be generated by randomly sampling from the possible values of θ in the full randomization distribution. When an approximation of the full randomization distribution is used to generate the P value for θ_{obtained} , the test is known as an *approximate randomization test*. In most real-world applications, the sheer size of the full randomization distribution dictates that you settle for an approximate randomization test. It is fairly easy to randomly sample from the full randomization distribution using available software or published code, for example, [4] and [7], or programming the test yourself. The random sampling is accomplished by randomly reassigning the Y values to the N units and then computing θ after this reshuffling of the data. This reshuffling procedure is repeated $b - 1$ times. The b th value of θ in the approximate randomization distribution is set equal to θ_{obtained} , and the P value for θ_{obtained} defined as $\#(\theta \geq \theta_{\text{obtained}})/b$. Because θ_{obtained} is one of the b values in the approximate randomization distribution, the smallest possible P value is $1/b$. Of course, the P value is a random variable because of the random nature of the random selection of possible values of θ . The larger b , the less the P value will vary from the exact P value (from using the full randomization distribution). For most applications, setting b between 5000 and 10 000 is sufficient. Little additional precision is gained using more than 10 000 randomizations.

The randomization test is sometimes called a *permutation test*, although some reserve that term to refer to a slightly different form of the randomization test, where the groups are randomly sampled from k larger and distinct populations. Under the null hypothesis assumption that these k populations are identical, the Y observations are exchangeable among populations (meaning simply that any Y score could be observed in any group with the same probability). Under this

null hypothesis, the P value can be constructed as described above for the randomization test.

The Kruskal–Wallis Test as a Randomization Test on Ranks

When you are interested in comparing a set of groups measured on some quantitative variable but the data are not normally distributed, many statistics textbooks recommend the **Kruskal–Wallis test**. The Kruskal–Wallis test is conducted by converting each Y value to its ordinal rank in the distribution of N values of Y . The sum of the ranks within group are then transformed into a test statistic H . For small N , H is compared to a table of critical values to derive an approximate P value. For large N , the P value is derived in reference to the χ^2 distribution with $k - 1$ degrees of freedom [8]. It can be shown that the Kruskal–Wallis test is simply a randomization test as described above, but using the ranks rather than the original Y values. With large N , the P value is derived by capitalizing on the fact that the full randomization distribution of H when analyzing ranks is closely approximated by the χ^2 distribution.

Assumptions and Selection of Test Statistic

Bootstrapping and randomization tests make fewer and somewhat weaker assumptions than does classical ANOVA about the source of the observed Y scores. Classical ANOVA assumes that the j th group of scores was randomly sampled from a normal distribution. Further, each of the k normal populations sampled is assumed to have a common variance, and, under the null hypothesis, a common mean.

The first bootstrapping approach assumes that the j th group of scores was randomly sampled from a population, of unknown form, that is well represented by the sample from that population. The k populations may differ in variability or shape, but under the null hypothesis, they are assumed to have a common mean. The second bootstrapping approach assumes that the k populations sampled potentially differ only in location (i.e., in their means). Under the null hypothesis of a common mean as well, the k samples can be treated as having been sampled from the same population, one well represented by the aggregate of the k samples.

Bootstrapping shares with the ANOVA a random sampling model of chance. The value of our test statistic will vary from one random sample of N units to another. By contrast, the model of chance for the randomization test does not depend on random sampling, only on the randomization to groups of a set of available units. The value of our test statistic now varies with the random allocation of this particular set of N units.

The two models of chance differ in their range of statistical inference. Random sampling leads to inferences about the larger populations sampled. Randomization-based inference is limited to the particular set of units that were randomized. However, randomization also leads to causal inference. Establishing causality, even over a limited number of units, can be more important than population inference. And, many behavioral science experiments are conducted using nonrandom collections of units.

ANOVA is often applied to the analysis of outcomes that clearly are not normal on the grounds that ANOVA is relatively 'robust' to violations of that assumption. For example, much behavioral science research is based on outcome variables that could not possibly be normally distributed, such as counts or responses that have an upper or lower bound such as rating scales. But why make assumptions in your statistical method you don't need to make in the first place, and that are almost certainly not warranted when there are methods available that work as well, can test the hypothesis of interest, and don't make those unrealistic assumptions?

Large departures from the assumptions, notably the assumption of variance equality, change the null hypothesis tested in ANOVA to equality of the population distributions rather than just equality of their means. So a statistically significant F ratio can result from differences in population variance, shape, mean, or any combination of these factors, depending on which assumption is violated. The problem is that the test statistic typically used in ANOVA based on a ratio of between-group variability to a pooled estimate of within-group variability is sensitive to group differences in variance as well as group differences in mean. For this reason, alternative test statistics such as Welch's F ratio and a corresponding degrees of freedom adjustment [10] as well as other methods, for example, [1] and [3], have been advocated as alternatives to traditional ANOVA

that allow you to relax the assumption of equality of variance.

So, the null hypothesis tested with ANOVA is a function of the test statistic used, and whether the assumptions used in the derivation of the test statistic are met. And the same can be said of resampling methods. If you cannot assume or have reason to reject the assumption of equality of the populations in variance and shape, the use of the traditional F ratio in ANOVA produces a test of equality of population distributions rather than population means. Similarly, the randomization test and the second version of the bootstrap described here using the pooled variance F ratio as the test statistic, or any monotonically related statistic such as the between groups sum of squares or mean square, yields a test of the null hypothesis of equality of the population distributions, unless you make some additional assumptions. In the case of the second approach to bootstrapping described earlier, if you desire a test of mean differences, you must also assume equality of the population distributions on 'nuisance' parameters such as **skew**, **kurtosis**, and variance.

In the case of randomization tests, you must assume that the manipulation does not affect the responses in any way whatsoever in order to test the null hypothesis of equality of the means. For example, if you have reason to believe that the manipulation might affect variability in the outcome variable, then you have no basis for interpreting a significant P value from a randomization test as evidence that the manipulation affects the means, because the F ratio quantifies departures from equality of the response distributions, not just the means.

The first bootstrapping method does not suffer near as much from this problem of choice of test statistic, although it is not immune to it. By conditioning the resampling for each group on the observed Y units in that group, the effect of population differences in variance and shape on the validity of the test as a test of group mean differences is reduced considerably. However, this method requires that you have substantial faith in each group's sample as representing the distribution of the population from which it was derived. The validity of this method of bootstrapping is compromised when one or more of the group sample sizes is small.

So even when using a resampling method, if you desire to test a precise null hypothesis about the group means, it is necessary to use a test statistic that is

sensitive only to mean differences. It is not always obvious what a test statistic is sensitive to without some familiarity with the derivation of and research on that statistic. Fortunately, as discussed above, there are many alternative test statistics developed for ANOVA that perform better when population variances differ [1, 3], or [9]. For example, the numerator of Welch's F ratio [10, also see 9] would be a sensible test statistic for comparing population means, given that it is much less sensitive to population differences in variability than the usual ANOVA F ratio. This test statistic is $\theta = \sum n_j/s_j^2 (\bar{Y}_j - \tilde{Y})^2$, where $\tilde{Y}$ is defined as $w^{-1} \sum n_j \bar{Y}_j / s_j^2$ and $w = \sum n_j / s_j^2$. But, resampling methods are a relatively new methodological tool, and continued research on their performance through simulation will no doubt clarify which test statistics are most suitable for which null hypothesis under which conditions.

References

- [1] Alexander, R.A. & Govern, D.M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity, *Journal of Educational Statistics* **19**, 91–101.
- [2] Aptech Systems. (2001). *GAUSS version 5.2* [computer program]. Maple Valley.
- [3] Brown, M.B. & Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of means, *Technometrics* **16**, 129–132.
- [4] Chen, R.S. & Dunlap, W.P. (1993). SAS procedures for approximate randomization tests, *Behavior Research Methods Instruments & Computers* **25**, 406–409.
- [5] Edgington, E. (1995). *Randomization Tests*, 3rd Edition, Dekker, New York.
- [6] Efron, B. & Tibshirani, R.J. (1991). *An Introduction to the Bootstrap*, Chapman & Hall, Boca Raton.
- [7] Hayes, A.F. (1998). SPSS procedures for approximate randomization tests, *Behavior Research Methods Instruments & Computers* **30**, 536–543.
- [8] Howell, D. (1997). *Statistical Methods for Psychology*, 4th Edition, Duxbury, Belmont.
- [9] James, G.S. (1951). The comparison of several groups of observations when the ratios of population variances are unknown, *Biometrika* **38**, 324–329.
- [10] Welch, B.L. (1951). On the comparison of several mean values: an alternative approach, *Biometrika* **38**, 330–336.

ANDREW F. HAYES

Optimal Design for Categorical Variables

MARTIJN P.F. BERGER

Volume 3, pp. 1474–1479

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Optimal Design for Categorical Variables

Introduction

Optimal designs enable behavioural scientists to obtain efficient parameter estimates and maximum power of their tests, thus reducing the required sample size and the costs of experimentation. **Multiple linear regression analysis** is one of the most frequently used statistical techniques to analyse the relation between quantitative and qualitative variables. Unlike quantitative variables, qualitative variables do not have a well-defined measurement scale and only distinguish categories, with or without an ordering. Nominal independent variables have no ordering among the categories, and **dummy variables** are usually used to describe their variation. Ordinal independent variables are usually numbered in accordance with the ordering of the categories and treated as if they are quantitative. The dependent variable can also be quantitative or qualitative. Although optimal design theory is applicable to models with both quantitative and qualitative variables, the results are usually quite different.

In this paper, we will briefly consider four cases, namely, optimal designs for **linear models** and **logistic regression** models, with and without dummy coding. First, the methodological basics of optimal design theory will be described by means of a linear regression model. Then, optimal designs for each of the four cases will be illustrated with concrete examples.

Finally, we will draw some conclusions and provide references for further reading.

Methodological Basics for Optimal Designs

Model. In describing the relationship of Y to the quantitative variables X_1 and X_2 , we consider the linear model for subject i :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad (1)$$

where β_0 is the intercept, and β_1 and β_2 are the usual effect parameters. This model assumes that the errors ε_i are independently normally distributed with mean zero and variance σ^2 , that is, $\varepsilon_i \sim N(0, \sigma^2)$.

Optimal Design. In the design stage of a study, we are free to sample subjects with different values for X_1 and X_2 . These choices influence the variance of the estimated effect parameters $\beta' = [\beta_1 \beta_2]$. Although the intercept may also be of interest, we will restrict ourselves to these two effect parameters. Because we usually want to estimate β_1 and β_2 as efficiently as possible, an ideal design would be a design with minimum variance of the estimated parameters. Such a design is called an *optimal design*. The variances and covariance of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ in model (1) are given by:

$$\text{var}(\hat{\beta}) = \sigma^2 \begin{bmatrix} \frac{1}{N(1-r_{12}^2)\text{var}(X_1)} & -\frac{\text{cov}(X_1, X_2)}{N(1-r_{12}^2)W} \\ -\frac{\text{cov}(X_1, X_2)}{N(1-r_{12}^2)W} & \frac{1}{N(1-r_{12}^2)\text{var}(X_2)} \end{bmatrix}, \quad (2)$$

where $W = \{\text{var}(X_1) \times \text{var}(X_2)\}$, and $\text{var}(X_1)$ and $\text{var}(X_2)$ are the variances of X_1 and X_2 , respectively, and $\text{cov}(X_1, X_2)$ and r_{12} are the corresponding covariance and correlation. $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ are on the main diagonal of the matrix in (2), and the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ is the off-diagonal element. Equation (2) shows that the $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ decrease as:

1. The variances of the variables $\text{var}(X_1)$ and $\text{var}(X_2)$, respectively, increase. This can be done by only selecting subjects with the most extreme values for X_1 and X_2 .
2. The correlation r_{12} between X_1 and X_2 decreases. For an orthogonal design, the correlation r_{12} is minimum, that is, $r_{12} = 0$.
3. The common error variance σ^2 decreases, that is, when more precise measurements of Y are obtained.
4. The total number of observations N increases. This is, in fact, the most often applied method to obtain sufficient **power** for finding real effects, but with extra costs of collecting data.

There is a trade-off among these effects. For example, reduction of measurement errors by a factor $(1 - (1/\sqrt{2})) = 0.29$ has the same effect on $\text{var}(\hat{\beta}_1)$ as doubling the total sample size N .

2 Optimal Design for Categorical Variables

Optimality Criterion. Although $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ can be minimized separately, model (1) is usually applied because interest is focused on the joint prediction of Y by X_1 and X_2 , and because hypotheses on both the regression parameters in β are tested. Moreover, the overall test for the fit of model (1) is based on all parameters in β , simultaneously. Thus, an optimal design for model (1) should take all the variances and covariance into account, simultaneously. This can be done by means of an optimality criterion. Different optimality criteria have been proposed in the literature. See [4] for a review. We will restrict ourselves to the determinant or D-optimality criterion, because of its advantages. This criterion is proportional to the volume of the simultaneous confidence region for the parameters in $\beta' = [\beta_1 \beta_2]$, thus giving it a similar interpretation as that of a **confidence interval** for a single parameter. Moreover, a D-optimal design is invariant under linear transformation of the scale of the independent variables, which is generally not true for the other criteria.

Consider several designs τ in a design space T , which are distinguished by the number of different levels of the independent variables X_1 and X_2 . A D-optimal design $\tau^* \in T$ is the design among all designs $\tau \in T$, for which the determinant of $\text{var}(\hat{\beta})$ is minimized:

$$\begin{aligned} \text{Minimize :} & \quad \det[\text{var}(\hat{\beta})] \quad \text{or} \\ \text{Minimize:} & \quad \frac{[\sigma^2]^2}{N^2(1 - r_{12}^2)\text{var}(X_1)\text{var}(X_2)}, \end{aligned} \quad (3)$$

Equation (3) shows that $\det[\text{var}(\hat{\beta})]$ is not only a function of N , but also of σ^2 , $\text{var}(X_1)$, $\text{var}(X_2)$, and of the correlation r_{12} . Again, we can see that, for example, doubling the sample size N has the same effect on the efficiency as reducing the error variance σ^2 by $1/2$, that is, by decreasing the errors ε_i by a factor of about 0.29.

Maximum power and efficiency. The $100(1 - \alpha)\%$ confidence ellipsoid for the p regression parameters in β is given by:

$$(\beta - \hat{\beta})' \text{var}(\hat{\beta})^{-1} (\beta - \hat{\beta}) \leq pMS_e F(df_1, df_2, \alpha), \quad (4)$$

where MS_e is an estimate of the error variance σ^2 and $F(df_1, df_2, \alpha)$ is the α point of the F distribution with df_1 and df_2 degrees of freedom.

Since $\det[\text{var}(\hat{\beta})]$ is proportional to the volume of this ellipsoid, minimizing this criterion will maximize the **power** of the simultaneous test of the hypothesis $\beta = 0$. Thus, a D-optimal design τ^* will result in maximum efficiency and power for a given sample size and estimate of the error variance σ^2 .

Relative efficiency. To compare the efficiency of a design τ with that of the optimal design τ^* , the following relative efficiency can be used:

$$\text{Eff}(\tau) = \left\{ \frac{\det[\text{var}(\hat{\beta}_{\tau^*})]}{\det[\text{var}(\hat{\beta}_{\tau})]} \right\}^{1/p}, \quad (5)$$

where p is the number of independent parameters in the parameter vector β . This ratio is always $\text{Eff}(\tau) < 1$, and indicates how much additional observations are needed to obtain maximum efficiency. If, for example, $\text{Eff}(\tau) = 0.8$, then $(0.8^{-1} - 1)100\% = 25\%$ more observations will be needed for design τ to become as efficient as the optimal design, and when each extra observation costs the same, then 25% cost reduction is feasible.

Examples

Multiple Regression with Ordinal Independent Variables

Consider, as an example, a study on the acquisition of new vocabulary. The design consists of four samples of pupils from the 8th, 9th, 10th, and 11th grades, respectively. Suppose that vocabulary growth also depends on IQ, and that three IQ classes are considered, namely, Class 1 (IQ range 90–100), Class 2 (IQ range 100–110), and Class 3 (IQ range 110–120). In total, 12 distinct samples of pupils are available. A similar study was reported by [9, pp. 453–456], but with a different design.

The vocabulary test score Y_i of pupil i can be described by the linear regression model with ordinal independent variables *Grade* and *IQ* class:

$$Y_i = \beta_0 + \beta_1 \text{Grade}_i + \beta_2 \text{IQ}_i + \varepsilon_i, \quad (6)$$

where β_0 is the intercept, β_1 and β_2 are the regression parameters for *Grade* and *IQ* class, respectively, and $\varepsilon_i \sim N(0, \sigma^2)$. We will treat the ordinal independent variables as being quantitative.

Five different designs for model (1) are displayed in Table 1. These designs are all balanced and based

Table 1 Five designs with 12 combinations of the levels of X_1 and X_2 for the vocabulary study

	Design τ_1		Design τ_2		Design τ_3		Design τ_4		Design τ_5	
	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
1	8	1	8	1	8	1	8	1	8	1
2	8	2	8	2	8	1	8	1	8	1
3	8	3	8	3	8	3	8	1	8	1
4	9	1	8	1	8	1	8	1	8	3
5	9	2	8	2	8	2	8	1	8	3
6	9	3	8	3	8	3	8	1	8	3
7	10	1	11	1	11	1	11	3	11	1
8	10	2	11	2	11	2	11	3	11	1
9	10	3	11	3	11	3	11	3	11	1
10	11	1	11	1	11	1	11	3	11	3
11	11	2	11	2	11	3	11	3	11	3
12	11	3	11	3	11	3	11	3	11	3
var(X)	1.25	0.6667	2.25	0.6667	2.25	0.8333	2.25	1.00	2.25	1.00
r_{12}	0.0000		0.0000		0.1833		1.0000		0.0000	
det[$\text{var}(\hat{\beta})$]	1.20		0.6667		0.5517		∞		0.4444	

Note: The D-optimality criterion is computed without N and σ^2 , that is, $\text{det}[\text{var}(\hat{\beta})]$ is divided by N^2/σ^4 .

on the 12 combinations of the levels of *Grade* and *IQ*, that is, the total sample size is $N = 12n$, where n is the number of pupils in each combination. Design τ_1 is the original design. Design τ_5 is the D-optimal design with minimum value for $\text{det}[\text{var}(\hat{\beta})] = 0.444$. As expected, this design has maximum variances $\text{var}(X_1)$ and $\text{var}(X_2)$, and minimum value $r_{12} = 0$. The relative efficiency of Design τ_1 is $\text{Eff}(\tau_1) = 0.6085$, indicating that Design τ_1 would need about 64% more observations to have maximum efficiency and maximum power. Designs τ_2 and τ_3 are both more efficient than Design τ_1 . Notice that Design τ_4 is not appropriate for this model because $r_{12} = 1$.

Multiple Regression with Nominal Independent Variables

Suppose that in the vocabulary growth study, the joint effect of *Grade* and three different *Instruction Methods* is investigated, and that two dummy variables Dum_1 and Dum_2 are used to describe the variation of the *Instruction Methods*. The regression model is:

$$Y_i = \beta_0 + \beta_1 \text{Grade}_i + \beta_2 \text{Dum}_{1i} + \beta_3 \text{Dum}_{2i} + \varepsilon_i. \quad (7)$$

In Table 2, the D-optimal design for model (7) is presented with n subjects for six optimal combinations of *Grade* levels and *Instruction Method* and

minimum value $\text{det}[\text{var}(\hat{\beta})] = 1.6875$. Three unequal group designs are also presented in Table 2, and it can be seen that efficiency loss is relatively small for these designs because $\text{Eff}(\tau) > 0.9$.

Logistic Model with Ordinal Independent Variables

Occasionally, the dependent variable may be dichotomous. For example, the following logistic model with *Grade* as predictor describes the probability P_i that subject i passes a vocabulary test:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 \text{Grade}_i. \quad (8)$$

The parameters β_0 and β_1 can be estimated by **maximum likelihood**, and a D-optimal design for model (8) can be found by minimizing the determinant of $\text{var}(\hat{\beta})$ (see [4], Chapter 18, for details). There is, however, a problem. For nonlinear models, $\text{var}(\hat{\beta})$ depends on the actual parameter values, and a D-optimal design for one combination of parameter values may not be optimal for other values. This problem is often referred to as the *local optimality problem* (see [4] for ways to overcome this problem).

It can be shown (see [5]) that a D-optimal design for model (8) has two distinct, equally weighted design points, namely, $x_i = \pm(1.5437 - \beta_0)/\beta_1$. This

4 Optimal Design for Categorical Variables

Table 2 The effect of unbalancedness on efficiency of a design for regression model (7) with quantitative variable *Grade* X_1 and qualitative variable *Instruction Method* X_2

Cells	Levels		Design weights			
	X_1	X_2	W_1	W_2	W_3	W_4
1	8	1	$\frac{1}{6}N$	$\frac{1}{4}N$	$\frac{1}{4}N$	$\frac{1}{5}N$
2	8	2	$\frac{1}{6}N$	$\frac{1}{6}N$	$\frac{1}{4}N$	$\frac{1}{10}N$
3	8	3	$\frac{1}{6}N$	$\frac{1}{12}N$	$\frac{1}{12}N$	$\frac{1}{5}N$
4	11	1	$\frac{1}{6}N$	$\frac{1}{4}N$	$\frac{1}{6}N$	$\frac{1}{5}N$
5	11	2	$\frac{1}{6}N$	$\frac{1}{6}N$	$\frac{1}{6}N$	$\frac{1}{10}N$
6	11	3	$\frac{1}{6}N$	$\frac{1}{12}N$	$\frac{1}{12}N$	$\frac{1}{5}N$
$\det[\text{var}(\hat{\beta})]$			1.6875	2.25	2.2345	1.9531
$\text{Eff}(\tau)$			1.0000	0.9086	0.9107	0.9524

Note: $\text{Eff}(\tau)$ is computed with $p = 3$ variables, that is, X_1 and two dummy variables for X_2 .

means that optimal estimation of β_0 and β_1 is obtained if $N/2$ pupils have a $P_i = 0.176$, and $N/2$ pupils have a $P_i = 0.824$ of passing the test. Of course, such a selection is often not feasible in practice. But, for a test designed for the 9th grade, a teacher may be able to select $N/2$ pupils from the 8th grade with about $P_i = 0.2$ and $N/2$ pupils from the 10th grade with about $P_i = 0.8$ probability of passing the test, respectively. Such a sample would be more efficient than one sample of N pupils from the 9th grade. The well known **item response theory (IRT)** models in educational and psychological measurement are comparable with model (8), and the corresponding optimal designs were studied by [5] and [12], among others.

Logistic Regression with Nominal Independent Variables

As already mentioned, an optimal design for logistic regression is only locally optimal, that is, for a given set of parameter values. If the probability of passing a vocabulary test is modeled with *Instruction Method* as only predictor, then the logistic regression model is:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 \text{Dum}_1 + \beta_2 \text{Dum}_2, \quad (9)$$

where Dum_1 and Dum_2 are two dummy variables needed to describe the variation of the three Instruction Methods. For parameter values $(\beta_0, \beta_1, \beta_2) \in [-10, 10]$, the D-optimal design for model (9) is a balanced design with an equal number of pupils assigned to each instruction method. It should, however, be emphasized that for other models or parameter values, a D-optimal design may not be balanced.

If the researcher expects that one of the instruction methods works better than the others and assigns 70% of the sample to this instruction method and divides the remaining pupils equally over the two other methods, then the $\text{Eff}(\tau) = 0.752$. This means that the sample size would have to be 33% larger to remain as efficient as the optimal design.

Summary and Further Reading

Although different models have different optimal designs, a rule-of-thumb may be that if the independent variables are treated as being quantitative, efficiency will increase if the variance of the independent variables increases, and if the design becomes more orthogonal. For qualitative independent variables, a design with equal sample sizes for all categories is often optimal. For non-linear models, the

optimal designs are only optimal locally, that is, for a given set of parameter values.

The number of studies on optimal design problems is increasing rapidly. Elementary reviews are given by [13] and [19]. A more general treatment is given by [4]. Optimal designs for correlated data were studied by [1] and [8]. Studies on optimal designs for random block effects are found in [3, 10], and [11], while [14] and [18] studied random effects models. Optimal designs for multilevel and random effect models with covariates are presented by [15, 16], and [17], among others. Optimal design for IRT models in educational measurement is studied by [2, 5, 6, 7] and [2].

References

- [1] Abt, M., Liski, E.P., Mandal, N.K. & Sinha, B.K. (1997). Correlated model for linear growth: optimal designs for slope parameter estimation and growth prediction, *Journal of Statistical Planning and Inference* **64**, 141–150.
- [2] Armstrong, R.D., Jones, D.H. & Kuncze, C.S. (1998). IRT test assembly using network-flow programming, *Applied Psychological Measurement* **22**, 237–247.
- [3] Atkins, J.E. & Cheng, C.S. (1999). Optimal regression designs in the presence of random block effects, *Journal of Statistical Planning and Inference* **77**, 321–335.
- [4] Atkinson, A.C. & Donev, A.N. (1992). *Optimum Experimental Designs*, Clarendon Press, Oxford.
- [5] Berger, M.P.F. (1992). Sequential sampling designs for the two-parameter item response theory model, *Psychometrika* **57**, 521–538.
- [6] Berger, M.P.F. (1998). Optimal design of tests with dichotomous and polytomous items, *Applied Psychological Measurement* **22**, 248–258.
- [7] Berger, M.P.F., King, C.Y. & Wong, W.K. (2000). Minimax D-optimal designs for item response theory models, *Psychometrika* **65**, 377–390.
- [8] Bishoff, W. (1993). On D-optimal designs for linear models under correlated observations with an application to a linear model with multiple response, *Journal of Statistical Planning and Inference* **37**, 69–80.
- [9] Bock, R.D. (1975). *Multivariate Statistical Methods*, McGraw-Hill, New York.
- [10] Cheng, C.S. (1995). Optimal regression designs under random block-effects models, *Statistica Sinica* **5**, 485–497.
- [11] Goos, P. & Vandebroek, M. (2001). D-optimal response surface designs in the presence of random block effects, *Computational Statistics and Data Analysis* **37**, 433–453.
- [12] Jones, D.H. & Jin, Z. (1994). Optimal sequential designs for on-line estimation, *Psychometrika* **59**, 59–75.
- [13] McClelland, G.H. (1997). Optimal design in psychological research, *Psychological Methods* **2**, 3–19.
- [14] Mentré, F., Mallet, A. & Baccar, D. (1997). Optimal design in random effects regression models, *Biometrika* **84**, 429–442.
- [15] Moerbeek, M., Van Breukelen, G.J.P. & Berger, M.P.F. (2001). Optimal experimental designs for multilevel models with covariates, *Communications in Statistics: Theory and Methods* **30**(12), 2683–2697.
- [16] Ouwens, J.N.M., Tan, F.E.S. & Berger, M.P.F. (2001). On the maximin designs for logistic random effects models with covariates, in *New Trends Models in Statistical Modelling, The Proceedings of the 16th International Workshop on Statistical Modelling*, B. Klein & L. Korshom, ed., Odense, Denmark, pp. 321–328.
- [17] Ouwens, J.N.M., Tan, F.E.S. & Berger, M.P.F. (2002). Maximin D-optimal designs for longitudinal mixed effects models, *Biometrics* **58**, 735–741.
- [18] Tan, F.E.S. & Berger, M.P.F. (1999). Optimal allocation of time points for the random effects model, *Communications in Statistics: Simulations* **28**(2), 517–540.
- [19] Wong, W.K. & Lachenbruch, P.A. (1996). Designing studies for dose response, *Statistics in Medicine* **15**, 343–360.

MARTIJN P.F. BERGER

Optimal Scaling

YOSHIO TAKANE

Volume 3, pp. 1479–1482

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Optimal Scaling

Suppose that (dis)similarity judgments are obtained between a set of stimuli. The dissimilarity between stimuli may be represented by a Euclidean distance model (see **Multidimensional Scaling**). However, it is rare to find the observed dissimilarity data measured on a ratio scale (see **Scales of Measurement**). It is more likely that the observed dissimilarity data satisfy only the ordinal scale level. That is, they are only monotonically related to the underlying distances (see **Monotonic Regression**). In such cases, we may consider transforming the observed data monotonically, while simultaneously representing the transformed dissimilarity data by a distance model. This process of simultaneously transforming the data, and representing the transformed data, is called optimal scaling [2, 8].

Let δ_{ij} denote the observed dissimilarity between stimuli i and j measured on an ordinal scale. Let d_{ij} represent the underlying Euclidean distance between the two stimuli represented as points in an A -dimensional Euclidean space. Let x_{ia} denote the coordinate of stimulus i on dimension a . Then, d_{ij} can be written as $d_{ij} = \left\{ \sum_{a=1}^A (x_{ia} - x_{ja})^2 \right\}^{1/2}$. (We use X to denote the matrix of x_{ia} , and sometimes write d_{ij} as $d_{ij}(X)$ to explicitly indicate that d_{ij} is a function of X .) Optimal scaling obtains the best monotonic transformation (m) of the observed dissimilarities (δ_{ij}), and the best representation ($d_{ij}(X)$) of the transformed dissimilarity ($m(\delta_{ij})$), in such a way that the squared discrepancy between them is as small as possible. Define the least squares criterion, $\phi = \sum_{i,j} (m(\delta_{ij}) - d_{ij}(X))^2$. We minimize this criterion with respect to m and X under suitable normalization restrictions on m or on X . This is called nonmetric multidimensional scaling [3], which played an important role in the development of ideas of optimal scaling.

We give an example of optimal scaling from nonmetric multidimensional scaling (MDS) (see **Multidimensional Scaling**). Rothkopf [4] reported stimulus confusion data among 36 Morse code signals. Shepard [5] analyzed his data by nonmetric MDS that allowed a monotonic transformation of the confusion probabilities, and a representation of the transformed data in a multidimensional Euclidean space. Figure 1 displays the derived

stimulus configuration. From this, we may deduce that the process mediating confusions among the signals is two-dimensional; one is the total number of components in the signals, and the other is the mixture rate of two kinds (dots and dashes) of components. Signals having more components tend to be located toward the top, and those having more dots tend to be located toward the left of the configuration. Figure 2 displays the optimal inverse monotonic transformation of the confusion probabilities. The derived optimal transformation looks very much like a negative exponential function, $p_{ij} = a \exp(-d_{ij})$, or possibly a Gaussian, $p_{ij} = a \exp(-d_{ij}^2)$, typically found in stimulus generalization data.

In the above example, the data involved are dissimilarity data, for which the distance model may be an appropriate choice. Other kinds of data may also be considered for optimal scaling. For example, the data may reflect the joint effects of two or more underlying factors. In this case, an **analysis of variance**-like additive model may be appropriate. As another example, preference judgments are obtained from a single subject on a set of objects (e.g., cars) characterized by a set of features (size, color, gas efficiency, roominess, etc.) In this case, a **regression-like linear model** that combines these features to predict the overall preference judgments, may be appropriate. As a third example, preference judgments are obtained from a group of subjects on a set of stimuli. In this case, a vector model of preference may be appropriate, in which subjects are represented as vectors, and stimuli as points in a multidimensional space, and subjects' preferences are obtained by projecting the stimulus points onto the subject vectors. This leads to a **principal component analysis**-like bilinear model [7]. Alternatively, subjects' ideal stimuli may be represented as (ideal) points, and it may be assumed that subjects' preferences are inversely related to the distances between subjects' ideal points and actual stimulus points. This is called unfolding (or ideal point) model [1] (see **Multidimensional Unfolding**).

Any one of the models described above may be combined with various types of data transformations, depending on the scale level on which observed data are assumed to be measured. Different levels of measurement scale allow different types of transformations, while preserving the essential properties of the information represented by numbers. In psychology, four major scale levels have traditionally

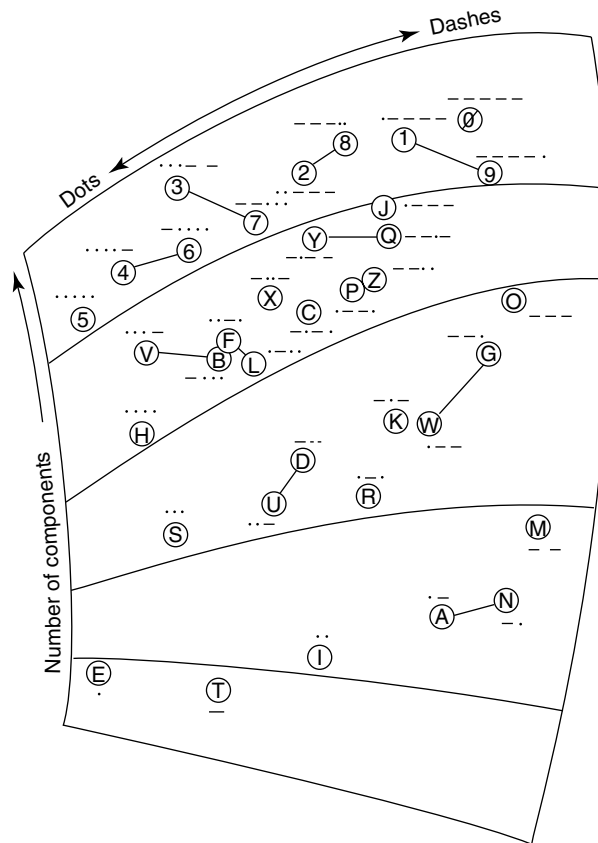


Figure 1 A stimulus configuration obtained by nonmetric multidimensional scaling of confusion data [4] among Morse code signals

been distinguished: nominal, ordinal, interval, and ratio [6]. In the nominal scale level, only the identity of numbers is considered meaningful (i.e., $x = y$ or $x \neq y$). Telephone numbers and gender (males and females coded as 1's and 0's) are examples of this level of measurement. In the nominal scale, any one-to-one transformation is permissible, since it preserves the identity (and nonidentity) between numbers. (This is called admissible transformation.) In the ordinal scale level, an ordering property of numbers is also meaningful (i.e., for x and y such that $x \neq y$, either $x > y$ or $x < y$, but how much larger or smaller is not meaningful). An example of this level of measurement, is the rank number given to participants in a race. In the ordinal scale, any monotonic (or order-preserving) transformation is admissible. In the interval scale level, the difference between two numbers is also meaningful. A difference in temperature

measured on an interval scale can be meaningfully interpreted (e.g., the difference between yesterday's temperature and today's is such and such), but because the origin (zero point) in the scale is arbitrary as the temperature is measured in Celsius or Fahrenheit, their ratio is not meaningful. In the interval scale, any affine transformation ($x' = ax + b$) is admissible. In the ratio scale level, a ratio between two numbers is also meaningful (e.g., temperature measured on the absolute scale where -273°C is set as the zero point). In the ratio scale, any similarity transformation ($x' = ax$) is admissible. In optimal scaling, a specific transformation of the observed data is sought, within each class of the admissible transformations consistent with the scale level on which observed data are assumed to be measured.

It is assumed that one of these transformations is tacitly applied in a data generation process. For

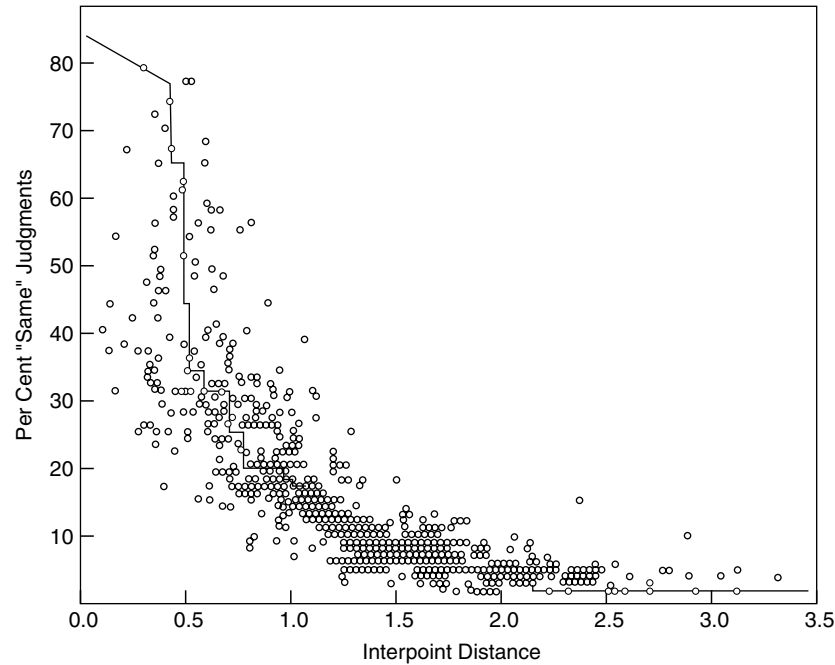


Figure 2 An optimal data transformation of the confusion probabilities

example, if observed dissimilarity data are measured on an ordinal scale, the model prediction, d_{ij} , is assumed error-perturbed, and then monotonically transformed to obtain the observed dissimilarity data, δ_{ij} . Optimal scaling reverses this operation by first transforming back δ_{ij} to the error-perturbed distance by m , which is then represented by the distance model, $d_{ij}(X)$.

References

- [1] Coombs, C.H. (1964). *A Theory of Data*, Wiley, New York.
- [2] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester.
- [3] Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**, 1–27.
- [4] Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired associate learning, *Journal of Experimental Psychology* **53**, 94–101.
- [5] Shepard, R.N. (1963). Analysis of proximities as a technique for the study of information processing in man, *Human Factors* **5**, 19–34.
- [6] Stevens, S.S. (1951). Mathematics, measurement, and psychophysics, in *Handbook of Experimental Psychology*, S.S. Stevens, ed., Wiley, New York, pp. 1–49.
- [7] Tucker, L.R. (1960). Intra-individual and inter-individual multidimensionality, in *Psychological Scaling: Theory and Applications*, H. Guliksen & S. Messick, eds., Wiley, New York, pp. 155–167.
- [8] Young, F.W. (1981). Quantitative analysis of qualitative data, *Psychometrika* **46**, 357–388.

YOSHIO TAKANE

Optimization Methods

ROBERT MICHAEL LEWIS AND MICHAEL W. TROSSET

Volume 3, pp. 1482–1491

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Optimization Methods

Introduction

The goal of optimization is to identify a best element in a set of competing possibilities. Optimization problems permeate statistics. Consider two familiar examples:

1. **Maximum likelihood estimation.** Let y denote a random sample, drawn from a joint probability density function p_θ , where $\theta \in \Theta$. Then, a maximum likelihood estimate of θ is any element of Θ at which the likelihood function, $L(\theta) = p_\theta(y)$, is maximal.
2. **Nonlinear regression.** Let $r_\theta(\cdot)$ denote a family of regression functions, where $\theta \in \Theta$. At input z_i , we observe output y_i . Then, a least squares estimate of θ is any element of Θ at which the sum of the squared residuals,

$$\sum_i [y_i - r_\theta(z_i)]^2, \quad (1)$$

is minimal.

Formally, let $f: X \rightarrow \mathbb{R}$ denote a real-valued *objective function* that measures the quality of elements of the set X . We will discuss the minimization of f ; this is completely general, as maximizing f is equivalent to minimizing $-f$.

Let $C \subseteq X$ denote the *feasible subset* that specifies which possibilities are to be considered. Then the *global optimization problem* is the following:

$$\text{Find } x_* \in C \text{ such that } f(x_*) \leq f(x) \text{ for all } x \in C. \quad (2)$$

Any such x_* is a *global minimizer*, and $f(x_*)$ is the *global minimum*.

In most cases, global optimization is extremely difficult. If X is a topological space, that is, if it is meaningful to speak of neighborhoods of $x \in X$, then we can pose the easier *local optimization problem*:

$$\text{Find } x_* \in C \text{ and a neighborhood } U \text{ of } x_* \text{ such that} \\ f(x_*) \leq f(x) \text{ for all } x \in C \cap U. \quad (3)$$

Any such x_* is a *local minimizer* and $f(x_*)$ is a *local minimum*.

Figure 1 illustrates local minimizers for functions of one and two variables. Figure 1(a) displays an objective function without constraints, i.e., with feasible subset $C = X = \mathbb{R}$. There are two local minimizers, indicated by asterisks. The local minimizer indicated by the asterisk on the right is also a global minimizer. Figure 1(b) displays an objective function with a square feasible subset. At each corner of the square, the function is decreasing, and one could find points with smaller objective function values by leaving the feasible subset. At each corner, however, the objective function is smaller at the corner than at nearby points within the square. Accordingly, each corner is a local constrained minimizer.

We will emphasize problems in which $X = \mathbb{R}^n$, a finite-dimensional Euclidean space, and neighborhoods are defined by Euclidean distance. In this setting, one can apply the techniques of analysis. For example, if f is differentiable and $C = X = \mathbb{R}^n$, then a local minimizer must solve the *stationarity equation* $\nabla f(x) = 0$. If f is convex and $\nabla f(x_*) = 0$, then x_* is a local minimizer. Furthermore, every local minimizer of a convex function is a global minimizer. If f is strictly convex, then f has at most one minimizer. The classical method of Lagrange multipliers provides necessary conditions for a local minimizer when C is defined by a finite number of equality constraints, that is, $C = \{x \in X: h(x) = 0\}$.

The Karush–Kuhn–Tucker (KKT) conditions extend the classical necessary conditions to the case of finitely many equality and inequality constraints. These conditions rely on the concept of *feasible directions*. The feasible subset displayed in Figure 2(a) is the dark-shaded circular region. For the point indicated on the boundary of the region, the feasible directions are all directions that lie strictly inside the half-plane lying to the left of the tangent to the circular region at the indicated point. Along any such direction, it is possible to take a step (perhaps only a very short step) from the indicated point and remain in the feasible subset.

To illustrate, let us examine one of the local constrained minimizers in Figure 1(b). In Figure 2(b), the shaded square region in the lower left hand corner is the feasible subset. The contour lines indicate different values of the objective function, extended beyond the feasible subset. The objective function

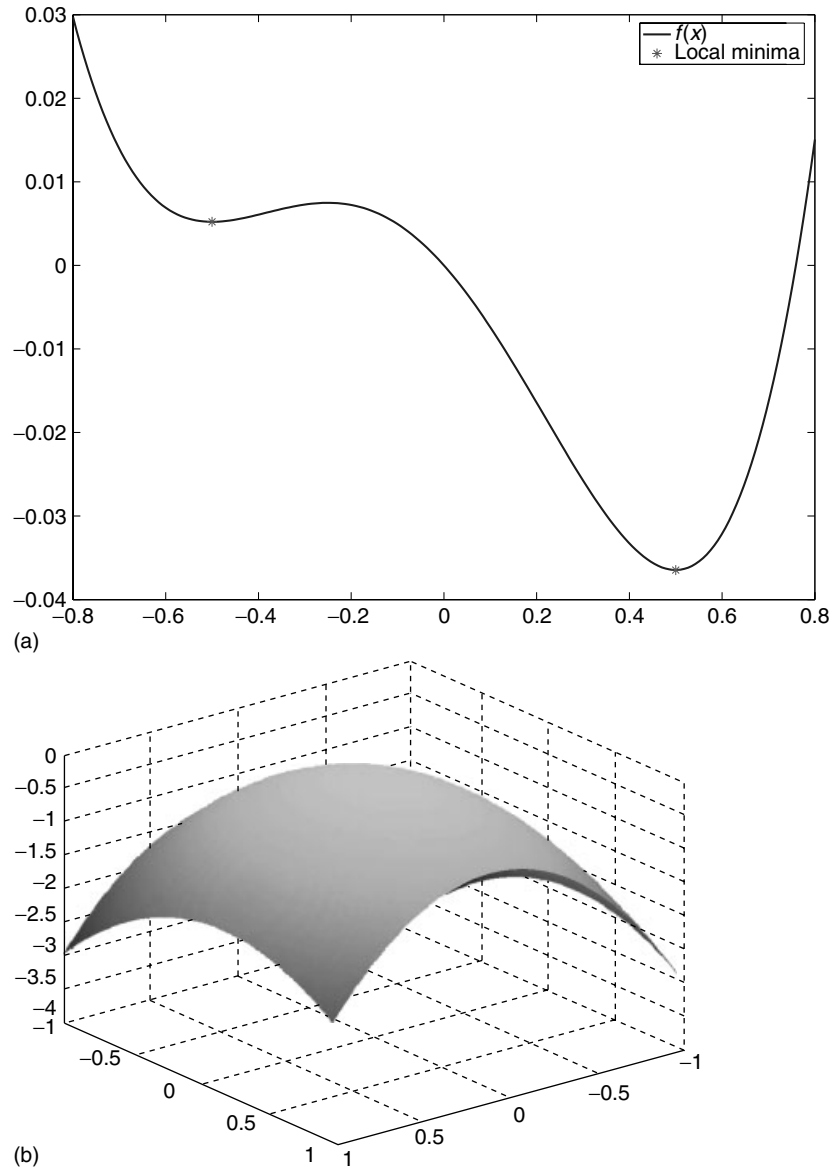


Figure 1 Unconstrained and constrained local minimizers

is decreasing as one moves from the lower left hand to the upper right hand corner at $(1, 1)$. A local constrained minimizer occurs at $(1, 1)$ because no feasible direction at $(1, 1)$ is a descent direction. This geometric idea is captured algebraically by the KKT conditions, which state that the direction of steepest descent can be written as a nonnegative combination of the normals pointing outward

from the constraints that hold at the local minimizer. In Figure 2(b), the direction of steepest descent is indicated by the solid arrow, while the normals to the binding constraints are indicated by dashed arrows.

Occasionally, mathematical characterizations of minimizers yield simple conditions from which minimizers can be deduced analytically. More often

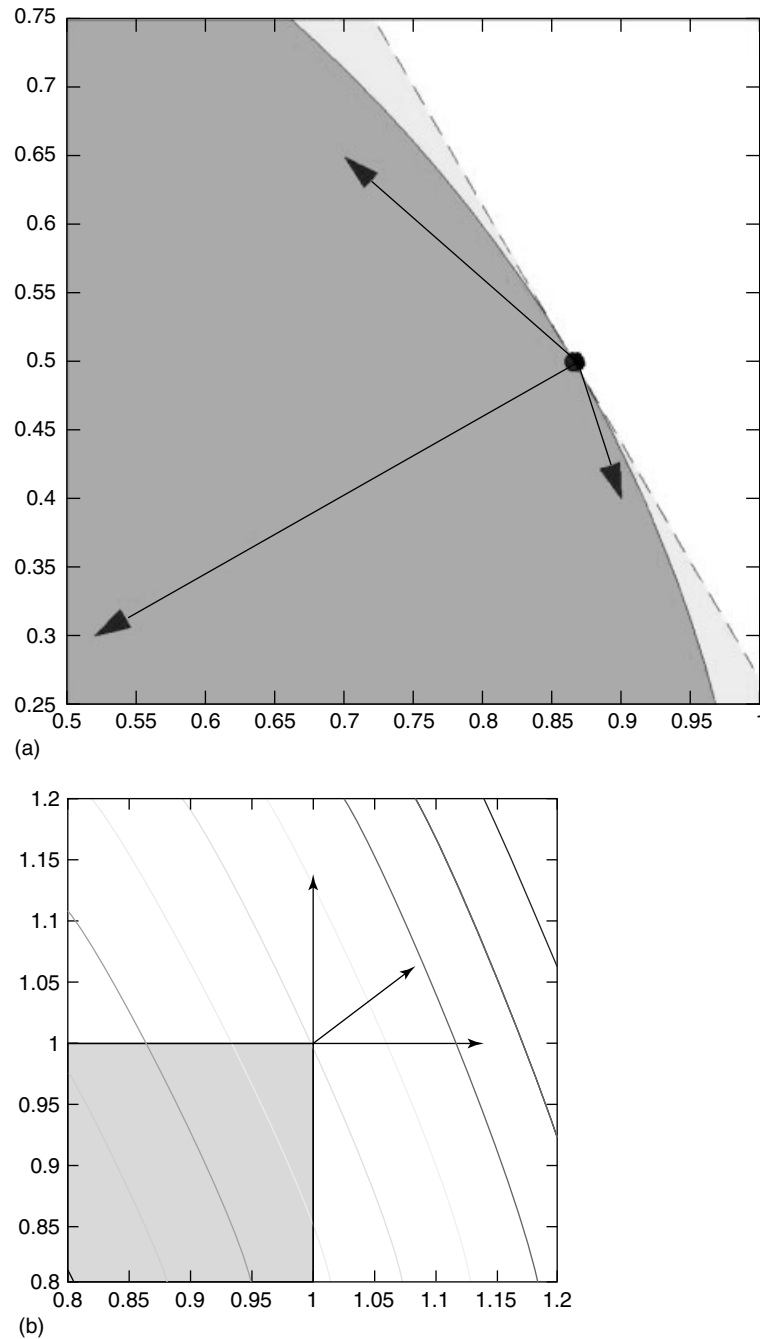


Figure 2 Feasible directions and the KKT conditions

in practice, the resulting conditions are too complicated for analytical deduction. In such cases, one must rely on iterative methods for numerical optimization.

Example

Before reviewing several of the essential ideas that underlie iterative methods for numerical optimization,

4 Optimization Methods

we consider an example that motivates such methods. Bates and Watts [1] fit a nonlinear regression model based on Newton's law of cooling to data collected by Count Rumford in 1798. This leads to an objective function of the form (1),

$$f(\theta) = \sum_i [y_i - 60 - 70 \exp(-\theta z_i)]^2 \quad (4)$$

where z measures cooling time and y_i is the temperature observed at time z_i . The goal is to find the value of the parameter $\theta \in \mathbb{R}$ that minimizes f . Negative values of θ are not physically plausible, suggesting that $C = [0, \infty)$ is a natural feasible subset.

The objective function (4) and its derivative are displayed in Figure 3. The location of the global minimizer, θ_* , is indicated by a vertical dotted line. Because $\theta_* > 0$, that is, the global minimizer lies in the strict interior of the feasible subset, we see that the constraint $\theta \geq 0$ is not needed.

The derivative of the objective function vanishes at θ_* . In theory, θ_* can be characterized as the unique solution of the stationarity equation $f'(\theta) = 0$; in practice, one cannot solve $f'(\theta) = 0$ to obtain a formula for θ_* . The practical impossibility of deriving such formulae necessitates iterative methods for numerical optimization.

How might an iterative method proceed? Suppose that we began by guessing that f might be small at $\theta_0 = 0.03$. The challenge is to improve on this guess. To do so, one might try varying

the value of θ in different directions, noting which values of θ result in smaller values of f . For example, $f(0.03) = 4931.369$. Upon incrementing θ_0 by 0.01, we observe $f(0.04) = 8463.455 > f(0.03)$; upon decrementing θ_0 by 0.01, we observe $f(0.02) = 1735.587 < f(0.03)$. This suggests replacing $\theta_0 = 0.03$ with $\theta_1 = 0.02$. The logic is clear, but it is not so clear how to devise an algorithm that is guaranteed to produce a sequence of values of θ that will converge to θ_* .

Alternatively, having guessed $\theta_0 = 0.03$, we might compute $f'(0.03) = 348599.6$. Because $f'(0.03) > 0$, to decrease f we should decrease θ . But by how much should we decrease θ ? Again, it is not clear how to devise an algorithm that is guaranteed to produce a sequence of values of θ that will converge to θ_* . One possibility, if θ_0 is sufficiently near θ_* , is to apply Newton's method for finding roots to the stationarity equation $f'(\theta) = 0$.

The remainder of this essay considers iterative methods for numerical optimization. Because there are many such methods, we are content to survey several key ideas that underlie unconstrained local optimization, constrained local optimization, and global optimization.

Unconstrained Local Optimization

We begin by considering methods for solving (3) when $C = X = \mathbb{R}^n$. Unless f is convex, there is no

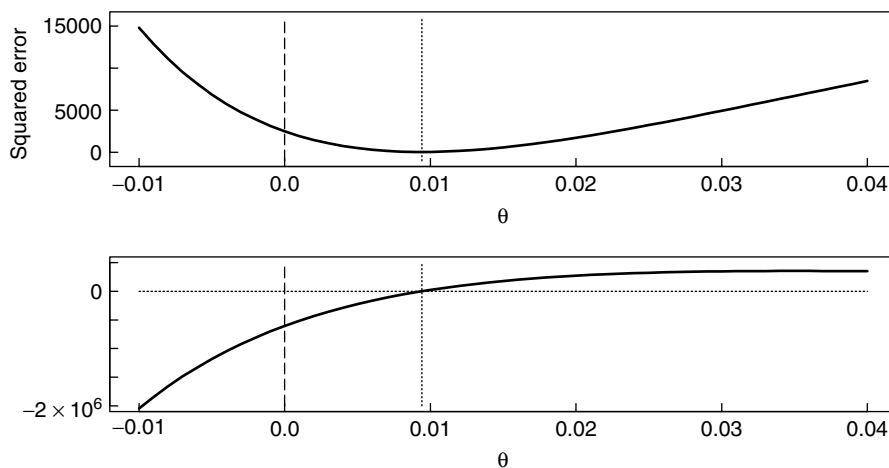


Figure 3 The objective function and its derivative for fitting a nonlinear regression model based on Newton's law of cooling to data collected by Count Rumford

practical way to guarantee convergence to a local minimizer. Instead, we seek algorithms that are guaranteed to converge to a solution of the stationarity equation, $\nabla f(x) = 0$. Because effective algorithms for minimizing f take steps that reduce $f(x)$, they generally do find local minimizers in practice and rarely converge to stationary points that are not minimizers. There are two critical concerns. First, can the algorithm find a solution when started from *any* point in $\mathbb{R}^n$? Algorithms that can are *globally convergent*. Second, will the algorithm converge rapidly once it nears a solution? Ideally, we prefer globally convergent algorithms with fast local convergence.

Global convergence can be ensured through *globalization techniques*, which place mild restrictions on the steps the algorithm may take. Some methods perform *line searches*, requiring a step in the prescribed direction to sufficiently decrease the objective function. Other methods restrict the search to a *trust region*, in which it is believed that the behavior of the objective function can be accurately modeled.

Generally, the more derivative information that an algorithm uses, the faster it will converge. If analytic derivatives are not available, then derivatives can be estimated using finite-differences [6] or complex perturbations [19]. Derivatives can also be computed by automatic differentiation techniques that operate directly on computer programs, applying the chain rule on a statement-by-statement basis [15, 8].

Most iterative methods for local optimization progress by constructing a model of the objective function in the neighborhood of a designated point (the current iterate), then using the model to help find a new point (the subsequent iterate) at which the objective function has a smaller value. It is instructive to classify these methods by the types of models that they construct.

0. Zeroth-order methods do not model the objective function. Instead, they simply compare values of the objective function at different points. Examples of such *direct search methods* include the Nelder–Mead simplex algorithm and the pattern search method of Hooke and Jeeves. The Nelder–Mead algorithm can be effective, but it can also fail, both in theory and in practice. In contrast, pattern search methods are globally convergent.
1. First-order methods construct local linear models of the objective function. Usually, this is

accomplished by using function values and first derivatives to construct the first-order Taylor polynomial at the current iterate. The classic example of a first-order method is the method of steepest descent, in which a line search is performed in the direction of the negative gradient.

2. Second-order methods construct local quadratic models of the objective function. The prototypical example of a second-order method is Newton’s method, which uses function values, first, and second derivatives to construct the second-order Taylor polynomial at the current iterate, then minimizes the quadratic model to obtain the subsequent iterate. (Newton’s method for optimization is equivalent to applying Newton’s method for finding roots to the stationarity equation.) When second derivatives are not available, *quasi-Newton methods* (also known as *variable metric methods* or *secant update methods*) use function values and first derivatives to update second derivative approximations. Various update formulas can be used; the most popular is BFGS (for Broyden, Fletcher, Goldfarb, and Shanno – the researchers who independently suggested it).

Zeroth-order methods are reviewed in [10]. Most surveys of unconstrained local optimization methods, e.g., [3, 4, 13, 14], emphasize first- and second-order methods. The local linear and quadratic models used by these methods are illustrated in Figure 4.

In Figure 4(a), the graph of the objective function is approximated by the line that is tangent at $x = 1/2$. This approximation suggests a descent direction; however, because it has no minimizer, it does not suggest the potential location of a (local) minimizer of the original function. When a linear approximation is used to search for the next iterate, it is necessary to control the length of the step, for example, by performing a line search.

In Figure 4(b), the same graph is approximated by the quadratic curve that is tangent at $x = 1/2$. This approximation has a unique minimizer that suggests itself as the potential location of a (local) minimizer of the original function. However, because the quadratic approximation is only valid near the point where it was constructed (here $x = 1/2$), it is still necessary to insure that the step suggested by the approximation does, in fact, decrease the value of the objective function.

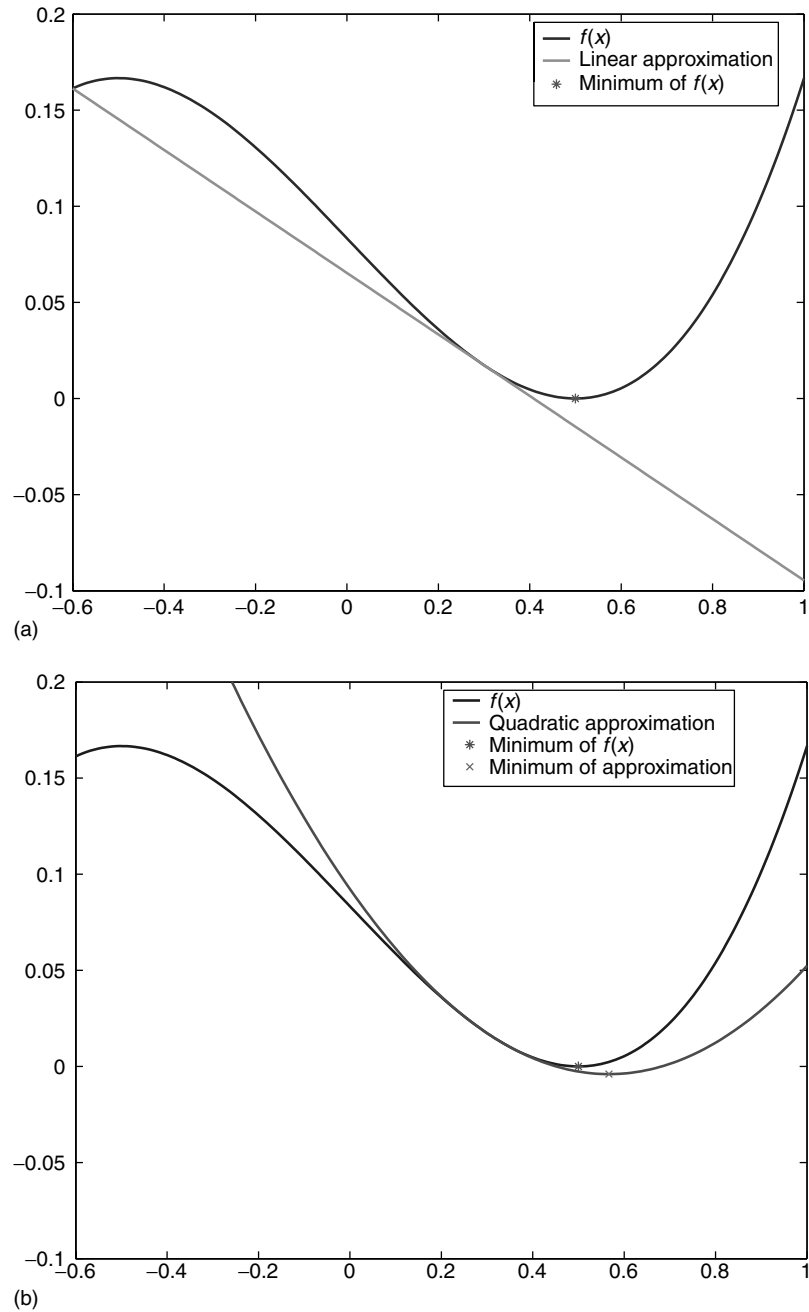


Figure 4 Local models of an objective function

Constrained Local Optimization

Next we discuss several basic strategies for finding local minimizers when $X = \mathbb{R}^n$, and the feasible set

C is a closed proper subset of X . In some cases, one can adapt unconstrained methods to accommodate constraints. For example, in seeking to minimize f with current iterate x_k , the method of steepest descent

would search for the next iterate, x_{k+1} , along the ray

$$\{x_k - t \nabla f(x_k) : t \geq 0\}. \quad (5)$$

This is a gradient method. But what if the ray in the direction of the negative gradient does not lie in the feasible set? If C is convex, there is an obvious remedy. Given $x \in X$, let $P_C(x)$ denote the point in C that is nearest x , the *projection* of x into C . Instead of searching in (5), a *gradient projection method* searches in

$$\{P_C(x_k - t \nabla f(x_k)) : t \geq 0\}.$$

Because each iterate produced by gradient projection lies in C , this is a *feasible-point method*.

Most algorithms for constrained local optimization proceed by solving a sequence of approximating optimization problems. One generic strategy is to approximate the constrained problem with a sequence of unconstrained problems. For example, suppose that $x \in C$ if and only if x satisfies the equality constraints $h_1(x) = 0$ and $h_2(x) = 0$. To approximate this constrained optimization problem with an unconstrained problem, we remove the constraints, allowing all $x \in X$, and modify the objective function so that $x \notin C$ are penalized, e.g.,

$$f_k(x) = f(x) + \rho_k (h_1(x) + h_2(x))^2, \quad \rho_k > 0.$$

The modified objective function is used in searching for the next iterate, x_{k+1} . Typically, $x_{k+1} \notin C$, so *penalty function methods* are *infeasible-point methods*. To ensure that $\{x_k\}$ will converge to a solution of the original constrained problem, ρ_k must be increased as optimization progresses.

Another opportunity to approximate a constrained problem with a sequence of unconstrained problems arises when the feasible set of the constrained problem is defined by a finite number of inequality constraints. An *interior point method* starts at a feasible point in the strict interior of C , and, thereafter, produces a sequence of iterates that may approach the boundary of C (on which local minimizers typically lie), but remain in its strict interior. This is accomplished by modifying the original objective function, adding a penalty term that grows near the boundary of the feasible region. For example, suppose that $x \in C$ if and only if x satisfies the inequality constraints $c_1(x) \geq 0$ and $c_2(x) \geq 0$. In the logarithmic barrier method, the modified objective function

$$f_k(x) = f(x) - \mu_k (\log c_1(x) + \log c_2(x)), \quad \mu_k > 0,$$

is used in searching for the next iterate, x_{k+1} . To ensure that $\{x_k\}$ will converge to a solution of the original constrained problem, μ_k must be decreased to zero as optimization progresses.

In many applications, the feasible region C is defined by finite numbers of equality constraints ($h(x) = 0$) and inequality constraints ($c(x) \geq 0$). There are important special cases of such problems for which specialized algorithms exist. The optimization problem is a *linear program* (LP) if both the objective function and the constraint functions are linear; otherwise, it is a *nonlinear program* (NLP). An NLP with a quadratic objective function and linear constraint functions is a *quadratic program* (QP).

An inequality constraint $c_j(x) \geq 0$ is said to be active at x_* if $c_j(x_*) = 0$. Active set strategies proceed by solving a sequence of approximating equality-constrained optimization problems. The equality constraints in the approximating problem are the equality constraints in the original problem, together with those inequality constraints that are believed to be active at the solution and, therefore, can be treated as equality constraints. The best-known example of an active set strategy is the simplex method for linear programming. Active set strategies often rely on clever heuristics to decide which inequality constraints to include in the working set of equality constraints.

Sequential quadratic programming (SQP) algorithms are among the most effective local optimization algorithms in general use. SQP proceeds by solving a sequence of approximating QPs. Each QP is typically defined by a quadratic approximation of the objective function being minimized, and by linear approximations of the constraint functions. The quadratic approximation is either based on exact second derivatives or on versions of the secant update approximations of second derivatives discussed in connection with unconstrained optimization. SQP is particularly well-suited for treating nonlinear equality constraints, and many SQP approaches use an active set strategy to produce equality-constrained QP approximations of the original problem.

For further discussion of constrained local optimization techniques, see [4, 5, 6, 13, 14].

Global Optimization

So far, we have emphasized methods that exploit local information, e.g., derivatives, to find local

minimizers. We would prefer to find global minimizers, but global information is often impossible to obtain. Convex programs are the exception: if f is a convex function and C is a convex set, then one can solve (2) by solving (3). Lacking convexity, one should embark on a search for a global minimizer with a healthy suspicion that the search will fail.

Why is global optimization so difficult? One often knows that an objective function is well-behaved locally, so that information about f at x_k conveys information about f in a neighborhood of x_k . However, one rarely knows how to infer behavior in one neighborhood from behavior in another neighborhood. The prototypical global optimization problem is discrete, that is, C is a finite set, and there is no notion of neighborhood in X . Of course, if C is finite then (2) can be solved by direct enumeration: evaluate f at each $x \in C$ and find the smallest $f(x)$. This is easy in theory, but C may be so large that it takes centuries in practice. For instance, the example in Figure 1(b) can be generalized to a problem on the unit cube in n variables, and each of the 2^n vertices of the cube is a constrained local minimizer. Thus, methods for global optimization tend to be concerned with clever heuristics for speeding up the search.

Most of the best-known methods for global optimization were originally devised for discrete optimization, then adapted (with varying degrees of success) for continuous optimization. The claims made on behalf of these methods are potentially misleading. For example, a method may be ‘guaranteed’ to solve (2), but the guarantee usually means something like: if one looks everywhere in C , then one will eventually solve (2); or, if one looks long enough, then one will almost surely find something arbitrarily close to a solution. Such guarantees have little practical value [11].

Because it seems impossible to devise efficient ‘off-the-shelf’ methods for global optimization, we believe that methods customized to exploit the specific structure of individual applications should be used or developed whenever possible. Nevertheless, a variety of useful general strategies for global optimization are available.

The most common benchmark for global optimization is the *multistart strategy* [17]. One defines neighborhoods, chooses a local search strategy, starts local searches from a number of different points, finds corresponding local minimizers, and declares the local minimizer at which the objective function is

smallest to be the putative global minimizer. The success of this strategy is highly dependent on how the initial iterates are selected: if one succeeds in placing an initial iterate in each basin of the objective function, then one should solve (2).

When using multistart, it may happen that a number of initial iterates are placed in the same basin, resulting in wasteful duplication as numerous local searches all find the same local minimizer. To reduce such inefficiency, it would seem natural to use the results of previous local searches to guide the placement of the initial iterates for subsequent local searches. Such memory-based strategies are the hallmark of a *tabu search*, a meta-heuristic that guides local search strategies [7].

The concept of a tabu search unifies various approaches to global optimization. For example, bounding methods begin by partitioning C into regions for which lower and upper bounds on f are available. These bounds are used to eliminate regions that cannot contain the global minimizer; then, the same process is applied to each of the remaining regions. The bounds may be obtained in various ways. Lipschitz methods [9] require knowledge of a global constant K such that $|f(x) - f(y)| \leq K\|x - y\|$ for all $x, y \in C$. Then, having computed $f(x)$, one deduces that $f(x) - \varepsilon \leq f(y) \leq f(x) + \varepsilon$ for all y within distance ε/K of x . Interval methods [16] compute bounds by means of interval arithmetic, which uses information about the range of possible inputs to an arithmetic statement to bound the range of outputs. This bounding process can be performed on a statement-by-statement level inside a computer program.

From the perspective of global optimization, one evident difficulty with a local search strategy is its inability to escape the basin in which it finds itself searching. One antidote to this difficulty is to provide an escape mechanism. This is the premise of *simulated annealing*. Suppose that x_k is the current iterate and that the local search strategy has identified a trial iterate, x_t , with $f(x_t) > f(x_k)$. Alone, the local search strategy would reject x_t and search for another trial iterate. Simulated annealing, gambling that perhaps x_t lies in a deeper basin than x_k , figuratively tosses a (heavily weighted) coin to decide whether or not to set $x_{k+1} = x_t$.

From a slightly different perspective [20], simulated annealing uses the Metropolis algorithm to construct certain nonstationary Markov chains/processes

on C , the behavior of which lead (eventually) to solutions of (2). This is a prime example of a random search strategy, in which stochastic simulation is used to sample C . Other well-known examples include genetic algorithms and evolutionary search strategies [18]. The effectiveness of random search strategies depends on how efficiently they exploit information about the problem that they are attempting to solve. Usually, C is extremely large, so that pure random search (randomly sample from C until one gets tired) is doomed to failure. More intelligent random search strategies try to sample from the most promising regions in C .

Conclusion

The discipline of numerical optimization includes more topics than one can possibly mention in a brief essay. We have emphasized nonlinear continuous optimization, for which a number of sophisticated methods have been developed. The study of nonlinear continuous optimization begins with the general ideas that we have sketched; however, more specialized methods such as the EM algorithm for maximum likelihood (see **Maximum Likelihood Estimation**) estimation and Gauss–Newton methods for nonlinear least squares will be of particular interest to statisticians. Furthermore, to each assumption that we have made corresponds a branch of numerical optimization dedicated to developing methods that can be used when that assumption does not hold. For example, nonsmooth optimization is concerned with nonsmooth objective functions, multiobjective optimization is concerned with multiple objective functions, and stochastic optimization is concerned with objective functions that cannot be evaluated, only randomly sampled. Each of these subjects has its own literature.

The reader who seeks to use the ideas described in this essay may discover that good software for numerical optimization is hard to find and harder to write. One attempt to address this situation is the network-enabled optimization system (NEOS) Server [2], whereby the user submits an optimization problem to be solved by state-of-the-art solvers running on powerful computing platforms. Many of the solvers available through NEOS are described in [12].

References

- [1] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*, John Wiley & Sons, New York.
- [2] Czyzyk, J., Mesnier, M.P. & Moré, J.J. (1996). *The Network-Enabled Optimization System (NEOS) Server*, Preprint MCS-P615-0996, Argonne National Laboratory, Argonne. The URL for the NEOS server is <http://www-neos.mcs.anl.gov/neos/>.
- [3] Dennis, J.E. & Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs. Republished by SIAM, Philadelphia, in 1996 as Volume 16 of *Classics in Applied Mathematics*.
- [4] Fletcher, R. (2000). *Practical Methods of Optimization*, 2nd Edition, John Wiley & Sons, New York.
- [5] Forsgren, A., Gill, P.E. & Wright, M.H. (2002). Interior methods for nonlinear optimization, *Society for Industrial and Applied Mathematics Review* **44**(4), 525–597.
- [6] Gill, P.E., Murray, W. & Wright, M.H. (1981). *Practical Optimization*, Academic Press, London.
- [7] Glover, F. & Laguna, M. (1997). *Tabu Search*, Kluwer Academic Publishers, Dordrecht.
- [8] Griewank, A. (2000). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, SIAM, Philadelphia. *Frontiers in Applied Mathematics*.
- [9] Hansen, P. & Jaumard, B. (1995). Lipschitz optimization, in *Handbook of Global Optimization*, R., Horst & P.M., Pardalos, eds, Kluwer Academic Publishers, Dordrecht, pp. 407–493.
- [10] Kolda, T.G., Lewis, R.M. & Torczon, V. (2003). Optimization by direct search: new perspectives on some classical and modern methods, *Society for Industrial and Applied Mathematics Review* **45**(3), 385–482.
- [11] *Mathematical Programming* A new algorithm for optimization, **3**, 124–128.
- [12] Moré, J.J. & Wright, S.J. (1993). *Optimization Software Guide*, Society for Industrial and Applied Mathematics, Philadelphia.
- [13] Nash, S.G. & Sofer, A. (1996). *Linear and Nonlinear Programming*, McGraw-Hill, New York.
- [14] Nocedal, J. & Wright, S.J. (1999). *Numerical Optimization*, Springer-Verlag, Berlin.
- [15] Rall, L.B. (1981). *Automatic Differentiation—Techniques and Applications*, Vol. 120, of *Springer-Verlag Lecture Notes in Computer Science* Springer-Verlag, Berlin.
- [16] Ratschek, H. & Rokne, J. (1995). Interval methods, in *Handbook of Global Optimization*, R., Horst & P.M., Pardalos, eds, Kluwer Academic Publishers, Dordrecht, pp. 751–828.
- [17] Rinnooy Kan, A.H.G. & Timmer, G.T. (1989). Global optimization, in *Optimization*, G.L., Nemhauser, A.H.G., Rinnooy Kan & M.J., Todd, eds, Elsevier, Amsterdam, pp. 631–662.

- [18] Schwefel, H.-P. (1995). *Evolution and Optimum Seeking*, John Wiley & Sons, New York. (See also **Markov Chain Monte Carlo and Bayesian Statistics**)
- [19] Squire, W. & Trapp, G. (1998). Using complex variables to estimate derivatives of real functions, *Society for Industrial and Applied Mathematics Review* **40**(1), 110–112.
- [20] Trosset, M.W. (2001). What is simulated annealing?, *Optimization and Engineering* **2**, 201–213.

ROBERT MICHAEL LEWIS AND
MICHAEL W. TROSSET

Ordinal Regression Models

SCOTT L. HERSHBERGER

Volume 3, pp. 1491–1494

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ordinal Regression Models

A number of regression models for analyzing ordinal variables have been proposed [2]. We describe one of the most familiar *ordinal regression* models: the *ordinal logistic model* (see **Logistic Regression**) [6]. The ordinal logistic model is a member of the family of **generalized linear models** [4].

Generalized Linear Models

Oftentimes, our response variable is not continuous. Instead, the response variable may be dichotomous, ordinal, or nominal, or may even be simply frequency counts. In these cases, the standard linear regression model (the *General Linear Model*) is not suitable for several reasons [4]. First, heteroscedastic and nonnormal errors are almost certain to occur when the response variable is not continuous. Second, the standard linear regression model (see **Multiple Linear Regression**) will often predict values that are impossible. For example, if the response variable is dichotomous, the linear model will predict scores that are less than 0 or greater than 1. Third, the functional form specified by the linear model will often be incorrect. For example, it is unlikely that extreme values of an explanatory variable will have the same effect on the response variable than more moderate values.

Consider a binary response variable Y (e.g., 0 = no and 1 = yes) and an explanatory variable X , which may be binary or continuous. In order to evaluate whether X has an effect on Y , we examine whether the probability of an event $\pi(x) = P(Y = 1|X = x)$ is associated with X by means of the generalized linear model

$$f(\pi(x)) = \alpha + \beta x, \quad (1)$$

where f is a *link function* relating Y and X , α is an intercept, and β is a regression coefficient for X . Link functions specify the correct relationship between Y and X : linear, logistic, lognormal, as well as many others. Where there are more than $k = 1$ explanatory variables, the model can be written as

$$f(\pi(x)) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \quad (2)$$

The two most common link functions for analyzing binary and ordinary response variables are the *logit link*,

$$f(\pi) = \log\left(\frac{\pi}{(1 - \pi)}\right), \quad (3)$$

and the *complementary log-log link*,

$$f(\pi) = \log(-\log(1 - \pi)). \quad (4)$$

The logit link represents the *logistic regression* model since

$$\log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \quad (5)$$

when

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{(1 + \exp(\alpha + \beta x))}. \quad (6)$$

The complementary log-log link function is an extreme-value regression model since

$$\log(-\log(1 - \pi(x))) = \alpha + \beta x \quad (7)$$

when

$$\pi(x) = 1 - \exp(-\exp(\alpha + \beta x)). \quad (8)$$

Other link functions include $f(\pi) = \mu$, the link function for linear regression; $f(\pi) = \Phi^{-1}(\mu)$, the link function for probit regression; and $f(\pi) = \log(\mu)$, the link function for Poisson regression (see **Generalized Linear Models (GLM)**).

The Ordinal Logistic Regression Model

Now let Y be a categorical response variables with $c + 1$ ordered categories. For example, we might have the response variable Y of performance, with the ordered categories of Y as 1 = fail, 2 = marginally pass, 3 = pass, and 4 = pass with honors. Let $\pi_k(x) = p(Y = k|X = x)$ be the probability of observing $Y = k$ given $X = x$, $k = 0, 1, \dots, c$. We can then formulate a model for ordinal response variable Y based upon the cumulative probabilities

$$\begin{aligned} f(\gamma_k(x)) &= p(Y \geq k|X = x) \\ &= a_k + \beta_k x, \quad k = 1, \dots, c. \end{aligned} \quad (9)$$

Instead of specifying one equation describing the relationship between Y and X , we specify k model

2 Ordinal Regression Models

equations and k regression coefficients to describe this relationship. Each equation specifies the probability of membership in one category versus membership in all other higher categories.

This model can be simplified considerably if we assume that the regression coefficient does not depend on the value of the explanatory variable, in this case $k = 1, 2$, or 3 , the equal slopes assumption of many models, including the ordinary least-squares regression model and the **analysis of covariance (ANCOVA)** model. The cumulative probability model with equal slopes is

$$\begin{aligned} f(\gamma_k(x)) &= p(Y \geq k | X = x) \\ &= \alpha_k + \beta x, \quad k = 1, \dots, c. \end{aligned} \quad (10)$$

Thus, we assume that for a link function f the corresponding regression coefficients are equal for each cut-off point k ; parallel lines are fit that are based on the cumulative distribution probabilities of the response categories. Each model has a different α but the same β .

Different models result from the use of different link functions. If the logit link function is used, we obtain the ordinal logistic regression model:

$$\begin{aligned} \log\left(\frac{\gamma_k(x)}{(1 - \gamma_k(x))}\right) &= \alpha_k + \beta x \\ \Leftrightarrow \gamma_k(x) &= \frac{\exp(\alpha_k + \beta x)}{(1 + \exp(\alpha_k + \beta x))}. \end{aligned} \quad (11)$$

Some writers refer to the ordinal logistic regression model using the logit link function as the proportional odds model [4].

Alternatively, if the complementary log–log link function is used, we have

$$\begin{aligned} \log(-\log(\gamma_k(x))) &= \alpha_k + \beta x \\ \Leftrightarrow \gamma_k(x) &= 1 - \exp(-\exp(\alpha_k + \beta x)). \end{aligned} \quad (12)$$

Some writers refer to the ordinal logistic regression model using the complementary log–log link as the discrete proportional hazards model [3].

In practice, the results obtained using either the logistic or complementary log–log link functions are substantively the same, only differing in scale. Arguments can be made, however, for preferring the use of the logistic function over the complementary log–log function. First, only the sign – and not the magnitude – of the regression coefficients change if the

order of the categories is reversed. Second, the ordinal logistic regression model has the important property of *collapsibility*. The parameters (e.g., regression coefficients) of models that have the property of collapsibility do not change in sign or magnitude even if some of the response categories are combined [7]. A third advantage is the relative interpretability of the ordinal logistic regression coefficients: $\exp(\beta)$ is an invariant odds ratio over all category transitions, and thus can be used as a single summary measure to express the effect of an explanatory variable on a response variable, regardless of the value of the explanatory variable.

Example Consider the following 2×5 contingency table (see Table 1) between party affiliation and political ideology, adapted from Agresti [1].

To model the relationship between party affiliation (the explanatory variable X , scored as Democratic = 0 and Republican = 1) and political ideology (the ordered response variable Y , scored from 1 = very liberal to 5 = very conservative), we used PROC LOGISTIC in SAS [5], obtaining these values for the α_k and β :

$$\begin{aligned} \alpha_1 &= -2.47 \\ \alpha_2 &= 1.47 \\ \alpha_3 &= 0.24 \\ \alpha_4 &= 1.07 \\ \beta &= .98. \end{aligned} \quad (13)$$

The **maximum likelihood estimate** of $\beta = .98$ means that for any level of political ideology, the estimated odds that a Democrat's response is in the liberal direction is $\exp(.98) = 2.65$, or more than two and one-half times the estimated odds for Republicans.

We can use the parameter values to calculate the cumulative probabilities for political ideology. The cumulative probabilities equal

$$p(Y \leq k | X = x) = \frac{\exp(\alpha_k + \beta x)}{1 + \exp(\alpha_k + \beta x)} \quad (14)$$

For example, the cumulative probability for a Republican having a very liberal political ideology is

$$p(Y \leq 1 | X = 0) = \frac{\exp(-2.47 + .98(0))}{1 + \exp(-2.47 + .98(0))} = .08, \quad (15)$$

Table 1 Relationship between political ideology and party affiliation

Party affiliation	Political ideology				
	Very liberal	Slightly liberal	Moderate	Slightly conservative	Very conservative
Democratic	80	81	171	41	55
Republican	30	46	148	84	99

whereas the cumulative probability for a Democrat having a very liberal political ideology is

$$p(Y \leq 1 | X = 1) = \frac{\exp(-2.47 + .98(1))}{1 + \exp(-2.47 + .98(1))} = .18. \quad (16)$$

References

- [1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis.*, Wiley, New York.
- [2] Ananth, C.V. & Kleinbaum, D.G. (1997). Regression models for ordinal responses: a review of methods and applications, *International Journal of Epidemiology* **26**, 1323–1333.
- [3] McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society: B* **42**, 109–142.
- [4] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models.*, Chapman & Hall, New York.
- [5] SAS Institute (1992). *SAS/STAT User's Guide, Version 9*, Vols. 1 and 2, SAS Institute, Inc, Cary.
- [6] Scott, S.C., Goldberg, M.S. & Mayo, N.E. (1997). Statistical assessment of ordinal outcomes in comparative studies, *Journal of Clinical Epidemiology* **50**, 45–55.
- [7] Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer, New York.

SCOTT L. HERSHBERGER

Outlier Detection

RAND R. WILCOX

Volume 3, pp. 1494–1497

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Outlier Detection

Dealing with Outliers

Outliers are a practical concern because they can distort commonly used descriptive statistics such as the sample mean and variance, as well as least squares regression; they can also result in low power relative to other techniques that might be used. Moreover, modern methods for detecting outliers suggest that they are common and can appear even when the range of possible values is limited. So, a practical issue is finding methods for dealing with outliers when comparing groups and studying associations among variables.

Another and perhaps more basic problem is finding good outlier detection techniques. A fundamental criterion when judging any outlier detection rule is that it should not suffer from masking, meaning that the very presence of outliers should not render a method ineffective at finding them. Outlier detection rules based on the mean and variance suffer from masking because the mean, and especially the variance, can be greatly influenced by outliers. What is needed are measures of variation and location that are not themselves affected by outliers. In the univariate case, a boxplot rule is reasonably effective because it uses the interquartile range which is not affected by the smallest and largest 25% of the data. Rules based on the median and a measure of variation called the **median absolute deviation** are commonly used as well. (If M is the median of $X_1, \dots, X_n$, the median absolute deviation is the median of the values $|X_1 - M|, \dots, |X_n - M|$.) Letting MAD represent the median absolute deviation, a commonly used method is to declare X_i an outlier if

$$\frac{|X_i - M|}{\text{MAD}/0.6745} > 2.24 \quad (1)$$

(e.g., [3] and [4]).

Detecting outliers among multivariate data is a substantially more complex problem, but effective methods are available (see **Multivariate Outliers**). To begin, consider two variables, say X and Y . A simple approach is to use some outlier detection method on the X values that avoids masking (such as a boxplot or the mad-median rule just described), and then do the same with the

Y values. This strategy is known to be unsatisfactory, however, because it does not take into account the overall structure of the data. That is, influential outliers can exist even when no outliers are detected among the X values, ignoring Y , and simultaneously, no outliers are found among the Y values, ignoring X . Numerous methods have been proposed for dealing with this problem, several of which appear to deserve serious consideration [3, 4]. Also see [1]. It is stressed, however, that methods based on the usual means and covariance matrix, including methods based on **Mahalanobis distance**, are known to perform poorly due to masking (e.g., [2]).

One of the best-known methods among statisticians is based on something called the *minimum volume ellipsoid* (MVE) estimator. The basic strategy is to search for that half of the data that is most tightly clustered together based on its volume. Then the mean and covariance matrix is computed using this central half of the data only, the idea being that this avoids the influence of outliers. Finally, a generalization of Mahalanobis distance is used to check for outliers. A related method is based on what is called the *minimum covariance determinant estimator* [2, 3, 4]. These methods represent a major advance, but for certain purposes, alternative methods now appear to have advantages. One of these is based on a collection of projections of the data. Note that if all points are projected onto a line, yielding univariate data, a univariate outlier detection could be applied. The strategy is to declare any point an outlier if it is an outlier for any projection of the data. Yet another strategy that seems to have considerable practical value is based on the so-called *minimum generalized variance* (MGV) method. These latter two methods are computationally intensive but can be easily performed with existing software [3, 4].

Next, consider finding a good measure of location when dealing with a single variable. A fundamental goal is finding a location estimator that has a relatively small standard error. Under normality, the sample mean is optimal, but under arbitrarily small departures from normality, this is no longer true, and in fact the sample mean can have a standard error that is substantially larger versus many competitors. The reason is that the population variance can be greatly inflated by small changes in the tails of a

2 Outlier Detection

distribution toward what are called *heavy-tailed distributions*. Such distributions are characterized by an increased likelihood of outliers relative to the number of outliers found when sampling from a normal distribution. Moreover, the usual estimate of the standard error of the sample mean, $\sqrt{(s^2/n)}$, where s^2 is the usual sample variance and n is the sample size, is extremely sensitive to outliers. Because outliers can inflate the standard error of the sample mean, a strategy when dealing with this problem is to reduce or even eliminate the influence of the tails of a distribution. There are two general approaches which are used: the first is to simply trim some predetermined proportion of the largest and smallest observations from the sample available, that is, use what is called a **trimmed mean**; the second is to use some measure of location that checks for outliers and removes or down weights them if any are found. In the statistical literature, the best-known example is a so-called **M-estimator of location**. Both trimmed means and M-estimators can be designed so that under normality, they have standard errors nearly as small as the standard error of the sample mean. But they have the advantage of yielding standard errors substantially smaller than the standard error of the sample mean as we move from a normal distribution toward distributions having heavier tails. In practical terms, power might be substantially higher when using a robust estimator.

An important point is that when testing hypotheses, it is inappropriate to simply apply methods for means to the data that remain after trimming or after outliers have been removed (e.g., [4]). The reason is that once extreme values are removed, the remaining observations are no longer independent under random sampling. When comparing groups based on trimmed means, there is a simple method for dealing with this problem which is based in part on Winsorizing the data to get a theoretically sound estimate of the standard error (see **Winsorized Robust Measures**). To explain Winsorizing, suppose 10% trimming is done. So, the smallest 10% and the largest 10% of the data are removed and the trimmed mean is the average of the remaining values. Winsorizing simply means that rather than eliminate the smallest values, their value is reset to the smallest value not trimmed. Similarly, the largest values that were trimmed are now set equal to the largest value not trimmed. The sample variance, applied to the Winsorized values, is called the *Winsorized variance*, and theory indicates that it

should be used when estimating the standard error of a trimmed mean.

When using M-estimators, estimates of the standard error take on a complicated form, but effective methods (and appropriate software) are available. Unlike methods based on trimmed means, the more obvious methods for testing hypotheses based on M-estimators are known to be unsatisfactory with small to moderate sample sizes. The only effective methods are based on some type of bootstrap method [3, 4]. Virtually all of the usual experimental designs, including one-way, two-way, and repeated measures designs, can be analyzed.

When working with correlations, simple methods for dealing with outliers among the marginal distributions include **Kendall's tau**, **Spearman's rho**, and a so-called *Winsorized correlation*. But these methods can be distorted by outliers, even when no outliers among the marginal distributions are found [3, 4]. There are, however, various strategies for taking the overall structure of the data into account. One is to simply eliminate (or down weight) any points declared outliers using a good multivariate outlier detection method. These are called *skipped correlations*. Another is to identify the centrally located points, as is done for example by the minimum volume ellipsoid estimator, and ignoring the points not centrally located, compute a correlation coefficient based on the centrally located values. A related method is to look for the half of the data that minimizes the determinant of the covariance matrix, which is called the *generalized variance*. Both of these methods are nontrivial to implement, but the software SAS, S-PLUS, and R have built-in functions that perform the computations (see **Software for Statistical Analyses**).

In terms of testing the hypothesis of a zero correlation, with the goal of establishing independence, again it is inappropriate to simply apply the usual hypothesis testing methods using the remaining data. This leads to using the wrong standard error, and if the problem is ignored, poor control over the probability of a Type I error results. There are, however, ways of dealing with this problem as well as appropriate software [3, 4]. One approach that seems to be particularly effective is a skipped correlation where a projection-type method is used to detect outliers, any outliers found are removed,

and Pearson's correlation is applied to the remaining data. When there are multiple correlations and the goal is to test the hypothesis that all correlations are equal to zero, replacing Pearson's correlation with Spearman's correlation seems to be necessary in order to control the probability of a Type I error.

As for regression, a myriad of methods has been proposed for dealing with the deleterious effects of outliers. A brief outline is provided here and more details can be found in [2, 3], and [4]. One approach, with many variations, is to replace the sum of squared residuals with some other function that reduces the effects of outliers. If, for example, the squared residuals are replaced by their absolute values, yielding the least absolute deviation estimator, protection against outliers among the Y values is achieved, but outliers among the X values can still cause serious problems. Rather than use absolute values, various M-estimators use functions of the residuals that guard against outliers among both the X and Y values. However, problems due to outliers can still occur [3, 4].

Another approach is to ignore the largest residuals when assessing the fit to data. That is, determine the slopes and intercept so as to minimize the sum of the squared residuals, with the largest residuals simply ignored. To ensure high resistance to outliers, a common strategy is to ignore approximately the largest half of the squared residuals. Replacing squared residuals with absolute values has been considered as well.

Yet another strategy, called *S-estimators*, is to choose values for the parameters that minimize some robust measure of variation applied to the residuals.

There are also two classes of correlation-type estimators. The first replaces Pearson's correlation with some robust analog which can then be used to estimate the slope and intercept. Consider p predictors, $X_1, \dots, X_p$, and let τ_j be any correlation between X_j , the j th predictor, and $Y - b_1X_1 - \dots - b_pX_p$. The other general class among correlation-type estimators chooses the slope estimates $b_1, \dots, b_p$ so as to minimize $\sum |\hat{\tau}_j|$. Currently, the most common choice for τ is Kendall's tau which (when $p = 1$) yields the Theil-Sen estimator.

Skipped estimators are based on the strategy of first applying some multivariate outlier detection method, eliminating any points that are flagged

as outliers, and applying some regression estimator to the data that remain. A natural suggestion is to apply the usual least squares estimator, but this approach has been found to be rather unsatisfactory [3, 4]. To achieve a relatively low standard error and high power under heteroscedasticity, a better approach is to apply the Theil-Sen estimator after outliers have been removed. Currently, the two best outlier detection methods appear to be the projection-type method and the MGV (minimum generalized variance) method; see [3, 4].

E-type skipped estimators (where E stands for error term) look for outliers among the residuals based on some preliminary fit, remove (or down weight) the corresponding points, and then compute a new fit to the data. There are many variations of this approach.

Among the many regression estimators that have been proposed, no single method dominates in terms of dealing with outliers and simultaneously yielding a small standard error. However, the method used can make a substantial difference in the conclusions reached. Moreover, least squares regression can perform very poorly, so modern robust estimators would seem to deserve serious consideration. More details about the relative merits of these methods can be found in [3] and [4]. While no single method can be recommended, among the many estimators available, skipped estimators used in conjunction with a projection or MGV outlier detection technique, followed by the Theil-Sen estimator, appear to belong to the class of estimators that deserves serious consideration.

One advantage of many modern estimators should be stressed. Even when the error term has a normal distribution, but there is heteroscedasticity, they can yield standard errors that are tens and even hundreds of times smaller than the standard error associated with ordinary least squares.

Finally, as was the case when dealing with measures of location, any estimator that reduces the influence of outliers requires special hypothesis testing methods. But even when there is heteroscedasticity, accurate **confidence intervals** can be computed under fairly extreme departures from normality [3, 4]. Currently, the best hypothesis testing methods are based on a certain type of percentile bootstrap method. The computations are long and tedious, but they can be done quickly with modern computers.

4 Outlier Detection

References

- [1] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, Wiley, New York.
- [2] Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression & Outlier Detection*, Wiley, New York.
- [3] Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*, Academic Press, San Diego.
- [4] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

RAND R. WILCOX

Outliers

RAND R. WILCOX

Volume 3, pp. 1497–1498

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Outliers

Outliers are values that are unusually large or small among a batch of numbers. Outliers are a practical concern because they can distort commonly used descriptive statistics such as the sample mean and variance. One result is that when measuring **effect size** using a standardized difference, large differences can be masked because of even one outlier (e.g., [4]).

To give a rough indication of one reason why, 20 observations were randomly generated from 2 normal distributions with variances equal to 1. The means were 0 and 0.8. Estimating the usual standardized difference yielded 0.84. Then the smallest observation in the first group was decreased from -2.53 to -4 , the largest observation in the second group was increased from 2.51 to 3, the result being that the difference between the means increased from 0.89 to 1.09, but the standardized difference decreased to 0.62; that is, what is generally considered to be a large effect size is now a moderate effect size. The reason is that the variances are increasing faster, in a certain sense, than the difference between the means. Indeed, with arbitrarily large sample sizes, large effect sizes (as originally defined by Cohen using a graphical point of view) can be missed when using means and variances, even with very small departures from normality (e.g., [4]).

Another concern is that outliers can cause the sample mean to have a large standard error compared to other estimators that might be used. One consequence is that outliers can substantially reduce power when using any hypothesis testing method based on means, and more generally, any least squares estimator. In fact, even a single outlier can be a concern (e.g., [4, 5]). Because modern outlier detection methods suggest that outliers are rather common, as predicted by Tukey [3], methods for detecting and dealing with them have taken on increased importance in recent years.

Describing outliers as unusually large or small values is rather vague, but there is no general agreement about how the term should be made more precise. However, some progress has been made in identifying desirable properties for outlier detection methods. For example, if sampling is from a normal distribution, a goal might be that the expected proportion of values declared outliers is relatively small. Another basic criterion is that an

outlier detection method should not suffer from what is called *masking*, meaning that the very presence of outliers inhibits a method's ability to detect them.

Suppose, for example, a point is declared an outlier if it is more than two standard deviations from the mean. Further, imagine that 20 points are sampled from a standard normal distribution and that 2 additional points are added, both having the value 10 000. Then surely the value 10 000 is unusual, but generally these two points are not flagged as outliers using the rule just described. The reason is that the outliers inflate the mean and variance, particularly the variance, the result being that truly unusual values are not detected.

Consider, for instance, the values 2, 2, 3, 3, 3, 4, 4, 4, 100 000, and 100 000. Surely, 100 000 is unusual versus the other values, but 100 000 is not declared an outlier using the method just described. The sample mean is $\bar{X} = 20\,002.5$, the sample variance is $s = 42\,162.38$, and it can be seen that $100\,000 - 20\,002.5 < 2 \times 42\,162.38$. What is needed are methods for detecting outliers that are not themselves affected by outliers, and such methods are now available (e.g., [1, 2, 5]). In the univariate case, the best-known approach is the **boxplot**, which uses the interquartile range. Methods based in part on the median are available and offer certain advantages [5].

Detecting outliers in multivariate data (*see Multivariate Outliers*) is a much more difficult problem. A simple strategy is, for each variable, to apply some univariate outlier detection method, ignoring the other variables that are present. However, among the combination of values, a point can be very deviant, while the individual scores are not. That is, based on the overall structure of the data, a point might seriously distort the mean, for example, or it might seriously inflate the standard error of the mean, even though no outliers are found when examining the individual variables. However, effective methods for detecting multivariate outliers have been developed [2, 5].

References

- [1] Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, Wiley, New York.
- [2] Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.
- [3] Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in *Contributions to Probability and Statistics*, S. Olkin, W. Ghurye, W. Hoeffding,

2 Outliers

- W. Madow & H. Mann, eds, Stanford University Press, Stanford.
- [4] Wilcox, R.R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy*, Springer, New York.
- [5] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

RAND R. WILCOX

Overlapping Clusters

MORVEN LEESE

Volume 3, pp. 1498–1500

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Overlapping Clusters

Many standard techniques in *cluster analysis* find partitions that are mutually exclusive (see **Hierarchical Clustering; k -means Analysis**). In some applications, mutually exclusive clusters would be a natural requirement, but in others, especially in psychological and sociological research, it is plausible that cases should belong to more than one cluster simultaneously. An example might be market research in which an individual person belongs to several consumer groups, or social network research in which people belong to overlapping groups of friends or collaborators. In the latter type of application, the individuals appearing in the intersections of the clusters are often of most interest. The overlapping cluster situation differs from that of **fuzzy clustering**: here, cases cannot be assigned definitely to a single cluster and cluster membership is described in terms of relative weights or probabilities of membership. However, both overlapping and fuzzy clustering are similar in that overlapping clustering relaxes the normal constraint that cluster membership should sum to 1 over clusters, and fuzzy clustering relaxes the constraint that membership should take only the values of 0 or 1.

Limited Overlap

Limited forms of overlap can be obtained from *direct data clustering* methods (see **Two-mode Clustering**) and *pyramids*. A particular type of direct data clustering method [4], sometimes known as the *two-way joining* or *block method*, clusters cases and variables simultaneously by reordering rows and columns of the data matrix so that similar columns and rows appear together and sets of contiguous rows (columns) form the clusters (see the package Systat for example). Overlapping clusters can be defined by extending a set of rows (columns) that forms a cluster, so that parts of adjoining clusters are incorporated. The pyramid is a generalization of the *dendrogram* (see **Hierarchical Clustering**) that can be used as a basis for clustering cases, since a cut through the pyramid at any given height gives rise to a set of ordered, overlapping clusters [3]. In both types of method, individuals can belong to at most two clusters.

Clumping and the B_k Technique

With unlimited overlap the problem is to achieve both an appropriate level of fit to the data and also a model that is simple to interpret. Two early examples of techniques that can allow potentially unlimited degrees of overlap are *clumping* and the B_k technique. Clumping [6] divides the data into two groups using a ‘cohesion’ function including a parameter controlling the degree of overlap. In the B_k technique [5], individuals are represented by nodes in a graph; pairs of nodes are connected that have a similarity value above some specified threshold. Each stage in a hierarchical clustering process finds the set of *maximal complete subgraphs* (the largest sets of individuals for which all pairs of nodes are connected), and these may overlap. The number of clusters can be restricted by choosing k such that a maximum of $k - 1$ objects belong to the overlap of any pair of clusters. Any clusters that have more than $k - 1$ objects in common are amalgamated. This method has been implemented in the packages CLUSTAN (version 3) and Clustan/PC.

Hierarchical Classes for Binary Data

A relatively widely used method is *hierarchical classes* [2], a two-mode method appropriate for binary attribute data that clusters both cases and attributes. It has been implemented as HICLASS software. The theoretical model consists of two hierarchical class structures, one for cases and one for attributes. Cases are grouped into mutually exclusive classes, called ‘bundles’, which in the underlying model have identical attributes. These can then be placed in a hierarchy reflecting subset/superset relations. Each case bundle has a corresponding attribute bundle. Above the lowest level are combinations of bundles, which may overlap. The model is fitted to data by optimizing a goodness of fit index for a given number of bundles (or rank), which has to be chosen by the investigator. Figure 1 shows a simple illustration of this for hypothetical data shown in Table 1, where an entry is 1 if the person exhibits the quality and 0 otherwise (adapted from [7]).

2 Overlapping Clusters

Table 1 Hypothetical data matrix of qualities exhibited by various people. Adapted from Rosenberg, S., Van Mechelen, I. & De Boeck, P. (1996) [7]

	Successful	Articulate	Generous	Outgoing	Hardworking	Loving	Warm
Father	1	1	0	0	1	0	0
Boyfriend	0	0	1	0	1	1	1
Uncle	0	0	0	1	0	1	1
Brother	0	0	0	1	0	1	1
Mother	0	0	1	1	1	1	1
Me now	0	0	1	1	1	1	1
Ideal me	1	1	1	1	1	1	1

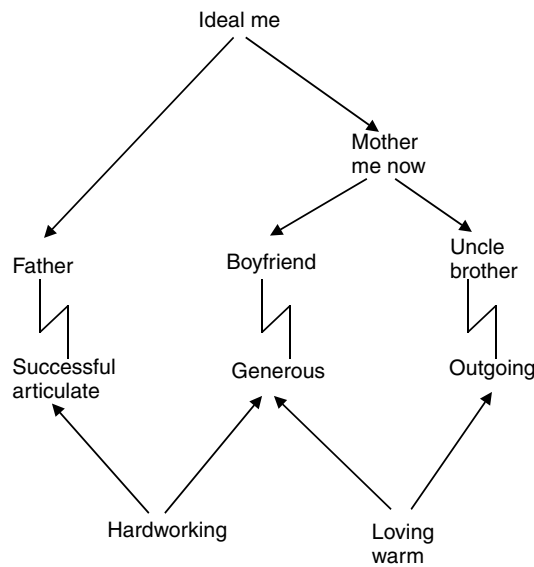


Figure 1 A hierarchical classes analysis of hypothetical data in Table 1. The ‘pure’ clusters, called ‘bundles’ at the bottom of the hierarchies are joined by zigzag lines. Clusters above that level in the hierarchy may contain lower level clusters and hence overlap, for example, (mother, me now, and boyfriend) and (mother, me now, uncle, brother)

Additive Clustering for Proximity Matrices

Additive clustering [1] fits a model to an observed proximity matrix (see **Proximity Measures**) such that the theoretical proximity between any pair of cases is the sum of the weights of those clusters containing that pair. The basic model has the reconstructed similarity between cases as

$$\hat{s}_{ij} = \sum_{k=1}^m w_k p_{ik} p_{jk} \quad (1)$$

where $p_{ik} = 1$ if case i is in cluster k and 0 otherwise, and w_k is a weight representing the ‘salience’ of cluster k . The goodness of fit is measured by the percentage of variance in the observed proximities explained by the model. Algorithms for fitting the values of w_k and p_{ik} include the original ADCLUS software, with an improved algorithm MAPCLUS.

Figure 2 illustrates the results of additive clustering in an analysis of data from a study of the social structure of an American monastery [1, 8]. The social relations between 18 novice monks were assessed in ‘sociograms’, in which the monks rated the highest four colleagues on four positive qualities and four negative qualities. The resulting similarity matrix was

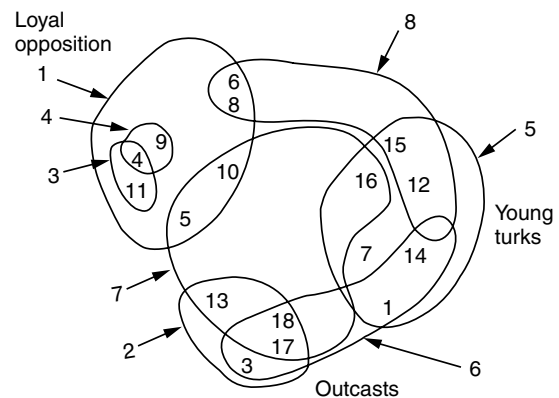


Figure 2 Overlapping clusters of monks found by additive clustering of data on interpersonal relations using MAPCLUS, superimposed on a nonmetric multidimensional scaling plot. Numbers with arrows indicate the cluster numbers, ranked in order of their weights. Other numbers denote individual monks. Text descriptions based on external data have been added. Adapted from Arabia, P. & Carroll, J.D. (1989) [1] and Sampson, S.F. (1968) [8]

analyzed using MAPCLUS and plotted on a two-dimensional representation, in which individuals that are similar are situated close together.

Summary

Methods for identifying overlapping clusters are not usually included in standard software packages. Furthermore, their results are often complex to interpret. For these reasons, applications are relatively uncommon. Nevertheless certain types of research questions demand their use, and for these situations, there are several methods available in specialist software.

References

- [1] Arabie, P. & Carroll, J.D. (1989). *Conceptions of Overlap in Social Structure*, in *Research Methods in Social Network Analysis*, L.C. Freeman, D.R. White & A.K. Romney, eds, George Mason University Press, Fairfax, pp. 367–392.
- [2] De Boeck, P. & Rosenberg, S. (1988). Hierarchical classes: model and data analysis, *Psychometrika* **53**, 361–381.
- [3] Diday, E. (1986). *Orders and Overlapping Clusters by Pyramids in Multidimensional Data Analysis*, J. De Leeuw, W. Heiser, J. Meulman & F. Critchley, eds, DSWO Press, Leiden.
- [4] Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New York.
- [5] Jardine, N. & Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications, *Computer Journal* **11**, 117–184.
- [6] Needham, R.M. (1967). Automatic classification in linguistics, *The Statistician* **1**, 45–54.
- [7] Rosenberg, S., Van Mechelen, I. & De Boeck, P. (1996). A hierarchical classes model: theory and method with applications in psychology and psychopathology. In: P. Arabie, L.J. Hubert & G. De Soete (eds) *Clustering and Classification* World Scientific Singapore, pp. 123–155.
- [8] Sampson, S.F. (1968). A novitiate in a period of change: an experimental and case study of social relationships. Doctoral dissertation Cornell University Microfilms No 69-5775.

MORVEN LEESE

***P* Values**

CHRIS DRACUP

Volume 3, pp. 1501–1503

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

P Values

The *P* value associated with a test statistic is the probability that a study would produce an outcome as deviant as the outcome obtained (or even more deviant) if the null hypothesis was true. Before the advent of high-speed computers, exact *P* values were not readily available, so researchers had to compare their obtained test statistics with tabled critical values corresponding to a limited number of *P* values, such as .05 and .01, the conventional significance levels. This fact may help to explain the cut and dried, reject/do not reject approach to significance testing that existed in the behavioral sciences for so long. According to this approach, a test statistic either was significant or it was not, and results that were not significant at the .05 level were not worthy of publication, while those that reached significance at that level were (*see Classical Statistical Inference: Practice versus Presentation*).

Statistical packages now produce *P* values as a matter of course, and this allows behavioral scientists to show rather more sophistication in the evaluation of their data. If one study yields $p = .051$ and another $p = .049$, then it is clear that the strength of evidence is pretty much the same in both studies (other things being equal). The arbitrary discontinuity in the *P* value scale that existed at .05 (and to a lesser extent at .01) is now seen for what it is: an artificial and unnecessary categorization imposed on a continuous measure of evidential support. The first sentence of the guideline of the American Psychological Association's Task Force on Statistical Inference [12] regarding hypothesis tests emphasizes the superiority of *P* values:

*It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual *P* value, or better still, a confidence interval. (page 599)*

The *P* value associated with a test statistic provides evidence that is relevant to the question of whether a real effect is present or not present in the observed data. Many behavioral scientists seem to believe that the *P* value represents the probability that the null hypothesis of no real effect is true, given the observed data. This incorrect interpretation has been encouraged by many textbook authors [2, 3, 11].

As outlined above, the *P* value actually conveys information about the probability of the observed outcome on the assumption that the null hypothesis is true. The fact that the observed outcome of a study would be unlikely if the null hypothesis was true does not, in itself, imply that it is unlikely that the null hypothesis is true. However, the two conditional probabilities are related, and, other things being equal, the smaller the *P* value, the more doubt is cast on the truth of the null hypothesis [4].

One and Two-tailed *P* values

In the description of a *P* value above, reference was made to 'an outcome as deviant as the outcome obtained (or even more deviant)'. Here, the deviation is from the expected value of the test statistic under the null hypothesis. But what constitutes an equally deviant or more deviant observation depends on the nature of the alternative hypothesis. If a one-tailed test is being conducted, then the *P* value must include all the outcomes between the one obtained and the end of the predicted tail. However, if a two-tailed test is being conducted, then the *P* value must include all the outcomes between the obtained one and the end of the tail in which it falls, plus all the outcomes in the other tail that are equally or more extreme (but in the other direction). Suppose a positive sample correlation, r , is observed. If this is subjected to a one-tailed test in which a positive correlation had been predicted, then the *P* value would be the probability of obtaining a sample correlation from r to $+1$ under the null hypothesis that the population correlation was zero. If the same sample correlation is subjected to a two-tailed test, then the *P* value would be the probability of obtaining a sample correlation from $-r$ to -1 under the null hypothesis that the population correlation was zero. The two-tailed *P* value would be twice as large as the one-tailed *P* value. The fact that the evidential value of a result appears to rest on a private and unverifiable decision to make a one-tailed prediction has led many behavioral scientists to turn their backs on one-tailed testing. A further argument against one-tailed tests is that, conducted rigorously, they require researchers to treat all results that are highly deviant in the nonpredicted direction as uninformative.

There are particular problems associated with the calculation of *P* values for test statistics with discrete

sampling distributions. One hotly debated issue is how the probability of the observed outcome should be included in the calculation of a *P* value. It could contribute to the *P* value entirely or in part (see [1]). Another issue concerns the calculation of two-tailed *P* values in situations where there are no equally deviant or more deviant outcomes in the opposite tail. Is it reasonable in such circumstances to report the one-tailed *P* value as the two-tailed one or should the two-tailed *P* value be doubled as it is in the case of a continuously distributed test statistic [9]? These issues are not resolved to everyone's satisfaction, as the cited sources testify.

P values and a False Null Hypothesis

When the null hypothesis is true, the sampling distribution of the *P* value is uniform over the interval 0 to 1, regardless of the sample size(s) used in the study. However, if the null hypothesis is false, then the obtained *P* value is largely a function of sample size, with an expected value that gets smaller as sample size increases. A null hypothesis is, by its nature, an exact statement, and, it has been argued, it is not possible for such a statement to be literally true [10]. For example, the probability that the mean of any two existing populations is *exactly* equal to the last decimal point, as the null hypothesis specifies, must be zero. Given this, some have argued that nothing can be learned from testing such a hypothesis as any arbitrarily low *P* value can be obtained just by running more participants. Some defense can be offered against such arguments when true experimental manipulations are involved [7], but the validity of the argument cannot be denied when it is applied to comparisons of preexisting groups or investigations of the correlation between two variables in some population. What information can *P* values provide in such situations?

The conviction that a particular null hypothesis must be untrue on *a priori* grounds does not fully specify what the true state of the world must be. The conviction that two conditions have different population means does not in itself identify which has the higher mean. Similarly, the conviction that a population correlation is not zero does not specify whether the relationship is positive or negative. Suppose that a study has been conducted that leads to the rejection

of the null hypothesis with some particular *P* value in favor of a two-tailed alternative hypothesis. In such circumstances, researchers usually go on to conclude that an effect exists in the direction indicated by the sample data. If this procedure is followed, then the probability that the test would lead to a conclusion in one direction given that the actual direction of the effect is in the other (a type 3 error [5, 6, 8]) cannot be as great as half the obtained *P* value. The *P* value, therefore, provides a useful index of the strength of evidence against a real effect in the opposite direction to the direction observed in the sample.

References

- [1] Armitage, P., Berry, G. & Matthews, J.N.S. (2002). *Statistical Methods in Medical Research*, 4th Edition, Blackwell Science, Oxford.
- [2] Carver, R.P. (1978). The case against statistical significance testing, *Harvard Educational Review* **48**, 378–399.
- [3] Dracup, C. (1995). Hypothesis testing: what it really is? *Psychologist* **8**, 359–362.
- [4] Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception, *Theory and Psychology* **5**, 74–98.
- [5] Harris, R.J. (1997). Reforming significance testing via three-valued logic, in *What if there were no Significance Tests?* L.L. Harlow, S.A. Mulaik & J.H. Steiger, eds, Lawrence Erlbaum Associates, Mahwah.
- [6] Kaiser, H.F. (1960). Directional statistical decisions, *Psychological Review* **67**, 160–167.
- [7] Krueger, J. (2001). Null hypothesis significance testing: on the survival of a flawed method, *American Psychologist* **56**, 16–26.
- [8] Leventhal, L. & Huynh, C.-L. (1996). Directional decisions for two-tailed tests: power, error rates, and sample size, *Psychological Methods* **1**, 278–292.
- [9] Macdonald, R.R. (1998). Conditional and unconditional tests of association in 2×2 tables, *British Journal of Mathematical and Statistical Psychology* **51**, 191–204.
- [10] Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy, *Psychological Methods* **5**, 241–301.
- [11] Pollard, P. & Richardson, J.T.E. (1987). On the probability of making Type I errors, *Psychological Bulletin* **102**, 159–163.
- [12] Wilkinson, L. and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations, *American Psychologist*, **54**, 594–604.

CHRIS DRACUP

Page's Ordered Alternatives Test

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Volume 3, pp. 1503–1504

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Page's Ordered Alternatives Test

Page's Test

Page's procedure [4] is a distribution free (*see Distribution-free Inference, an Overview*) test for an ordered hypothesis with $k > 2$ related samples. It takes the form of a **randomized block design**, with k columns and n rows. The null hypothesis is $H_0 : m_1 = m_2 = \dots = m_k$. This is tested against the alternative hypothesis:

$$\begin{aligned} H_1 : m_i &\leq m_j, \text{ for all } i < j \text{ and} \\ m_i &< m_j \text{ for at least one } i < j; \\ \text{for } i, j &\text{ in } 1, 2, \dots, k. \end{aligned}$$

The ordering of treatments, of course, is established prior to viewing the results of the study. Neave and Worthington [3] provided a Match test as a powerful alternative to Page's test.

Procedure

The data are ranked from 1 to k for each row. The ranks of each of the k columns are totaled. If the null hypothesis is true, the ranks should be evenly distributed over the columns, whereas if the alternative is true, the rank's sums should increase with the column index.

Assumptions

It is assumed that the rows are independent and there are no tied observations in a row. Average ranks are typically applied to ties.

Test Statistic

Each column rank sum is multiplied by the column index. The test statistic is

$$L = \sum_{i=1}^k i R_i, \quad (1)$$

where i is the column index, $i = 1, 2, 3, \dots, k$, and R_i is the rank sum for the i th column. Because the

rank sums are expected to increase directly as the column number increases, and the greater rank sums are multiplied by larger numbers, L is expected to be largest under the alternative hypotheses. The null hypothesis is rejected when $L \geq$ the critical value.

Large Sample Sizes

The mean of L is

$$\mu = \frac{nk(k+1)^2}{4}, \quad (2)$$

and the standard deviation is

$$\sigma = \sqrt{\frac{nk^2(k+1)(k^2-1)}{144}}. \quad (3)$$

For a given α , the approximate critical region is

$$L \geq \mu + z\sigma + \frac{1}{2}. \quad (4)$$

Monte Carlo simulations conducted by Fahoome and Sawilowsky [2] and Fahoome [1] indicated that the large sample approximation requires a minimum sample size of 11 for $\alpha = 0.05$, and 18 for $\alpha = 0.01$.

Example

Page's statistic is calculated with Samples 1 to 5 in the table below (Table 1), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. It is computed as a one-tailed test with $\alpha = 0.05$.

The rows are ranked, with average ranks assigned to tied ranks (Table 2).

Table 1 Sample data

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
20	11	9	34	10
33	34	14	10	2
4	23	33	38	32
34	37	5	41	4
13	11	8	4	33
6	24	14	26	19
29	5	20	10	11
17	9	18	21	21
39	11	8	13	9
26	33	22	15	31
13	32	11	35	12
9	18	33	43	20
33	27	20	13	33
16	21	7	20	15
36	8	7	13	15

2 Page's Ordered Alternatives Test

Table 2 Sample data

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
	4	3	1	5	2
	4	5	3	2	1
	1	2	4	5	3
	3	4	2	5	1
	4	3	2	1	5
	1	4	2	5	3
	5	1	4	2	3
	2	1	3	4.5	4.5
	5	3	1	4	2
	3	5	2	1	4
	3	4	1	5	2
	1	2	4	5	3
	4.5	3	2	1	4.5
	3	5	1	4	2
	5	2	1	3	4
Total	48.5	47.0	33.0	52.5	44.0

The column sums are as follows: $R_1 = 48.5$, $R_2 = 47.0$, $R_3 = 33.0$, $R_4 = 52.5$, $R_5 = 44.0$. The statistic, L , is the sum of $iR_i = 671.5$, where $i = 1, 2, 3, 4, 5$. $L = 1 \times 48.5 + 2 \times 47.0 + 3 \times 33.0 + 4 \times 52.5 + 5 \times 44.0 = 48.5 + 94.0 + 99.0 + 210.0 + 220.0 = 671.5$.

The large sample approximation is calculated with $\mu = 675$ and $\sigma = 19.3649$. The approximation is $675 + 1.64485(19.3649) + 0.5 = 707.352$. Because $671.5 < 707.352$, the null hypothesis cannot be rejected on the basis of the evidence from these samples.

References

- [1] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.
- [2] Fahoome, G. & Sawilowsky, S. (2000). Twenty non-parametric statistics, in *Annual Meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.
- [3] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.
- [4] Page, E.B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks, *Journal of the American Statistical Association*, **58**, 216–230.

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Paired Observations, Distribution Free Methods

VANCE W. BERGER AND ZHEN LI

Volume 3, pp. 1505–1509

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Paired Observations, Distribution Free Methods

Nonparametric Analyses of Paired Observations

Given two random samples, $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$, each with the same number of observations, the data are said to be paired if the first observation of the first sample is naturally paired with the first observation of the second sample, the second observation of the first sample is naturally paired with the second observation of the second sample, and so on. That is, there would need to be a natural one-to-one correspondence between the two samples. In behavioral studies, paired samples may occur in the form of a 'before and after' experiment. That is, individuals or some other unit of measurement would be observed on two occasions to determine if there is a difference in the value of some variable of interest. For example, researchers may want to determine if there is a drop in unwanted pregnancies following an intervention that consists of relevant educational materials that are being disseminated. It is possible to randomize (*see* **Randomization**) some communities or schools to receive this intervention and others not to, but an alternative design is the one that uses each school or community as its own control.

That is, one would make the intervention available to all the schools in the study, and then compare their rate of unwanted pregnancies before the intervention to the corresponding rate after the intervention. In this way, the variability across schools is effectively eliminated. Because comparisons are most precise when the maximum number of sources of extraneous variation is eliminated, it is true in general that paired observations can be used to eliminate those sources of extraneous variation that fall within the realm of variability across subjects. Note that actual differences may not exist between two populations, yet the presence of extraneous, sampling, sources of variation (confounding) may cause the illusion of such a difference. Conversely, true differences may be masked by the presence of extraneous factors [4].

Consider, for example, two types of experiment that could be conducted to compare two types of sunscreen. One method would be to select two different samples of subjects, one of which uses

sunscreen A, while the other uses sunscreen B. It may turn out, either by chance or for some more systematic reason such as self-selection bias, that most of the individuals who use sunscreen A are naturally less sensitive to sunlight. In such a case, a conclusion that individuals who received sunscreen A had less sun damage would not be attributable to the sunscreen itself. It could just as easily reflect the underlying differences between the groups. Randomization is one method that can be used to ensure the comparability of the comparison groups, but while randomization does, eliminate self-selection bias and some other types of biases, it cannot, in fact, ensure that the comparison groups are comparable even in distribution [1].

Another method that can be used to compare the types of sunscreen is to select one sample of subjects and have each subject in this sample receive both sunscreens, but at different times. The unit of measurement would then be the combination of a subject and a time point, as opposed to the more common case in which the unit of measurement is the subject without consideration of the time point. One could, for example, randomly assign half of the subjects to receive sunscreen A first and sunscreen B later, whereas the other half of the subjects is exposed to the sunscreens in the reverse order. This is a classical **crossover design**, which leads to paired observations. Specifically, the two time points for a given subject are paired and can be compared directly. One concern with this design is the potential for carryover effects (*see* **Carryover and Sequence Effects**).

Yet another design would consist of again selecting one sample of subjects and having each subject in this sample receive both sunscreens, but now the exposure to the sunscreens is simultaneous. It is not time that distinguishes the exposure of a given individual to a given sunscreen but rather the side of the face. For example, each subject could be randomized to either sunscreen A applied to the left side of the face and sunscreen B applied to the right side of the face or sunscreen B applied to the left side of the face and sunscreen A applied to the right side of the face. After a specified length of exposure to the sun, the investigator would measure the amount of damage to each half of the face, possibly by recording a lesion count. This design also leads to paired observations, as the two sides of a given individual's face are paired.

2 Paired Observations, Distribution Free Methods

If the side of the face to which sunscreen A was applied tended to be less damaged overall, regardless of whether this was the left side or the right side, then one could fairly confidently attribute this result to sunscreen A, and not to preexisting differences between the groups of subjects randomized to the pairings of sides of the face and sunscreens, because both sunscreens were applied to equally pigmented skin [4]. Note that if this design is used in a multinational study, then it might be prudent to stratify the randomization by nation. Not only could different nations have different levels of exposure to the sun (contrast Greenland, for example, to Ecuador), but also some nations mandate that drivers be situated in the left side of the car (for example, the United States), whereas other nations mandate that drivers be situated in the right side of the car (for example, the United Kingdom). Clearly, this will have implications for which side of the face has a greater level of exposure to the sun.

By using paired observations, Burgess [3] conducted a study to determine weight loss, body composition, body fat distribution, and resting metabolic rate in obese subjects before and after 12 weeks of treatment with a very-low-calorie diet (VLCD) and to compare hydrodensitometry with bioelectrical impedance analysis. The women's weights before and after the 12-week VLCD treatment were measured and compared. This is also a type of paired observations study. Likewise, Kashima et al. [8] conducted a study of parents of mentally retarded children in which a media-based program presented, primarily through videotapes and instructional manuals, information on self-help skill teaching. In this research study, paired observations were obtained. Before and after the training program, the Behavioral Vignettes Test was administered to the parents. We see that paired observations may be obtained in a number of ways [4] as follows.

1. The same subjects may be measured before and after receiving some treatment.
2. Pairs of twins or siblings may be assigned randomly to two treatments in such a way that members of a single pair receive different treatments.
3. Material may be divided equally so that one half is analyzed by one method and one half is analyzed by the other.

The techniques that can be used to analyze paired observations can be classified as parametric or nonparametric on the basis of the assumptions they require for validity. A classical parametric statistical method, for example, is the paired t Test (see **Catalogue of Parametric Tests**). The nonparametric alternatives include Fisher's randomization test, the McNemar Chi-square test, the sign test, and the Wilcoxon Signed-Rank test (see **Distribution-free Inference, an Overview**). The paired t Test is used to determine if there is a significant difference between two samples, but it exploits the natural pairing of the data to reduce the variability. Specifically, instead of considering the variability within the X and Y samples separately, we consider the difference d_i between the paired observations for the i th subject as the variable of interest and compute the variability of d_i . This will often lead to a reduction in variability. The paired t Test assumes or requires (for validity) the following.

1. Independence.
2. Normality.

Each of these assumptions requires some elaboration. Regarding the first, we note that it is not required that the X_i 's be independent of the Y_i 's. In fact, if they were, then this would defeat the purpose of the pairing. Rather, it is the differences, or the d_i 's, that need to be independent of each other. Regarding the second assumption, we note that the normality of X_i and Y_i are not required because the analysis depends on only the d_i 's. However, it is required that the d_i 's follow the normal distribution.

The following example shows a case in which the X 's and Y 's are not normally distributed but the d 's are at least close. A researcher wanted to find out if he could get closer to a squirrel than the squirrel was to the nearest tree before the squirrel would start to run away [11]. The observations follow in Table 1. The normal QQ plots (see **Probability Plots**) of the X 's, Y 's, and d 's are shown in Figure 1, Figure 2, and Figure 3, respectively. The distribution of the distances from the squirrel to the person (X) appears to be reasonably normally distributed, but the distances from the squirrel to the tree (Y) are far from being normally distributed. However, the differences of the paired observations (d) appear to meet the required normality condition.

Table 1 Distances (in inches) from person and from tree when squirrel started to run. Reproduced from Witmer, S. (2003). *Statistics for the Life Sciences*, 3rd Edition, Pearson Education, Inc. [11]

Observations	From person X	From tree Y	Difference $d = X - Y$
1	81	137	-56
2	178	34	144
3	202	51	151
4	325	50	275
5	238	54	184
6	134	236	-102
7	240	45	195
8	326	293	33
9	60	277	-217
10	119	83	36
11	189	41	148
Mean	190	118	72
SD	89	101	148

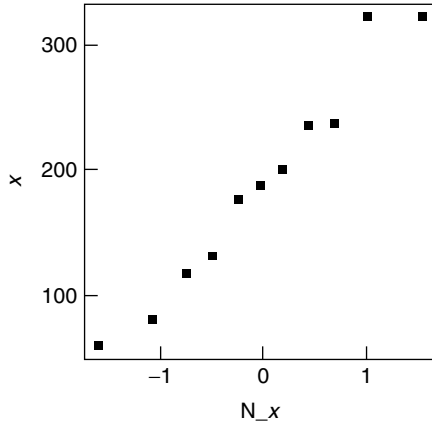


Figure 1 QQ plot of variable X

In general, the null hypothesis can often be expressed as follows.

$$H_0 : E[X|i] = E[Y|i]. \quad (1)$$

That is, given the i th block, the expected values of two distributions are the same. This conditional null hypothesis can be simplified by considering the differences:

$$H_0 : E(d) = 0 \quad (2)$$

To test this null hypothesis, we can use the one-sample t Test. The test statistics based on the t

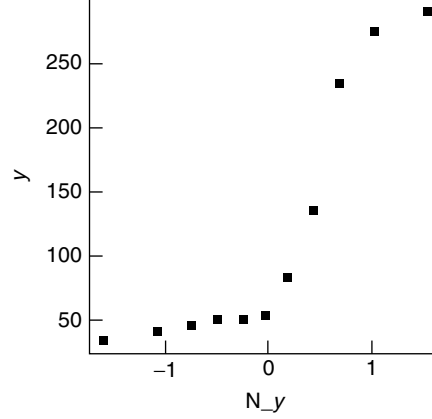


Figure 2 QQ plot of variable Y

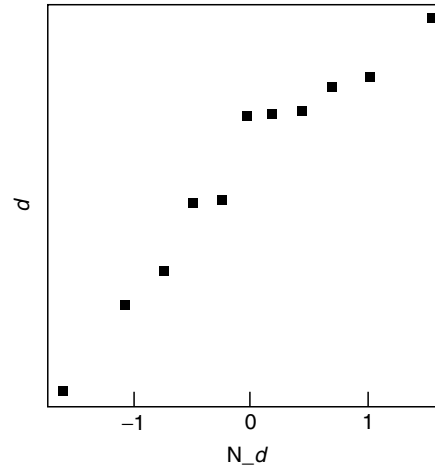


Figure 3 QQ plot of difference d

distribution are:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}, \quad \text{where } \bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$$

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}, \quad (3)$$

The rejection region is determined using the t distribution with $n - 1$ degrees of freedom, where n is the number of pairs. If the pairing is ignored, then an analysis of differences in means is not

4 Paired Observations, Distribution Free Methods

valid because it assumes that the two samples are independent when in fact they are not.

Even when the pairing is considered, this t Test approach is not valid if either the normality assumption or the independence assumption is unreasonable, which is often the case. Hence, it is often appropriate to instead use nonparametric analyses. **Fisher** introduced the first nonparametric test, the widely used Fisher randomization test, in his analysis of Darwin's data in 1935 [5] (see **Randomization Based Tests**). Darwin planted 15 pairs of plants, one cross-fertilized and the other self-fertilized, over four pots. The number of plant pairs varied from pot to pot. Darwin's experimental results follow in Table 2.

That is, Fisher argued that under the null hypothesis, the cross- and self-fertilized plants in the i th pair have heights that are random samples from the same distribution, so this means that each difference between cross- and self-fertilized plant heights could have appeared with a positive or negative sign with equal probability. There were 15 paired observations, leading to 15 differences, and $\sum_{i=1}^{15} d_i = 314$. Also, there were a total of $2^{15} = 32768$ possible arrangements of signs with the 15 differences obtained. Assuming that all of these configurations are equally likely, only 863 of these arrangements give a total difference of 314 or more, thus giving a probability of $863/32768 = 2.634\%$

Table 2 Darwin's results for the heights of cross- and self-fertilized plants of *Zea mays*, reported in 1/8th's of an inch. Reproduced from Fisher, R.A. (1935). *The Design of Experiments*, 1st Edition, Oliver & Boyd, London [5]

Pot	Cross-fertilized	Self-fertilized	Difference
I	188	139	49
	96	163	-67
	168	160	8
II	176	160	16
	153	147	6
	172	149	23
III	177	149	28
	163	122	41
	146	132	14
	173	144	29
	186	130	56
IV	168	144	24
	177	102	75
	184	124	60
	96	144	-48

for that one-side test [7]. That is, the P value is 0.026.

Another nonparametric technique that can be used to analyze paired observations is the Wilcoxon Signed-Rank test (see **Wilcoxon-Mann-Whitney Test**). This test can be used in any situation in which the d 's are independent of each other and come from a symmetric distribution; the distribution need not be normal. The null hypothesis of 'no difference between population' can be stated as:

$$H_0 : E(d) = 0. \quad (4)$$

The Wilcoxon Signed-Rank test is conducted as follows. Compute the differences between two treatments in each experimental subject, rank the differences according to their magnitude (without regard for sign), then reattach the sign to each rank. Finally, sum the signed ranks to obtain the test statistic W . Because this procedure is based on ranks, it does not require making any assumptions about the nature of the population. If there is no true difference between the means of the units within a pair, then the ranks associated with the positive changes should be similar in both number and magnitude to the ranks associated with the negative changes. So the test statistic W should be close to zero if H_0 is true. On the other hand, W will tend to be a large positive number or a large negative number if H_0 is not true. This is a nonparametric procedure, which means that it is valid in more situations than the parametric t Test is [2].

It is generally recommended that the Wilcoxon Signed-Rank test be used instead of the paired t Test if the normal distributional assumption does not hold or the sample size is small (for example, $n < 30$), but the comparison between these two analyses actually differ not only in the reference distribution (exact versus based on the normal distribution) (see **Exact Methods for Categorical Data**) but also with regard to the test statistic (mean difference vs mean ranks). This means that the usual recommendation is suspect for two reasons. First, given the availability of an exact test, it is hard to imagine justifying an approximation to that exact test on the basis of the goodness of the approximation [2]. Second, the t Test is not an approximation to the Wilcoxon test anyway. So a better recommendation would be to use an exact test whenever it is feasible to do so, regardless of how normally distributed the data appear to be, and to exercise discretion in the selection of the

test statistic. If interest is in the ranks, then use the Wilcoxon test. If the raw data are of interest, then use the exact version of the t Test (i.e., use the t Test statistic, but refer it to its exact permutation reference distribution, as did Fisher). These are not the only reasonable choices, as other scores, including normal scores, could also be used.

The **sign test** is another robust nonparametric test that can be used as an alternative to the paired t Test. The sign test gets its name from the fact that pluses and minuses, rather than numerical values, are considered, counted, and compared. This test is valid in any situation in which the d 's are independent of each other. The sign test focuses on the median rather than mean. So the null hypothesis is:

$$H_0 : \text{The median difference is zero.}$$

An alternative way of stating the null hypothesis is:

$$H_0 : p(x_i > y_i) = p(x_i < y_i) = 0.5. \quad (5)$$

That is,

$$H_0 : p(+) = p(-) = 0.5. \quad (6)$$

So, the sign test is distribution free. Its validity does not depend on any conditions about the form of the population distribution of the d 's. The test statistic is the number of plus signs, which follows the **binomial distribution** (see **Catalogue of Probability Density Functions**) with parameters n (the sample size excluding ties) and $p = 0.5$ (the null probability) if H_0 is true.

The McNemar Chi-square test studies the paired observations on the basis of a dichotomous variable, and is often used in case-control studies. The data can be displayed as a 2×2 table as follows.

		Control Outcome	
Case Outcome		+	-
		a	b
	-	c	d

The test statistics is:

$$\text{McNemar } \chi^2 = \frac{(|b - c| - 1)^2}{b + c}. \quad (7)$$

Under the null hypothesis, this test statistic follows the χ^2 distribution with one degree of freedom.

Besides these classical techniques, some more modern techniques have been developed on the basis of paired observations. Wei proposed the idea of a test for interchangeability with incomplete paired observations [10]. Gross and Lam [6] studied the paired observations from a survival distribution and developed hypothesis tests to investigate the equality of mean survival times when the observations came in pairs, for example, length of a tumor remission when a patient received a standard treatment versus length of a tumor remission when the same patient, at a later time, received an experimental treatment. Lipscomb and Gray studied a connection between paired data analysis and regression analysis for estimating sales adjustments [9].

References

- [1] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias, *Journal of Modern Applied Statistical Methods* **2**(1), 80–86.
- [2] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**, 74–82.
- [3] Burgess, N.S. (1991). Effect of a very-low-calorie diet on body composition and resting metabolic rate in obese men and women, *Journal of the American Dietetic Association* **91**, 430–434.
- [4] Daniel, W.W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences*, 7th Edition, John Wiley & Sons, New York.
- [5] Fisher, R.A. (1935). *The Design of Experiments*, 1st Edition, Oliver & Boyd, London.
- [6] Gross, A.J. & Lam, C.F. (1981). Paired observations from a survival distribution, *Biometrics* **37**(3), 505–511.
- [7] Jacquez, J.A. & Jacquez, G.M. (2002). Fisher's randomization test and Darwin's data – a footnote to the history of statistics, *Mathematical Biosciences* **180**, 23–28.
- [8] Kashima, K.J., Baker, B.L. & Landen, S.J. (1988). Media-based versus professionally led training for parents of mentally retarded children, *American Journal on Mental Retardation* **93**, 209–217.
- [9] Lipscomb, J.B. & Gray, J.B. (1995). A connection between paired data analysis and regression analysis for estimating sales adjustments, *The Journal of Real Estate Research* **10**(2), 175–183.
- [10] Wei, L.J. (1983). Test for interchangeability with incomplete paired observations, *Journal of the American Statistical Association* **78**, 725–729.
- [11] Witmer, S. (2003). *Statistics for the Life Sciences*, 3rd Edition, Pearson Education, Inc.

VANCE W. BERGER AND ZHEN LI

Panel Study

JAAK BILLIET

Volume 3, pp. 1510–1511

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Panel Study

In a typical **cross-section survey**, people are interviewed just once. By contrast, in a panel study, exactly the same people are interviewed repeatedly at multiple points in time. Panel surveys are much more suitable for the study of attitude change and for the study of causal effects than repeated cross-sections or surveys with retrospective questions. Statistical analysis of turn-over tables that emerge from repeated interviewing of the same respondents may reveal the amount of individual-level attitude change that occurred between the waves of the survey. This individual level change is unobservable in repeated cross-sections [12].

The strength of panel studies is eminently illustrated by the *Intergenerational Panel Study of Parents and Children* (IPS) of the University of Michigan. This is an eight wave 31-year intergenerational panel study (1961–1993) of young men, young women, and their families [1]. These data are ideally suited to study the relationship between attitudes, family values, and behavior for several reasons. The data are longitudinal and span the entire lives of the children, allowing insights into causation by relying on temporal ordering of attitudes and subsequent behavior. Moreover, these data include measures of children's attitudes and their mother's attitudes in multiple domains at crucial moments during their life. Further, the data contain measures of children's subsequent cohabitation, marriage, and childbearing patterns, as well as their education and work behavior. Finally, the data contain detailed measurement of the families in which the children were raised [2].

Whereas panel studies provide unique opportunities, there are also dark sides. Long-term panel studies are complicated by the expense and difficulty of finding respondents who have moved in intervening years (location of respondents). Sometimes, there is a considerable amount of nonresponse at the first wave, which may rise to over 30% because of the long-term engagement that respondents are unwilling to make. However, more frequently, problems occur during the second or later waves. Drop out from the original sample units in subsequent waves may lead to considerable bias. For example, in three waves of the Belgian Election surveys (1991–1999), it is shown that the nonresponse in the second and third waves of the panel is not random and can be

predicted by some background characteristics, task performance [8], and attitudes of both interviewers and respondents in previous waves [9]. In this panel survey, we found that the less interested and more politically alienated voters were more likely to drop out and that, consequently, the population estimates on the willingness to participate in future elections were strongly biased toward increasing willingness to participate over time.

Simple assessment of the direction of nonresponse bias can be done by comparing the first-wave answers given by respondents who were successfully reinterviewed to the first-wave answers of those who drop out. But this is only a preliminary step. In current research, much of the attention is paid to the nature of nonignorable nonresponse [7] and to statistical ways of correcting the estimates [3]. Refreshment of the panel sample with new respondents in subsequent waves [4] and several weighting procedures and imputation methods are proposed in order to adjust the statistical estimates for attrition bias [3, 10] (*see Missing Data*). The availability of a large amount of information about the nonrespondents in later waves of a panel is certainly a major advantage of the panel over 'fresh' nonresponse. Designers of longitudinal surveys should seriously consider adding variables that are useful to predict location, contact difficulty, and cooperation propensity in order to improve the understanding of processes that produce nonresponse in later waves of panel surveys. This may also lead to the reduction of nonresponse rates and to more effective adjustment procedures [7].

High costs for locating respondents who have moved, wrong respondent selection, and the problems related to panel attrition are not the only deficits of panels. Panel interviews are repeated measurements that are not independent at the individual level. The high expectations regarding the methodological strength of panel studies for analyzing change are somewhat tempered by this. Panel studies within quasi-experimental designs, aimed at studying the effects of a specific independent variable (or intervention), are confronted with the presence of unobserved intervening variables, or by the effect of the first observation on subsequent measurement. Some of these challenges, but not all, can be solved by an appropriate research design like the '*simulated pretest–posttest design*' [5]. In the case of a series of repeated measurements over time, confusion is possible about the time order of a causal effect, if no

stringent assumptions are made about the duration of an effect. The unreliability of the measurements is another problem that needs consideration by those who are interested in change. Even a small amount of random error in repeated measurements may lead to faulty conclusions about substantial change in turnover tables even when nothing has changed [5, 6]. In order to distinguish unreliability (random error) from real (systematic) change in the measurements, at least three wave panels are necessary unless repeated measurements within each wave are used [11].

References

- [1] Axinn, W.G. & Arland, T. (1996). Mothers, children and cohabitation: the intergenerational effects of attitudes and behavior, *American Sociological Review* **58**(2), 233–246.
- [2] Barber, J.S., Axinn, W.G. & Arland, T. (2002). The influence of attitudes on family formation processes, in *Meaning and Choice-Value Orientations and Life Course Decisions*, Monograph 37, R. Lesthaeghe, ed., NIDI-CBGS Publications, The Hague and Brussels, pp. 45–95.
- [3] Brownstone, D. (1998). Multiple imputation methodology for missing data, non-random response, and panel attrition, in *Theoretical Foundations of Travel Choice Modeling*, T. Gärling, I. Laitila & K. Westin, eds, Elsevier Biomedical, Amsterdam, pp. 421–450.
- [4] Brownstone, D., Thomas, F.G. & Camilla, K. (2002). Modeling nonignorable attrition and measurement error in panel surveys: an application to travel demand modeling, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little, eds, John Wiley & Sons, New York, pp. 373–387.
- [5] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Chicago.
- [6] Hagenaars, J.A. (1975). *Categorical Longitudinal Data: Loglinear Panel, Trend, and Cohort Analysis*, Sage Publications, Newbury Park.
- [7] Lepkowski, J.M. & Mick P.C. (2002). Nonresponse in the second wave of longitudinal household surveys, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little, eds, John Wiley & Sons, pp. 259–272.
- [8] Loosveldt, G. & Carton, A. (2000). An empirical test of a limited model for panel refusals, *International Journal of Public Opinion Research* **13**(2), 173–185.
- [9] Pickery, J., Loosveldt, G. & Carton, A. (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys, *Sociological Methods and Research* **29**(4), 509–523.
- [10] Rubin, D.B. & Elaine, Zanutto (2002). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge & R.J.A. Little, eds, John Wiley & Sons, New York, pp. 389–402.
- [11] Saris, W.E. & Andrews, F.M. (1991). Evaluation of measurement instruments using a structural equation approach, in *Measurement Errors in Surveys*, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman, eds, John Wiley & Sons, New York, pp. 575–599.
- [12] Weinsberg, H.F., Krosnick J.A. & Bowen B.D. (1996). *An Introduction to Survey Research, Polling, and Data Analysis*, 3rd Edition, Sage Publications, Thousand Oaks.

JAAK BILLIET

Paradoxes

VANCE W. BERGER AND VALERIE DURKALSKI

Volume 3, pp. 1511–1517

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Paradoxes

Although scientists have spent centuries defining mathematical systems, self-contradictory conclusions continue to arise in certain scenarios. Why, for example, do larger data sets sometimes give opposite conclusions from smaller data sets, and could the larger data sets be incorrect? How can a nonsmoker's risk of heart disease be lower overall if a particular smoker who exercises fares better than a particular nonsmoker who exercises, and a particular smoker who does not exercise fares better than a particular nonsmoker who does not exercise? How is it that various baseline adjustment methods can give different results, or that the adjustment applied to artificially small P values because of multiplicity can result in even smaller P values? How can a data set have a low kappa statistic (*see* **Rater Agreement – Kappa**) and yet have high values of observer agreement? How can (almost) each element in a population be above the population mean? Below are some of the more common paradoxes encountered in observational and experimental studies and explanations of why they may occur in actual practice.

Simpson's Paradox

Baker and Kramer [1] provide the following hypothetical example. Among males, there are 120/200 (60%) responses to A and 20/50 (40%) responses to B, while among females there are 95/100 (95%) responses to A and 221/250 (85%) responses to B. Overall, then, there are 215/300 (72%) responses to A and 241/300 (80%) responses to B, so females respond better than males to either treatment, and A is better than B in each gender, yet B appears to be better than A overall. This apparent paradox occurs because gender is associated with both treatment and outcome. Specifically, women have both a higher survival rate than men and a better chance to receive Treatment B, so ignoring gender makes Treatment B look better than Treatment A, when in fact it is not.

A reversal of the direction of an association between two variables when a third variable is controlled is referred to as Simpson's paradox (*see* **Two by Two Contingency Tables**) [25], and is commonly found in both observational and experimental studies

that partition data. The implications of this paradox are that different ways of partitioning the data can produce different correlations that appear to be discordant with the initial correlations. Once the invalidity is revealed, it can give evidence to a causal relationship.

The Birth To Ten (BTT) Study, conducted in South Africa, provides one example of the potential for intuitive reasoning to be incorrect [20]. A birth cohort was formed to identify cardiovascular risk factors in children living in an urban environment in South Africa. A survey to collect information on health-related issues was performed when the children were 5 years of age. One measured association is between the presence/absence of medical aid in the cohort that responded at 5 years and a comparative cohort that had children but was not in the initial survey ('No Trace' cohort). The overall **contingency table** of the presence or absence of medical aid between the two cohorts shows with statistical significance that the probability of the 5-year cohort having medical aid is lower than the probability of the 'Not Traced' cohort having medical aid. However, when the population is partitioned by race, the reverse is found.

There is a higher probability of having medical aid for the 5-year cohort in each racially homogeneous population, which is contradictory to the conclusion based on the overall population. The contradiction is referred to as the 'reversal of inequalities' or 'Simpson's paradox', and appears in the hypothetical data presented by Heydtmann [12], the BBT data set [20], and other scenarios presented in the medical literature [2, 22, 24]. It can also be shown that even if the numbers uniformly increase (retaining all relative proportions), the contradictions observed in these fractions remain constant [17]. This means that the best method for avoiding Simpson's paradox is not to increase the sample size but rather to carefully research the current literature when designing your study and consult with experts in the therapeutic field to define appropriate outcomes and prognostic variables, and tabulate the data within each level of each key predictor.

Lord's Paradox

In experimental studies that compare two treatments (e.g., placebo vs. experimental treatment), often the investigator is interested in observing the effect of

an intervention on a specific continuous outcome variable (*see* **Clinical Trials and Intervention Studies**). Experimental units (i.e., subjects) are randomly assigned to each treatment group and a comparison of the two groups, while controlling for baseline factors, is conducted. There are different yet related methods for controlling for baseline factors, which under certain situations can yield different conclusions. This occurrence is commonly referred to as **Lord's Paradox** [15, 16]. For example, in depression studies, one may define the primary outcome variable as Beck's Depression Index (BDI). One would be interested in comparing BDI scores between two groups after an intervention.

Upon study completion, the two groups may be compared with an unadjusted analysis (i.e., a comparison based on a two-sample *t* Test of the means). However, as Lord points out in his review of a weight gain study in college students [15], this approach ignores the potential correlation between outcome and baseline variables. Adjustments for baseline are meant to control for pretreatment differences and remove the variation in response on the basis of baseline factors. This is important for obtaining precise estimates and treatment comparisons [15, 25]. Options for adjustment include: (1) subtracting the baseline from the outcome variable (i.e., analyzing the change score, or delta), (2) dividing the outcome variable by the baseline (i.e., analyzing the percent change), or (3) using baseline as a covariate in an **analysis of covariance** (ANCOVA) [4, 5] or a suitable **nonparametric alternative**. Approach 1 leads to the descriptive statement 'On average the outcome decreases *X* amount for Treatment Group A, whereas it decreases *Y* amount for Treatment Group B'.

Approach 2 concludes 'On average the outcome decreases *X*% for Treatment Group A, whereas it decreases *Y*% for Treatment Group B', and Approach 3 concludes 'On average the outcome for Treatment Group A decreases as much after intervention as that of Treatment Group B for those subjects with the same baseline'. Senn [25] discusses the pros and cons of each of these approaches and suggests that ANCOVA is the 'best' approach. However, for this discussion we will focus on reasons why related approaches can give different results rather than which approach is most appropriate.

Many authors offer insight into why results may vary based on technique of choice [13, 15, 16, 29]. Using pulmonary function test data, Kaiser [13]

demonstrates different statistical inferences on the same data set between percent change and change score results. Wainer [29] illustrates how the three adjustment options can alter the conclusion of heart rate changes between young and old rats and blur the causal effect of the treatment. He concludes that the level of discrepancy between methods depends on different untestable assumptions such as 'if an intervention was not applied then the response would be similar to baseline'. There are many scenarios where this specific assumption does not hold owing to placebo effects or unobservable variables. For this reason alone, a certain level of judgment is needed when choosing the adjustment approach. The goal should be to choose the method that allows baseline to be independent (or close to independent) of the adjusted response. If dependence is present, then there is a potential for decreased sensitivity in estimating treatment differences and a decrease in useful summary statistics.

Correlation, Dependence, and the Ecological Fallacy

Consider a study of age and income conducted on two groups, one of which was young professionals, less than 5 years from earning their medical or law degree. The other group consists of older unskilled laborers, near to retirement age. It is reasonable to suppose that within each group the trend would be toward increased income with increased age, but any such effect within the groups would be overwhelmed by the association across the groups if the two groups are combined. Overall, then, the older the worker, the more likely the worker would be to fall into the lower income group. Hence, income and age are inversely associated, in contrast to the within-group direction of the association. This is nonsensical correlation due to pooling [11].

Ecologic studies in which the experimental or observational unit is a defined population are conducted in order to collect aggregate measures and provide a summary of the individuals within a group. This design can be found in studies where it is too expensive to collect individual data or primary interest is in population responses (i.e., air pollution studies or intersection car wrecks). Conclusions based on ecologic studies should be interpreted with caution when expressed on an individual level. Otherwise, the

researcher may fall prey to the ecological fallacy, in which an association observed between variables on an aggregate level does not necessarily represent the association that exists at an individual level [27].

Davis [8] discusses an ecological fallacy seen in car-crash rates and the positive correlation with variation in speed. A number of studies show a direct causal relationship between driving speeds and crash rate and conclude that both slower and faster drivers are more likely to have a car wreck than those who drive within the average speed limit. The unit of analysis for these studies is a group of individuals at a specific intersection or a specific highway segment. Davis [8] discusses the potential role of the ecologic fallacy in these studies and illustrates through a series of hypothetical examples that correlations seen in group analyses do not always provide support for conclusions at the individual level. In particular, whether the individual risk is monotonically increasing, decreasing, or U-shaped, the group correlation between the risk of a car crash and variance in speed is positive. A hypothetical data set illustrates what may occur under this fallacy. Suppose that a researcher is observing three inner-city hospitals to understand the relationship between tumor size and age of patient. Five patients are observed at each hospital. The mean ages in the three hospitals are 45.8, 47.0, and 53.0 years, respectively. Respective incidences of tumors ≥ 6 mm are 0.4, 0.6, and 0.8, indicating a positive correlation between tumor size and age (Figure 1). However, when observing the individual data (Figure 2), it appears that patients under the age of 50 have larger size tumors.

Piantadosi et al. [23] expand on this topic and offer a concise summary and reasoning of why the

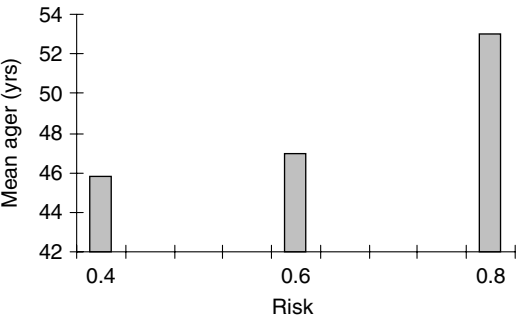


Figure 1 Mean age and incidence of tumor size ≥ 6 mm

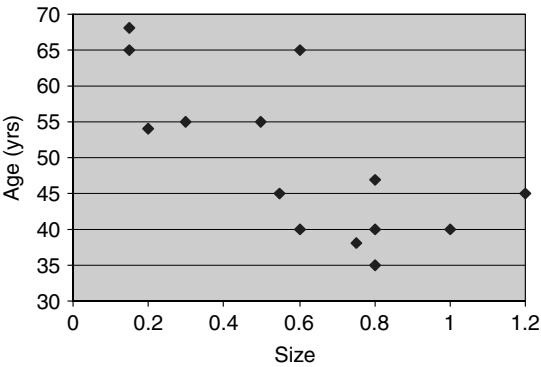


Figure 2 Individual age and tumor size

fallacy exists. The conclusion is that despite the fact that aggregate data are somewhat easier to collect (in certain scenarios), correlations based on ecologic data should be interpreted with caution since they may not represent the associations that truly exist at the individual level. This is another way of saying that correlated variables may be conditionally uncorrelated. Another example would be height and weight, which tend to be correlated, but when conditioning on both the individual and the day, the height tends to be fairly constant, and certainly independent of the weight. It is also possible for uncorrelated variables to be conditionally correlated. This could be a variation on Simpson’s paradox, with pooling masking a common correlation across strata (but instead of reversing it, simply resulting in no correlation). It could also represent compensation, with the direction of association varying across the strata.

Meehl’s paradox is also rooted in correlation, except that it involves multiple correlation (*see R-squared, Adjusted R-squared*). The issue here is that two **binary variables**, neither one of which predicts a binary outcome at all, can together predict it perfectly. Von Eye [28] offers an example with two binary predictors, each assuming the values true or false, and a binary outcome, taking the values yes or no. There are then $2 \times 2 \times 2 = 8$ possible sets of values, and these eight patterns have frequencies as follows:

TTY	20	TTN	0	TFY	0	TFN	20
FTY	0	FTN	20	FFY	20	FFN	0

Neither binary predictor is predictive of the outcome, but the outcome is yes if the two predictors

agree, and no otherwise. So in combination, the two predictors predict the outcome perfectly, truly a case of the whole being greater than the sum of the parts.

High Proportion of Agreement yet Low Kappa

Cohen's Kappa (*see Rater Agreement – Kappa*) is a chance-adjusted measure of agreement between two or more observers [10]. Although it is used in a variety of interrater agreement observer studies, there are situations that present relatively low kappa values yet high proportions of overall agreement. This paradox is due to the prevalence of the observed value and the imbalance in marginal totals of the contingency table [7, 9, 14]. If the prevalence of the variable of interest is low, then it should be expected that agreement will be highly skewed. Thus, it is important to take prevalence into consideration when designing your study and defining the sampled population.

The Two-sided P value Equals the One-sided P value

One-sided tests are based on one-sided rejection regions. That is, the rejection regions consists of the observed value plus either outcomes to the right of (larger values than) the observed value or outcomes to the left of (smaller values than) the observed value, but not both. Two-sided tests, in contrast, allow values to be on either side of the observed value, as long as they are more extreme, or further away from what one might expect under the null hypothesis. The critical values for a two-sided test are derived from the specified type I error rate (α) such that the error rate can be divided in half for a two-sided test ($\alpha/2$). When dealing with symmetrical distributions, the two-sided P value is two times the one-sided P value, and so for any given alpha level, the two-sided test is the more conservative one. This is generally understood, and the doubling that applies to symmetric distributions appears to be universally accepted (or assumed), because in practice it is common to conduct a one-sided test at a significance level that is half the usual two-sided one. That is, if a two-sided test would be performed at the 0.05 level, then the corresponding one-sided test would be performed at the 0.025 level [3]. Yet,

in the grand scheme of things, very few distributions are technically symmetric.

Some distributions are, in fact, very far from symmetric. Neuhauser [21] discussed one such hypothetical data set, in which there were seven observations overall, and two treatment groups, denoted A and B. In Group A, the observations were 7, 5, and 5. In Group B, the observations were 4, 3, 3, and 2. This is already the most extreme configuration possible with this set of numbers. That is, any permutation of these numbers will result in a less extreme deviation from the null expected outcome of common means across the groups. As such, the critical region of a permutation test (*see Permutation Based Inference*) consists of only the observed outcome, with null probability .018, and this is true whether we consider a one-sided test or a two-sided test. So either way, the P value is .018, yet this appears to be much more impressive if the test was planned as two-sided.

Same Data, Different Conclusion

Suppose that one wishes to test to see if a given population, say those who seek help in quitting smoking, are as likely to be male as female, against the alternative hypothesis that males are better represented among this population. But resources are scarce, so the study needs to be as small as possible. One could sample nine members of the population, ascertain the gender of each, and tabulate the number of males. The probability of finding all nine males, if the genders are truly equally likely, is $1/512$. The probability of finding eight males, if the genders are truly equally likely, is $9/512$. The probability of finding eight or nine males, if the genders are truly equally likely, is $10/512$, which is less than the customary significance level of 0.05 (and it is even less than half of it, 0.025, which is relevant as this is a one-sided test), so this is one way to construct a valid test of the hypothesis. Suppose that one team of investigators uses this test, finds eight males, and reports a one-sided P value of $10/512$.

Another research team decides to stop after six subjects, for an interim analysis. The probability of all six being male, under the null hypothesis, is $1/64$. In fact, they found only five males among the first six, and so continued on to nine subjects altogether. Of these nine, eight were males, the same as for the previously mentioned study. But now the P value

has to account for the interim analysis. The critical region consists of the observed outcome, eight of nine males, plus the more extreme outcome, six of six males (as opposed to nine of nine males in the previous experiment) which is still more extreme, but no longer a possible outcome. This probability is then $1/64 + (6/64)(1/8) = 14/512$. This means that the same data are observed but a different P value is reported. The reason for this apparent paradox is that a different system was used for ranking the outcomes by extremity. In the first experiment, 8/9 was second place to only 9/9, but in the second experiment 8/9 was behind 6/6, even if the 6/6 had continued and turned into 6/8 or even 6/9. So 6/6 could have turned into 6/9 had there been no interim analysis, and this would certainly have been ranked as less extreme than 8/9, yet an implication of the interim analysis and its associated stopping rule is that 6/6 is ranked more extreme than 8/9, no matter what the unobserved three outcomes would have been.

Curious Adjusted P values

Consider a study comparing two active treatments, A and B, and a control group, C. One wishes to compare A to B, and A to C. Suppose that the data are as follows: Three successes out of four (75%) subjects with A, one of four (25%) with B, and zero of 2000 (0%) with C. Without adjusting for multiplicity, the raw P value for comparing A to B is .9857. However, we would want to adjust for the multiple tests, to keep this P value from being artificially low. The adjustment performed by PROC MULTTEST in SAS would produce an adjusted P value of .0076. Keep in mind the purpose for the adjustment and the expectation that the P value would become larger after adjustment. Instead, it became smaller, and in fact so much smaller that it went from being nowhere near significant to highly significant. So what went wrong? As explained by Westfall and Wolfinger [30], the reference distribution is altered by consideration of the third group, and this creates the paradox.

What Does the Mean Mean?

It is possible, if a calibration system is obsolete, for every member of a population to be above average, or above the mean. For example, the IQ was initially developed so that an IQ of 100 was

considered average, but it is unlikely that an IQ of 100 represents the average today. Standard ranges may be established for any quantity on the basis of a sample relevant to that quantity, but as time elapses, the mean may shift to the point that each subject is above the mean. Even if no time elapses, it is still possible for almost every subject to be on one side of the mean. For example, the income distribution is skewed by a few very large values. This will tend to drive up the mean to the point that almost nobody has an income that is above average. Finally, the same set of data may give rise to two different yet equally valid values for the mean. Consider, for example, a school with four classes. One class has 30 students and the other three classes have 10 students each. What, then, is the average classroom size? Counting classes gives an average of $(30 + 10 + 10 + 10)/4 = 15$, but suppose that there is no overlap between students taking any two classes. That is, there are 60 students in all and each takes exactly one class. One could survey these students, asking each one about the size of their class. Now 30 of these students would answer '30', whereas the other 30 would answer '10'. The mean would then be $[(30)(30) + (10)(10) + (10)(10) + (10)(10)]/60 = 20$, to reflect the weighting applied to the class sizes.

Miscellaneous Other Paradoxes

The one-sample runs test is designed to detect departures from randomness in a series of observations. It does so by comparing the observed lengths of runs (consecutive observations with the same value) to what would be expected under the null hypothesis of randomness. Any repeating pattern would certainly constitute a deviation from randomness, so one would expect it to be detected. Yet, an observed sequence of runs of length two, AABBAABBAABB... , would not lead to the runs test rejecting the null hypothesis of randomness [19]. See also [18].

A triad is a departure from the transitive property, in which $A > B > C$ implies that $A > C$. So, for example, in the game 'rock, paper, scissors', the triad occurs because rock beats scissors, which beats paper, which in turn beats the rock. One might think that transitivity [6] applies to pairs of random variables in the sense that $P\{A > B\} > .5$ and $P\{B > C\} > .5$ jointly imply that $P\{A > C\} > .5$. However, consider three die, fair in the sense that each face occurs with

probability $1/6$. Die A has one 1 and five 4s. Die B has one 6 and five 3s. Die C has three 2s and three 5s. Now $P\{A > B\} = 25/36 > .5$, $P\{B > C\} = 7/12 > .5$, and $P\{C > A\} = 7/12 > .5$ [26].

References

- [1] Baker, S.G. & Kramer, B.S. (2001). Good for women, Good for men, Bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies, *Journal of Women's Health and Gender-Based Medicine* **10**, 867–872.
- [2] Berger, V.W. (2004a). Valid adjustment of randomized comparisons for binary covariates, *Biometrical Journal* **46**, 589–594.
- [3] Berger, V.W. (2004b). On the generation and ownership of alpha in medical studies, *Controlled Clinical Trials* **25**, 613–619.
- [4] Berger, V.W. (2005). Nonparametric adjustment techniques for binary covariates, *Biometrical Journal* in press.
- [5] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004a). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**(1), 48–55.
- [6] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004b). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**(1), 48–55.
- [7] Cicchetti, D.V. & Feinstein, A.R. (1990). High agreement but low kappa. II. Resolving the paradoxes, *Journal of Clinical Epidemiology* **43**, 551–558.
- [8] Davis, G. (2002). Is the claim that 'variance kills' an ecological fallacy? *Accident Analysis and Prevention* **34**, 343–346.
- [9] Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa. I. The problems of two paradoxes, *Journal of Clinical Epidemiology* **43**, 543–549.
- [10] Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, John Wiley & Sons.
- [11] Hassler U., Thadewald, T. Nonsensical and biased correlation due to pooling heterogeneous samples, *The Statistician* **52**(3), 367–379.
- [12] Heydtmann, M. (2002). The nature of truth: Simpson's paradox and the limits of statistical data, *The Quarterly Journal Medicine* **95**, 247–249.
- [13] Kaiser, L. (1989). Adjusting for baseline: change or percentage change? *Statistics in Medicine* **8**, 1183–1190.
- [14] Lantz, C.A. & Nebenzahl, E. (1996). Behavior and Interpretation of the κ Statistic: Resolution of the Two Paradoxes, *Journal of Clinical Epidemiology* **49**(4), 431–434.
- [15] Lord, F.M. (1967). A paradox in the interpretation of group comparisons, *Psychological Bulletin* **69**(5), 304–305.
- [16] Lord, F.M. (1969). Statistical adjustments when comparing preexisting groups, *Psychological Bulletin* **72**, 336–337.
- [17] Malinas, G. & Bigelow, J. Simpson's paradox, in *The Stanford Encyclopedia of Philosophy* (Spring 2004 Edition), Edward N. Zalta, ed., forthcoming URL = <http://plato.stanford.edu/archives/spr2004/entries/paradox-simpson/>.
- [18] McKenzie, D.P., Onghena, P., Hogenraad, R., Martindale, C. & MacKinnon, A.J. (1999). Detecting patterns by one-sample runs test: paradox, explanation, and a new omnibus procedure, *The Journal of Experimental Education* **67**(2), 167–179.
- [19] Mogull, R.G. (1994). The one-sample runs test – A category of exception Mogull RG, *Journal of Educational and Behavioral Statistics* **19**(3), 296–303.
- [20] Morrell, C.H. (1999). Simpson's paradox: an example from a longitudinal study in South Africa, *Journal of Statistics Education* **7**(3), 1–4.
- [21] Neuhauser, M. (2004). The choice of alpha for one-sided tests, *Drug Information Journal* **38**, 57–60.
- [22] Neutel, C.I. (1997). The potential for Simpson's paradox in drug utilization studies, *Annals of Epidemiology* **7**(7), 517–521.
- [23] Piantadosi, S., Byar, D. & Green, S. (1988). The ecological fallacy, *American Journal of Epidemiology* **127**(5), 893–904.
- [24] Reintjes R. de Boer A., van Pelt W. & Mintjes-de Goot J. (2000). Simpson's paradox: an example from hospital epidemiology, *Epidemiology* **11**(1), 81–83.
- [25] Senn, S. (1997). *Statistical Issues in Drug Development*, John Wiley & Sons, England.
- [26] Szekely, G.J. (2003). Solution to problem #28: irregularly spotted dice, *Chance* **16**(3), 54–55.
- [27] Szklo, M. & Nieto, F.J. (2000). *Epidemiology: Beyond the Basics*, Aspen Publishers.
- [28] Von Eye, A. (2002). *Configural Frequency Analysis*, Lawrence Erlbaum, Mahwah.
- [29] Wainer, H. (1991). Adjusting for differential base rates: lord's paradox again, *Psychological Bulletin* **109**(1), 147–151.
- [30] Westfall P.H., Wolfinger R.D., Closed Multiple Testing Procedures and PROC MULTTEST, <http://support.sas.com/documentation/periodicals/obs/obswww23>, accessed 2/2/04.

VANCE W. BERGER AND VALERIE DURKALSKI

Parsimony/Occham's Razor

STANLEY A. MULAİK

Volume 3, pp. 1517–1518

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Parsimony/Occham's Razor

Parsimony or simplicity as a desirable feature of theories and models was first popularized by the Franciscan William Ockham (1285–1347) in the principle ‘Entities are not to be multiplied unnecessarily’. In other words, theoretical explanations should be as simple as possible, evoking the fewest explanatory entities as possible. The nineteenth-century physicist Heinrich Hertz listed parsimony as one of several desirable features for physical theories [4]. However, a rationale for this principle was not given. Later, the philosopher of science Karl Popper suggested it facilitated the falsifiability of theories, but did not provide a clear demonstration [7]. L. L. Thurstone [8], the American factor analyst, advocated parsimony and simplicity in factor analytic theories.

Mulaik [6] showed that in models in which estimation of parameters is necessary to provide values for unknown parameters, the number of dimensions in which data is free to differ from the model of the data is given by the degrees of freedom of the model, $df = D - m$, where D is the number of data elements to fit and m is the number of estimated parameters. (In **confirmatory factor analysis** and **structural equation models**, D is the number of nonredundant elements of a variance–covariance matrix to which the model is fit.) Each dimension corresponding to a degree of freedom corresponds to a condition by which the model can fail to fit and thus be disconfirmed. Parsimony is the fewness of parameters estimated relative to the number of data elements, and is expressed as a ratio $P = m/D$. Zero is the ideal case where nothing is estimated. James, Mulaik, and Brett [3] suggested that an aspect of model quality could be indicated by a ‘parsimony ratio’: $PR = df/D = 1 - P$. As this ratio approaches unity, the model is increasingly disconfirmable. They further suggested that

this ratio could be multiplied by ‘goodness-of-fit’ indices that range between 0 and 1 to combine fit with parsimony in assessing model quality, Q . For example, in structural equation modeling one can obtain $Q_{CFI} = PR * CFI$ or $Q_{GFI} = PR * GFI$, where CFI is the comparative fit index of Bentler [1] and GFI the ‘goodness-of-fit’ index of LISREL [5]. (see **Goodness of Fit; Structural Equation Modeling: Software**) It can also be adapted to the RMSEA index [2] as $Q_{RMSEA} = PR * \exp(-RMSEA)$. D in PR for the CFI is $p(p - 1)/2$ and for the GFI and $RMSEA$ it is $p(p + 1)/2$, where p is the number of observed variables. A $Q > 0.87$ indicates a model that is both highly parsimonious and high in degree of fit.

References

- [1] Bentler, P.M. (1990). Comparative fit indexes in structural models, *Psychological Bulletin* **107**, 238–246.
- [2] Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit, in *Testing Structural Equation Models*, K.H. Bollen & J.S. Long, eds, Sage Publications, Newbury Park, 136–162.
- [3] James, L.R., Mulaik, S.A. & Brett, J.M. (1982). *Causal Analysis: Assumptions, Models and Data*, Sage Publications, Beverly Hills.
- [4] Janik, A. & Toulmin, S. (1973). *Wittgenstein's Vienna*, Simon and Schuster, New York.
- [5] Jöreskog, K.G. & Sörbom, D. (1987). *LISREL V: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, National Educational Resources, Chicago.
- [6] Mulaik, S.A. (2001). The curve-fitting problem: An objectivist view, *Philosophy of Science* **68**, 218–241.
- [7] Popper, K.R. (1961). *The Logic of Scientific Discovery*, (translated and revised by the author), Science Editions, (Original work published 1934), New York.
- [8] Thurstone, L.L. (1947). *Multiple Factor Analysis*, University of Chicago Press, Chicago.

STANLEY A. MULAİK

Partial Correlation Coefficients

BRUCE L. BROWN AND SUZANNE B. HENDRIX

Volume 3, pp. 1518–1523

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Partial Correlation Coefficients

Partial correlation answers the question ‘what is the correlation coefficient between any two variables with the effects of a third variable held constant?’ An example will illustrate. Suppose that one wishes to know how well high-school grades correlate with grades in college, but with *potential ability* (scholastic aptitude) held constant; that is, for a group of students of equal ability, is there a strong tendency for students who do well in high school to also do well in college?

A direct way to do this would be to gather data on a group of students who are all equal in scholastic aptitude, and then examine the correlation coefficient between high school grades and college grades within this group. However, it would obviously not be easy to find a group of students who are equal in scholastic aptitude. Partial correlation is the statistical way of solving this problem by estimating from ordinary data what this correlation coefficient would be.

Suppose, in this example, that the correlation coefficient between high school grades (variable X_1) and college grades (variable X_2) is found to be very high ($r_{12} = 0.92$). However, both of these variables are also found to be quite highly correlated with scholastic aptitude (variable X_3), with the correlation between aptitude and high school grades being $r_{13} = 0.73$, and the correlation between aptitude and college grades being $r_{23} = 0.76$. These three correlation coefficients are referred to as *zero-order correlations*. A *zero-order correlation* is a correlation coefficient that does not include a control variable (*see Correlation; Correlation and Covariance Matrices*). Partial correlation that includes one control variable (variable X_3 above) is referred to as a *first-order correlation*.

The partial correlation coefficient (a first-order correlation) is a straightforward function of these three zero-order correlations:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}} \quad (1)$$

The symbol $r_{12.3}$ is read as ‘the partial correlation coefficient between variable X_1 and variable X_2 with variable X_3 held constant’. For the example just

described, this calculates as:

$$\begin{aligned} r_{12.3} &= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \\ &= \frac{0.92 - (0.73) \times (0.76)}{\sqrt{1 - (0.73)^2}\sqrt{1 - (0.76)^2}} \\ &= \frac{0.92 - 0.5548}{\sqrt{1 - 0.5329}\sqrt{1 - 0.5776}} \\ &= \frac{0.3652}{\sqrt{0.4671}\sqrt{0.4224}} = \frac{0.3652}{(0.6834)(0.6499)} \\ &= \frac{0.3652}{0.4419} = 0.82 \end{aligned} \quad (2)$$

From this calculation, it can be concluded that the correlation between high school grades and college grades drops from 0.92 to 0.82 when it is adjusted for the effects of scholastic aptitude. In other words, 0.82 is what the correlation between high school grades and college grades would be expected to be for a group of students who are all equal in scholastic aptitude. This partial correlation coefficient removes the effects of scholastic aptitude from the correlation between high school grades and college grades. It therefore represents the determinative influence of those factors not related to scholastic aptitude – perhaps such things as diligence, responsibility, academic motivation, and so forth – that are common to both high school grades and college grades.

Partial Correlation and Regression

It can be demonstrated that partial correlation is closely related to linear regression analysis (*see Multiple Linear Regression*). In fact, the partial correlation coefficient $r_{12.3}$ of 0.82 that was just demonstrated is equivalent to the correlation between the two sets of residuals that are produced by regressing X_1 onto X_3 , and X_2 onto X_3 .

This will be illustrated with ‘simplest case’ data; that is, a data set with few observations and simple numerical structure will be used to demonstrate that the partial correlation coefficient (as calculated from the above formula) is equivalent to the correlation coefficient between the two sets of residuals. For this demonstration, only six students for whom there are high school grades, college grades, and scholastic aptitude scores will be used. They are shown in Table 1.

2 Partial Correlation Coefficients

Table 1 High school GPA, college GPA, and SAT scores for six students

	High school GPA X_1	College GPA X_2	SAT scores X_3
Student A	3.64	3.65	661
Student B	3.43	3.51	523
Student C	3.20	3.38	477
Student D	2.80	2.96	569
Student E	2.17	3.10	523
Student F	1.56	1.99	385

In actual practice, sample sizes would usually be considerably larger. The small sample size here is to make the demonstration clear and understandable. To further simplify the demonstration, these three variables are first converted to standard score (Z score) form. The means and standard deviations of these three variables are

$$\bar{X}_1 = \frac{\sum X}{n} = \frac{3.64 + 3.43 + 3.20 + 2.80 + 2.17 + 1.56}{6} = \frac{16.80}{6} = 2.80$$

$$\begin{aligned}\sigma_1 &= \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n-1}} \\ &= \sqrt{\frac{50.237 - 16.80^2/6}{6-1}} \\ &= \sqrt{\frac{50.237 - 47.040}{5}} = \sqrt{\frac{3.197}{5}} \\ &= \sqrt{0.6394} = 0.800\end{aligned}$$

$$\bar{X}_2 = \frac{\sum X}{n} = \frac{3.65 + 3.51 + 3.38 + 2.96 + 3.10 + 1.99}{6} = \frac{18.59}{6} = 3.10$$

$$\begin{aligned}\sigma_2 &= \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n-1}} \\ &= \sqrt{\frac{59.399 - 18.59^2/6}{6-1}} \\ &= \sqrt{\frac{59.399 - 57.598}{5}} = \sqrt{\frac{1.801}{5}} \\ &= \sqrt{0.3601} = 0.600\end{aligned}$$

$$\begin{aligned}\bar{X}_3 &= \frac{\sum X}{n} = \frac{661 + 523 + 477 + 569 + 523 + 385}{6} \\ &= \frac{3138}{6} = 523 \\ \sigma_3 &= \sqrt{\frac{\sum X^2 - (\sum X)^2/n}{n-1}} \\ &= \sqrt{\frac{1683494 - 3138^2/6}{6-1}} \\ &= \sqrt{\frac{1683494 - 1641174}{5}} = \sqrt{\frac{42320}{5}} \\ &= \sqrt{8464} = 92\end{aligned}\quad (3)$$

For each of the three variables, the Z scores are created by subtracting the mean of each from the raw values to get deviation scores, and then dividing each deviation score by the standard deviation. The Z scores for these three variables are shown in Table 2.

Now that the data are in Z score form, it is relatively easy to obtain predicted scores and residual scores on high school grade point average (Z_1) and college GPA (Z_2), as predicted by regression from SAT scores (Z_3). The correlation coefficient between high school GPA and college GPA is found from the product of Z scores to be

$$\begin{aligned}r_{12} &= \frac{\sum Z_1 Z_2}{n-1} \\ &= \frac{1.0505 \times 0.9193 + 0.7879 \times 0.6860 + 0.5002 \times 0.4694 + 0.0000 \times (-0.2305) + (-0.7879) \times 0.0028 + (-1.5507) \times (-1.8469)}{5} \\ &= \frac{4.6027}{5} = 0.92\end{aligned}\quad (4)$$

Table 2 Standardized high school GPA, college GPA, and SAT scores for six students

	High school GPA Z_1	College GPA Z_2	SAT scores Z_3
Student A	1.0505	0.9193	1.50
Student B	0.7879	0.6860	0.00
Student C	0.5002	0.4694	-0.50
Student D	0.0000	-0.2305	0.50
Student E	-0.7879	0.0028	0.00
Student F	-1.5507	-1.8469	-1.50

The correlation coefficient between high school GPA and SAT is

$$r_{13} = \frac{\sum Z_1 Z_3}{n-1} = \frac{1.0505 \times 1.50 + 0.7879 \times 0.00 + 0.5002 \times (-0.50) + 0.0000 \times 0.50 + (-0.7879)0.00 + (-1.5507) \times (-1.50)}{5} = \frac{3.6517}{5} = 0.73 \quad (5)$$

The correlation coefficient between college GPA and SAT is

$$r_{23} = \frac{\sum Z_2 Z_3}{n-1} = \frac{0.9193 \times 1.50 + 0.6860 \times 0.00 + 0.4694 \times (-0.50) + (-0.2305) \times 0.50 + 0.0028 \times 0.00 + (-1.8469) \times (-1.50)}{5} = \frac{3.7993}{5} = 0.76 \quad (6)$$

These three zero-order correlation coefficients are the same ones from which the partial correlation between high school GPA and college GPA with scholastic aptitude test (SAT) scores held constant was found from the partial correlation formula (1) to be 0.82. It can be now shown that this same partial correlation coefficient can be obtained by finding the correlation coefficient between the residual scores of high school GPA predicted from SAT scores, and the residual scores of college GPA predicted from SAT scores. The high school GPA scores predicted from SAT scores are found from the *standardized regression function*:

$$\hat{Z}_{1.3} = r_{13} Z_3 \quad (7)$$

In other words, the standardized predicted high school GPA scores that are predicted on the basis of SAT scores ($\hat{Z}_{1.3}$) are found by multiplying each of the standardized SAT scores (Z_3) by the correlation coefficient between these two variables (r_{13}) (see **Standardized Regression Coefficients**).

$$\hat{Z}_{1.3} = r_{13} Z_3 = (0.730) \times \begin{bmatrix} 1.50 \\ 0.00 \\ -0.50 \\ 0.50 \\ 0.00 \\ -1.50 \end{bmatrix} = \begin{bmatrix} 1.0955 \\ 0.0000 \\ -0.3652 \\ 0.3652 \\ 0.0000 \\ -1.0955 \end{bmatrix} \quad (8)$$

The *residual scores* for high school GPA are obtained by subtracting these predicted scores from the actual high school GPA standardized scores:

$$e_{1.3} = Z_1 - \hat{Z}_{1.3} = \begin{bmatrix} 1.0505 \\ 0.7879 \\ 0.5002 \\ 0.0000 \\ -0.7879 \\ -1.5507 \end{bmatrix} - \begin{bmatrix} 1.0955 \\ 0.0000 \\ -0.3652 \\ 0.3652 \\ 0.0000 \\ -1.0955 \end{bmatrix} = \begin{bmatrix} -0.0450 \\ 0.7879 \\ 0.8654 \\ -0.3652 \\ -0.7879 \\ -0.4552 \end{bmatrix} \quad (9)$$

Similarly, the college GPA predicted scores (from SAT) are found to be

$$\hat{Z}_{2.3} = r_{23} Z_3 = (0.760) \times \begin{bmatrix} 1.50 \\ 0.00 \\ -0.50 \\ 0.50 \\ 0.00 \\ -1.50 \end{bmatrix} = \begin{bmatrix} 1.1398 \\ 0.0000 \\ -0.3799 \\ 0.3799 \\ 0.0000 \\ -1.1398 \end{bmatrix} \quad (10)$$

and the residual scores for college GPA are

$$e_{2.3} = Z_2 - \hat{Z}_{2.3} = \begin{bmatrix} 0.9193 \\ 0.6860 \\ 0.4694 \\ -0.2305 \\ 0.0028 \\ -1.8469 \end{bmatrix} - \begin{bmatrix} 1.1398 \\ 0.0000 \\ -0.3799 \\ 0.3799 \\ 0.0000 \\ -1.1398 \end{bmatrix} = \begin{bmatrix} -0.2205 \\ 0.6860 \\ 0.8493 \\ -0.6104 \\ 0.0028 \\ -0.7071 \end{bmatrix} \quad (11)$$

4 Partial Correlation Coefficients

The partial correlation coefficient of 0.82 between high school GPA and college GPA with SAT scores held constant can now be shown to be equivalent to the zero-order correlation coefficient between these two sets of residuals. The covariance between these two sets of residuals is

$$\begin{aligned} \text{cov}(e_{1.3}e_{2.3}) &= \frac{\sum e_{1.3}e_{2.3}}{n-1} \\ &= \frac{(-0.0450)(-0.2205) + 0.7879 \times 0.6860 \\ &\quad + 0.8654 \times 0.8493 + (-0.3652)(-0.6104) \\ &\quad + (-0.7879)0.0028 + (-0.4552)(-0.7071)}{5} \\ &= \frac{1.8280}{5} = 0.3656 \end{aligned} \quad (12)$$

Since the **Pearson product moment correlation coefficient** is equal to the covariance divided by the two standard deviations, the correlation between these two sets of residuals can be obtained by dividing this covariance by the two standard deviations. The residual scores are already in deviation score form, so the standard deviation for the first set of residuals is

$$\begin{aligned} \sigma_{e1} &= \sqrt{\frac{\sum e_{1.3}^2}{n-1}} = \sqrt{\frac{2.3330}{6-1}} \\ &= \sqrt{\frac{2.3330}{5}} = \sqrt{0.4666} = 0.6831 \end{aligned} \quad (13)$$

and the standard deviation for the second set of residuals is

$$\begin{aligned} \sigma_{e2} &= \sqrt{\frac{\sum e_{2.3}^2}{n-1}} = \sqrt{\frac{2.1131}{6-1}} \\ &= \sqrt{\frac{2.1131}{5}} = \sqrt{0.4226} = 0.6501 \end{aligned} \quad (14)$$

The zero-order correlation between the two sets of residuals is therefore found as

$$\begin{aligned} r_{12.3} &= \frac{\text{cov}(e_{1.3}e_{2.3})}{\sigma_{e1}\sigma_{e2}} \\ &= \frac{0.3656}{(0.6831)(0.6501)} = \frac{0.3656}{0.4441} = 0.82 \end{aligned} \quad (15)$$

which is indeed equivalent to the partial correlation coefficient as calculated by the received formula. It can be demonstrated algebraically that this equivalence holds in general.

The calculations involved in obtaining partial correlations (as shown above) are simple enough to be easily accomplished using a spreadsheet, such as Microsoft Excel, Quattro Pro, ClarisWorks, and so on. They can also be accomplished using computer statistical packages such as SPSS and SAS (see **Software for Statistical Analyses**). Landau and Everitt [7] clearly outline how to calculate partial correlations using SPSS (pp. 94–96).

Partial Correlation in Relation to Other Methods

Partial correlation is related to, and foundational to many statistical methods of current interest, such as **structural equation modeling** (SEM) and causal modeling. Using partial correlation as a method to determine the nature of causal relations between observed variables has a long history, as suggested by Lazarsfeld [8], and presented in more detail in by Simon [9] and Blalock [2, 3]. Partial correlation is also closely related, both conceptually and also mathematically, to **path analysis** (see [5]). Path analysis methods were first developed in the 1930s by Sewall Wright [11–14] (see also [10]), and they also are fundamental to the developments underlying structural equations modeling.

In addition, partial correlation mathematics is central to the computational procedures in traditional factor analytic methods [4] (and see **Factor Analysis: Exploratory**). They are also central to the algebraic rationale underlying stepwise multiple regression. That is, partial correlation mathematics is the conceptual basis for removing the effects of each predictive variable in a stepwise regression, in order to determine the additional predictive contribution of the remaining variables. One of the more interesting applications of partial correlation methods is in demonstrating **mediation** [1, 6]. That is, partial correlation methods are used to demonstrate that the causal effects of one variable on another are in fact mediated by yet another variable (a third one) – that without this third variable, the correlative relationship vanishes, or at least is substantially attenuated. Thus, partial correlation is useful not only as a method of statistical control but also as a way of testing theories of causation.

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] Blalock, H.M. (1961). *Causal Inferences in Non-Experimental Research*, University of North Carolina Press, Chapel Hill.
- [3] Blalock, H.M. (1963). Making causal inferences for unmeasured variables from correlations among indicators, *American Journal of Sociology* **69**, 53–62.
- [4] Brown, B.L., Hendrix, S.B. & Hendrix, K.A. (in preparation). *Multivariate for the Masses*, Prentice-Hall, Upper Saddle River.
- [5] Edwards, A.L. (1979). *Multiple Regression and the Analysis of Variance and Covariance*, W. H. Freeman, San Francisco.
- [6] Judd, C.M. & Kenny, D.A. (1981). Process analysis: estimating mediation in treatment evaluations, *Evaluation Review* **5**, 602–619.
- [7] Landau, S. & Everitt, B.S. (2003). *A Handbook of Statistical Analyses Using SPSS*, Chapman and Hall, Boca Raton.
- [8] Lazarsfeld, P.F. (1955). The interpretation of statistical relations as a research operation, in *The Language of Social Research*, P.F. Lazarsfeld & M. Rosenberg, eds, The Free Press, New York.
- [9] Simon, H.A. (1957). *Models of Man*, Wiley, New York.
- [10] Tukey, J.W. (1954). Causation, regression, and path analysis, in *Statistics and Mathematics in Biology*, O. Kempthorne, T.A. Bancroft, J.M. Gowen & J.L. Lush, eds, Iowa State College Press, Ames.
- [11] Wright, S. (1934). The method of path coefficients, *Annals of Mathematical Statistics* **5**, 161–215.
- [12] Wright, S. (1954). The interpretation of multivariate systems, in *Statistics and Mathematics in Biology*, O. Kempthorne, T.A. Bancroft, J.M. Gowen & J.L. Lush, eds, Iowa State College Press, Ames.
- [13] Wright, S. (1960a). Path coefficients and path regressions: alternative or complementary concepts? *Biometrics* **16**, 189–202.
- [14] Wright, S. (1960b). The treatment of reciprocal interaction, with or without lag, in path analysis, *Biometrics* **16**, 423–445.

BRUCE L. BROWN AND SUZANNE B. HENDRIX

Partial Least Squares

PAUL D. SAMPSON AND FRED L. BOOKSTEIN

Volume 3, pp. 1523–1529

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Partial Least Squares

Introduction

Partial Least Squares (PLS), as described in this article, is a method of statistical analysis for characterizing and making inferences about the relationship(s) among two or more causally related blocks of variables in terms of pairs of **latent variables**. The notion of causality in the preceding sentence is meant in its scientific, not its statistical, sense. PLS is not a tool for asserting or testing causation, but for *calibrating* it (see [10]) in contexts already known to be causal in Pearl's sense [14]. The informal idea of a 'block' of variables is used in the sense of a vector of measurements that together observe a range of relevant aspects of the causal nexus under study. This discussion of PLS is presented in the context of an application in the field of fetal alcohol teratology. It concerns the characterization of 'alcohol-related brain damage' as expressed in the relationship of a multivariate characterization of prenatal exposure to alcohol through maternal drinking to a large battery of neurobehavioral performance measures in a prospective study of children of approximately 500 women with a broad range of drinking patterns during pregnancy. That the effect of alcohol on the brain or on behavior is causal is amply known in advance from a plethora of studies in animals and also from human clinical teratology.

This PLS approach to defining and making inference about latent variables is quite distinct from other methods of **structural equation modeling (SEM)** and from multivariate methods such as **canonical correlation analysis**. In this context, the latent variable models that underlie two-block PLS analysis can be shown to parameterize a range of reduced-rank regression models [1] as shown by Wegelin et al. [19].

SEM typically draws inferences about latent variables – without enabling the estimation/computation of explicit scores for these constructs – under strong assumptions about the error structure in measurement models for the observable ('manifest') variables representing each latent variable. PLS computes latent variable scores without any measurement model assumptions. As will become clear from the symmetries of the equations following, PLS is an 'undirected' technique in the sense of the directed

acyclic graph (DAG) literature (see **Graphical Chain Models**), without specific tools to assess the direction of causation between blocks its tools show to be associated. Nevertheless, PLS is completely consistent with the critique of SEM that can be mounted from that more coherent point of view [17], in that its reports are consistent with the full range of causal reinterpretations that might model one single data stream. Wegelin et al. [19] show that all feasible directed path diagrams for two-block latent variable models are equivalent or indistinguishable in the sense that they parameterize the same set of covariance matrices. Once causation is presumed, the tie between paired latent variables in PLS is in most ways equivalent to a multivariate version of the $E(Y|do(X))$ operation of Pearl [14] in the situation where X is a vector of alternate indicators of the same latent score (so that one can't 'do' $X_1, X_2, \dots$ all at the same time).

Canonical correlation analysis (and generalizations to multiple blocks), as the name suggests, computes scores to optimize between-block correlation of the composites (linear combinations of the manifest variables), but these bear no interpretation as latent variables (see **Canonical Correlation Analysis**).

PLS must also be distinguished from a number of other related statistical analysis methods reported in the literature under the same name. The twentieth-century Swedish econometrician Herman Wold developed PLS as an ordinary least squares–based approach for modeling paths of causal relationship among any number of 'blocks' of manifest variables [9, 20] (see **Path Analysis and Path Diagrams**). There appear to have been relatively few recent applications of this type of PLS in the behavioral sciences; see [13] and [7].

The largest body of literature on Partial Least Squares is found in the field of chemometrics, with early work associated with Herman Wold's son, Svante. A perspective on early PLS development was provided by Svante Wold in 2001 in a special issue of a chemometrics journal [21]. The chemometric variants of PLS include the situation where one 'block' consists of a single dependent variable, and PLS provides an approach to coping with a multiple regression problem with highly multicollinear predictor blocks. In this context it is often compared with methods such as ridge regression and principal components regression (see **Principal Component Analysis**).

Increasingly common in applications of PLS in chemometrics, bioinformatics, and the behavioral sciences is the two-block case, using the algebra described below, but where one of the blocks of variables is a design matrix defining different classes to be discriminated. Recent behavioral science applications include the discrimination of dementias [5], the differentiation of fronto-temporal dementia from Alzheimer's using FDG-PET imaging [8], and the characterization of schizophrenia in terms of the relationship of MRI brain images and neuropsychological measures [11]. PLS has also been applied to examine differences between neurobehavioral experimental design conditions in a study using multivariate event related brain potential data [6]. For recent discussion of applications to microarray data, see [12] (see **Microarrays**). Barker and Rayens [2] argue that PLS is to be preferred over principal component analysis (PCA) when discrimination is the goal and dimension reduction is needed.

Algebra and Interpretation of two-block PLS

The scientist for whom PLS is designed is faced with data for $i = 1, \dots, n$ subjects on two (or more) lists or blocks of variables, $X_j, j = 1, \dots, J$ and $Y_k, k = 1, \dots, K$. In the application below, $J = 13$ X 's will represent different measures of the alcohol consumption of pregnant women, and $K = 9$ Y 's will be a range of measurements of young adult behavior, limited here for purposes of illustration to nine subscales of the WAIS IQ test. Write $\mathbf{C}$ for the $J \times K$ covariance matrix of the X 's by the Y 's. In many applications, the variables in a block are incommensurate, and so are scaled to variance 1. If both blocks of variables are standardized to variance 1, this will be a matrix of correlations. Analyses of standardized and unstandardized blocks are in general different; analyses of mean-centered versus uncentered blocks are in most respects exactly the same.

The central computation can be phrased in any of several different ways. All involve the production of a vector of coefficients $\mathbf{A}_1 = (A_{11}, \dots, A_{1J})$, one for each X_j , together with a vector of coefficients $\mathbf{B}_1 = (B_{11}, \dots, B_{1K})$, one for each Y_k . The elements of either vector, $\mathbf{A}_1$ or $\mathbf{B}_1$, are scaled so as to have squares that sum to 1. These make up the

first pair of singular vectors of the matrix $\mathbf{C}$ from computational linear algebra. Accompanying this pair of vectors is a scalar quantity, d_1 , the first singular value of $\mathbf{C}$, such that all of the following assertions are true (in fact, all four are exactly algebraically equivalent):

- The scaled outer product $d_1 \mathbf{A}_1 \mathbf{B}_1'$, is the best (least-squares) fit of all such matrices to the between-block covariance matrix $\mathbf{C}$. The goodness of this fit serves as a figure of merit for the overall PLS analysis. It is characterized as 'the fraction of summed squared covariance explained' in the model $\mathbf{C} = d_1 \mathbf{A}_1 \mathbf{B}_1' + \text{error}$.
- The vector $\mathbf{A}_1$ is the 'central tendency' of the K columns of $\mathbf{C}$ thought of as patterns of covariance across the J variables X_j and can be computed as the first uncentered, unstandardized principal component of the columns of $\mathbf{C}$. Likewise, the vector $\mathbf{B}_1$ is the first principal component of the rows of $\mathbf{C}$, the central tendency they share as J patterns of covariance across the K variables Y_k .
- The latent variables $LV_{1X} = \sum_{j=1}^J A_{1j} X_j$ and $LV_{1Y} = \sum_{k=1}^K B_{1k} Y_k$ have covariance d_1 , and this is the greatest covariance of any pair of such linear combinations for which the coefficient vectors are standardized to norm 1, $\mathbf{A}_1' \mathbf{A}_1 = \mathbf{B}_1' \mathbf{B}_1 = 1$. These latent variables are not like those of an SEM (e.g., LISREL or EQS see **Structural Equation Modeling: Software**); they are ordinary linear combinations of the observed data, not factors, and they have exact values. The quantity PLS is optimizing is the covariance between these paired LV's, *not* the variance explained in the course of predictions of the outcomes individually or collectively.
- The elements A_{1j} of the vector $\mathbf{A}_1$ are proportional to the covariances of the corresponding X -block variable X_j with the latent variable LV_{1Y} representing the Y 's, the other block in the analysis; and, similarly, the elements B_{1k} of the vector $\mathbf{B}_1$ are proportional to the covariances of the corresponding Y -block variable Y_k with the latent variable LV_{1X} representing the X 's. When it is known *a priori* that a construct that the X 's share causes changes in a construct that the Y 's share, or vice versa, these coefficients may be called *salience*s: each A_{1j} is the salience of the variable X_j for the latent variable representing the Y -block, and each B_{1k} is the salience of the

variable Y_k for the latent variable representing the X -block.

A PLS analysis thus combines the elements of two (or more) blocks of variables into composite scores, usually one per block, that sort variables (by saliences A_{1j} or B_{1k}) and also subjects (by scores $LV_{1X} = \sum_{j=1}^J A_{1j}X_j$ and $LV_{1Y} = \sum_{k=1}^K B_{1k}Y_k$). The resulting systematic summary of the pattern of cross-block covariance thus links characterization of subjects with characterization of variables. The computations of PLS closely resemble other versions of cross-block averaging in applied statistics, such as **correspondence analysis**.

Subsequent pairs of salience or coefficient vectors, A_2 and B_2 , A_3 and B_3 , and so on, and corresponding latent variables, explaining residual covariance structure after that explained by the first pair of singular vectors, can be computed as subsequent pairs of singular vectors from the complete singular-value decomposition (SVD) of the cross-block covariance matrix, $C = ADB'$, where $A = [A_1, A_2, \dots]$ is the $J \times M$ matrix of orthonormal left singular vectors, $B = [B_1, B_2, \dots]$ is the $K \times M$ matrix of orthonormal right singular vectors, and D is the $M \times M$ diagonal matrix of ordered singular values, $M = \min\{J, K\}$.

There are no distributional assumptions in PLS; hence inference in PLS goes best by applications of bootstrap analysis and permutation testing [4] (see **Permutation Based Inference**). The primary hypotheses of interest concern the *structure* of the relationship between the X and Y blocks – the patterning of the rows and columns of C – as these calibrate the assumed causal relationship. The significance of this structure is assessed in terms of the leading singular values (covariances of the latent variable pairs) with respect to a null model of zero covariance between the X and Y blocks. More formally, assuming random sampling and a true population cross-covariance matrix Σ with singular values $\delta_1 \geq \delta_2 \geq \dots \geq \delta_M$, we ask whether the structure of the covariance matrix can be modeled in terms of a small number, p , of singular vector pairs (or latent variables); that is: $\delta_1 \geq \delta_2 \geq \dots \geq \delta_p > \delta_{p+1} = \dots = \delta_M = 0$. Without altering the covariance structure within the several blocks separately, one permutes case orders of all the blocks independently and examines sample singular values, or sums of the first few singular values, with respect to the permutation distribution.

When the PLS model for the SVD structure has been found to be statistically significant, one examines the details of the structure, the saliences or singular-value coefficients of A_1 and B_1 , by computing bootstrap standard errors. These inferential computations effectively take into account the excess of number of variables over sample size in this data set. For instance, all the P values in the example to follow pertain to analysis of all 13 prenatal alcohol measures, all 9 IQ subscale outcomes, or their combination, at once; no further adjustment for multiple comparisons is necessary.

PLS Applications in Behavioral Teratology

Teratologists study the developmental consequences of chemical exposures and other measurable aspects of the intrauterine and developmental environment. In human studies, the magnitude of teratological exposures cannot be experimentally controlled. Analysis by PLS as presented here rewards the investigator who conscientiously observes a multiplicity of alternative exposure measures along with a great breadth of behavioral outcomes. This strategy is appropriate whenever there is a plausible single factor ('latent brain damage') underlying a wide range of child outcomes, and a corresponding ranking of children by extent of deficit. Streissguth et al. [18] and Bookstein et al. [3] discuss the PLS methodology in detail as applied to this type of 'dose-response' behavioral teratology study. The application here is a simple demonstration relating prenatal alcohol exposure to an assessment of IQ at 21 years of age. See [15] for a similar application using IQ data at 7 years of age. Longitudinal PLS analyses are discussed in [18].

Sample. During 1974–1975, 1529 women were interviewed in their fifth month of pregnancy regarding health habits. Most were married (88%) and had graduated from high school (again 88%). Approximately 500 of these generally low risk families, chosen by participation in prenatal care and by prenatal maternal report of alcohol use during pregnancy, have been followed as part of a longitudinal prospective study of child development [18]. Follow-up examinations of morphological, neurocognitive, motor, and other real life outcomes were conducted at 8 months, 18 months, 4 years, 7 years, 14 years and 21 years.

Thirteen measures of prenatal alcohol exposure were computed from the prenatal maternal interviews.

4 Partial Least Squares

Table 1 Two-block PLS analysis: prenatal alcohol exposure by WAIS IQ

Prenatal alcohol measure	Salience	Standard error	IQ subscale score	Salience	Standard error
AAP	0.23	0.06	INFOSS	0.34	0.08
AAD	0.10	0.07	DGSPSS	0.29	0.10
MOCCP	0.13	0.06	ARTHSS	0.52	0.07
MOCCD	0.04	0.07	SIMSS	0.09	0.12
DOCCP	0.42	0.04	PICCSS	0.24	0.11
DOCCD	0.33	0.04	PICASS	0.07	0.13
BINGEP	0.33	0.05	BLKDSS	0.46	0.07
BINGED	0.24	0.05	OBJASS	0.33	0.08
MAXP	0.35	0.03	DSYMSS	0.24	0.10
MAXD	0.26	0.04			
QFVP	0.36	0.04			
QFVD	0.27	0.04			
ORDEXC	0.19	0.05			

94% of summed squared covariance in the cross-covariance matrix explained by this first pair of LV scores. Correlation between Alcohol and IQ LV scores is -0.22 .

Salience is a bias-corrected bootstrap mean salience from 1000 bootstrap replications. Standard error is the bootstrap standard error. Prenatal alcohol measures: AA: ave. ounces of alcohol per day; MOCC: ave monthly drinking occasions; DOCC: ave drinks per drinking occasion; BINGE: indicator of at least 5 drinks on at least one occasion, MAX: maximum drinks on a drinking occasion; QFV: 5-point ordered quantity-frequency-variability scale; ORDEXC: 5-point *a priori* ranking of consumption for study design. 'P' and 'D' in alcohol measure names refer to the periods prior to recognition of pregnancy and during pregnancy, respectively.

These measures reflect the level and pattern of drinking, including binge drinking, during two time periods: the month or so prior to pregnancy recognition and mid-pregnancy (denoted by 'P' and 'D', respectively, in the variable names of Table 1). Each of these scores was gently (monotonically) transformed to approximately linearize their relationships with a net neurobehavioral outcome [16].

For purposes of illustration the analysis here considers just 9 subscales of the WAIS-R IQ test assessed at 21 years of age. A recurring theme in PLS analysis is that it is usually inadvisable to base analyses only on common summary scales, such as full-scale, 'performance' or 'verbal' IQ, regardless of the established reliability of these scales. The dimensions (or 'scales') of the multivariate outcome most relevant for characterizing the cross-block associations are not necessarily well-represented by the conventional summary scores. In this case the nine subscales are represented without further summarization as a *block* of the nine original raw WAIS scores.

Partial Least Squares for Exposure-behavior Covariances

Because of incommensurabilities among the exposure measures together with differences in variance among

the IQ measures, both the 13-variable prenatal alcohol exposure block $X_j, j = 1, \dots, 13$ and the 9-variable psychological outcome block $Y_k, k = 1, \dots, 9$ were scaled to unit variance. In other words, this PLS derives from the singular-value decomposition of a 13×9 correlation matrix and results in a series of pairs of linear combinations, of which only the first will merit consideration here: (LV_{1X}, LV_{1Y}) . The A_1 's, the coefficients of LV_{1X} , jointly specify the pattern of maternal drinking behavior during pregnancy with greatest relevance for IQ functioning as a whole, in the sense of maximizing cross-block covariance. At the same time, the profile given by the B_1 's, the coefficients of LV_{1Y} is the IQ profile of greatest relevance for prenatal alcohol exposure.

Table 1 summarizes the PLS analysis. The first singular value, 1.107, represents 94% of the total (summed) squared covariance in the cross-covariance matrix. The structure of this first dimension of the SVD is highly significant as the value 1.107 was not exceeded by the first singular value in any of 1000 random permutations of the rows of the alcohol block with respect to the IQ block. Almost all the prenatal alcohol measures have notable salience, but the two drinking frequency scores (AA: average ounces

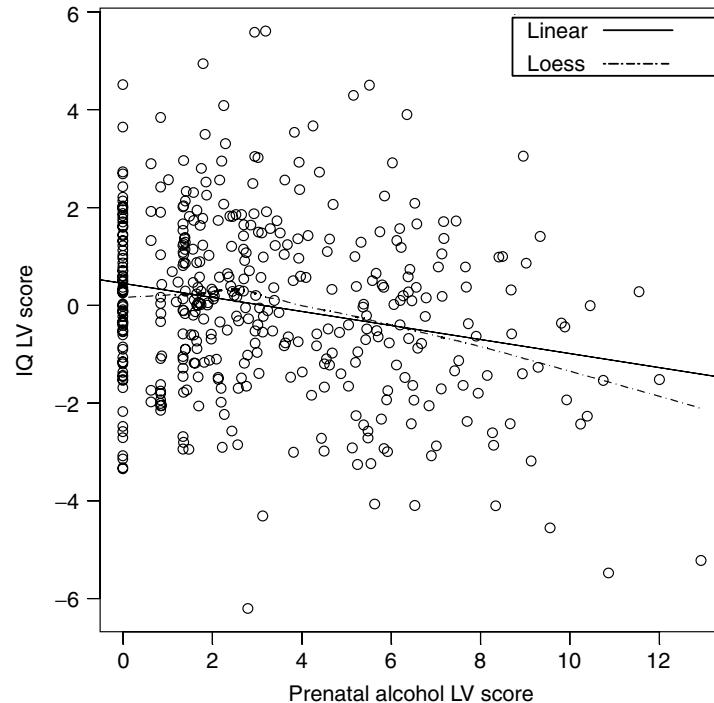


Figure 1 Scatterplot for the relationship of the IQ LV score and the Alcohol LV score, with linear and smooth regressions

of alcohol per day; MOCC: average monthly drinking occasions) have relatively low salience for IQ in comparison with the other binge-related measures, notably DOCC: average drinks per drinking occasion. The measures of drinking behavior prior to pregnancy recognition (P) are also somewhat more salient than the measures reflecting mid-pregnancy (D). The arithmetic (ARTH) and block design (BLKD) scales of the WAIS-R are clearly the most salient reflections of prenatal alcohol exposure. While the LV scores computed from these saliences, (LV_{1Y}) are very highly correlated with full-scale IQ ($r = 0.978$), this IQ LV score is notably more highly correlated with the Alcohol LV score (LV_{1X}) than is full-scale IQ.

Figure 1 presents a scatterplot of the Alcohol and IQ LV scores that correlate -0.22 . In applications, when the relationship depicted in this figure is challenged by adjustment for other prenatal exposures, demographics, and postnatal environmental factors using conventional multiple regression analysis, alcohol retains a highly significant correlation with the IQ LV after adjustment, although it does

not explain as much variance in the IQ LV as is explained by demographic and socioeconomic factors.

Remarks

In other applications, the PLS technique can focus the investigator's attention on profiles of saliences (causal relevance) across aspects of measurement in multimethod studies (*see Multitrait–Multimethod Analyses*), can detect individual cases with extreme scores on one or more LV's, can reveal subgrouping patterns by graphical enhancement of LV scatterplots that were carried out on covariance structures pooled over those subgroupings, and the like. PLS is not useful for theory-testing, as its permutation tests do not meet even the mild strictures of the DAG literature, let alone the standards of causation arising in the neighboring domains from which the multidisciplinary data typical of PLS applications tend to emerge. It takes its place alongside **biplots** and correspondence analysis in any toolkit of exploratory data analysis wherever empirical theory is weak and point

predictions of case scores or coefficients correspondingly unlikely.

References

- [1] Anderson, T.W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions, *Annals of Statistics* **27**, 1141–1154.
- [2] Barker, M. & Rayens, W. (2003). Partial least squares discrimination, *Journal of Chemometrics* **17**, 166–173.
- [3] Bookstein, F.L., Sampson, P.D. & Streissguth, A.P. (1996). Exploiting redundant measurement of dose and developmental outcome: new methods from the behavioral teratology of alcohol, *Developmental Psychology* **32**, 404–415.
- [4] Good, P.I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer, New York.
- [5] Gottfries, J., Blennow, K., Wallin, A. & Gottfries, D.G. (1995). Diagnosis of dementias using partial least-squares discriminant-analysis, *Dementia* **6**, 83–88.
- [6] Hay, J.F., Kane, K.A., West, K. & Alain, D. (2002). Event-related neural activity associated with habit and recollection, *Neuropsychologia* **40**, 260–270.
- [7] Henningsson, M., Sundbom, E., Armelius, B.A. & Erdberg, P. (2001). PLS model building: a multivariate approach to personality test data, *Scandinavian Journal of Psychology* **42**, 399–409.
- [8] Higdon, R., Foster, N.L., Koepp, R.A., DeCarli, C.S., Jagust, W.J., Clark, C.M., Barbas, N.R., Arnold, S.E., Turner, R.S., Heidebrink, J.L. & Minoshima, S. (2004). A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging, *Statistics in Medicine* **23**, 315–326.
- [9] Jöreskog, K.G. & Wold, H., eds (1982). *Systems Under Indirect Observation: Causality, Structure, Prediction*, North Holland, New York.
- [10] Martens, H. & Naes, T. (1989). *Multivariate Calibration*, Wiley, New York.
- [11] Nestor, P.G., Barnard, J., O'Donnell, B.F., Shenton, M.E., Kikinis, R., Jolesz, F.A., Shen, Z., Bookstein, F.L. & McCarley, R.W. (1997). Partial least squares analysis of MRI and neuropsychological measures in schizophrenia, *Biological Psychiatry* **41**(7 Suppl 1), 16S.
- [12] Nguyen, D.V. & Rocke, D.M. (2004). On partial least squares dimension reduction for microarray-based classification, *Computational Statistics & Data Analysis* **46**, 407–425.
- [13] Pagès, J. & Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements, *Chemometrics and Intelligent Laboratory Systems* **58**, 261–273.
- [14] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, New York.
- [15] Sampson, P.D., Streissguth, A., Barr, H. & Bookstein, F. (1989). Neurobehavioral effects of prenatal alcohol. Part II. Partial least squares analyses, *Neurotoxicology and Teratology* **11**, 477–491.
- [16] Sampson, P.D., Streissguth, A.P., Bookstein, F.L. & Barr, H.M. (2000). On categorizations in analyses of alcohol teratogenesis, *Environmental Health Perspectives* **108**(Suppl. 2), 421–428.
- [17] Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction and Search*, 2nd Edition, MIT Press, Cambridge.
- [18] Streissguth, A.P., Bookstein, F.L., Sampson, P.D. & Barr, H.M. (1993). *The Enduring Effects of Prenatal Alcohol Exposure on Child Development: Birth Through Seven Years, a Partial Least Squares Solution*, The University of Michigan Press, Ann Arbor.
- [19] Wegelin, J.A., Packer, A., & Richardson, T.S. (2005). Latent models for cross-covariance, *Journal of Multivariate Analysis* in press.
- [20] Wold, H. (1985). Partial least squares, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds., Wiley, New York, pp. 581–591.
- [21] Wold, S. (2001). Personal memories of the early PLS development, *Chemometrics and Intelligent Laboratory Systems*, **58**, 83–84.

PAUL D. SAMPSON AND FRED L. BOOKSTEIN

Path Analysis and Path Diagrams

STEVEN M. BOKER AND JOHN J. MCARDLE

Volume 3, pp. 1529–1531

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Path Analysis and Path Diagrams

History

Path analysis was first proposed by Sewall Wright [12] as a method for separating sources of variance in skeletal dimensions of rabbits (see [10, 11], for historical accounts). He extended this method of *path coefficients* to be a general method for calculating the association between variables in order to separate sources of environmental and genetic variance [14]. Wright was also the first to recognize the principle on which path analysis is based when he wrote, ‘The correlation between two variables can be shown to equal the sum of the products of the chains of path coefficients along all of the paths by which they are connected’ [13, p. 115].

Path analysis was largely ignored in the behavioral and social sciences until Duncan [3] and later Goldberger [4] recognized and brought this work to the attention of the fields of econometrics and psychometrics. This led directly to the development of the first modern **structural equation modeling** (SEM) software tools ([5] (see **Structural Equation Modeling: Software**)).

Path diagrams were also first introduced by Wright [13], who used them in the same way as they are used today; albeit using drawings of guinea pigs rather than circles and squares to represent variables (see [6], for an SEM analysis of Wright’s data). Duncan argued that path diagrams should be ‘...isomorphic with the algebraic and statistical properties of the postulated system of variables...’ ([3], p. 3). Modern systems for path diagrams allow a one-to-one translation between a path diagram and computer scripts that can be used to fit the implied structural model to data [1, 2, 9].

Path Diagrams

Path diagrams express the regression equations relating a system of variables using basic elements such as squares, circles, and single- or double-headed arrows (see Figure 1). When these elements are correctly combined in a path diagram, the algebraic relationship between the variables is completely specified and the

predicted covariance matrix between the measured variables can be unambiguously calculated [8]. The (Reticular Action Model) RAM method is discussed below as a recommended practice for constructing path diagrams.

Figure 2 is a path diagram expressing a bivariate regression equation with one outcome variable, y , such that

$$y = b_0 + b_1x + b_2z + e \quad (1)$$

where e is a residual term with a mean of zero. There are four single-headed arrows, b_0 , b_1 , b_2 , and b_3 pointing into y and likewise four terms are added together on the right hand side of 1. In a general linear model, one would use a column of ones to allow the estimation of the intercept b_0 . Here a constant variable is denoted by a triangle that maps onto that column of ones. For simplicity of presentation, the two predictor variables x and z in this example have means of zero. If x and z had nonzero means, then single-headed arrows would be drawn from the triangle to x and from the triangle to z .

There are double-headed variance arrows for each of the variables x , z , and e on the right hand side of the equation, representing the variance of the predictor variables, and residual variance respectively. The constant has, by convention, a nonzero variance term

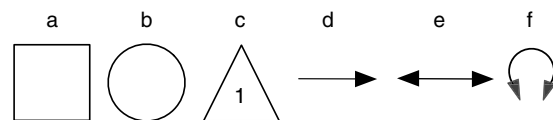


Figure 1 Graphical elements composing path diagrams. a. Manifest (measured) variable. b. Latent (unmeasured) variable. c. Constant with a value of 1. d. Regression coefficient. e. covariance between variables. f. Variance (total or residual) of a variable

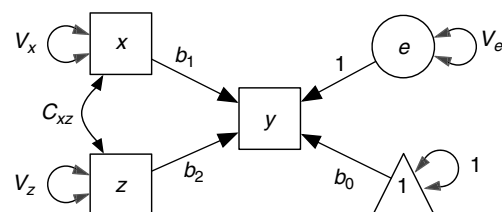


Figure 2 Bivariate regression expressed as a path diagram (see **Multivariate Multiple Regression**)

2 Path Analysis and Path Diagrams

fixed at the value 1.0. While this double-headed arrow may seem counterintuitive since it is not formally a variance term, it is required in order to provide consistency to the path tracing rules described below. In addition, there is a double-headed arrow between x and z that specifies the potential covariance, C_{xz} between the predictor variables.

Multivariate models expressed as a system of linear equations can also be represented as path diagrams. As an example, a factor model (*see* **Factor Analysis: Confirmatory**) with two correlated factors is shown in Figure 3. Each variable with one or more single-headed arrows pointing to it defines one equation in the system of simultaneous linear equations. For instance, one of the six simultaneous equations implied by Figure 3 is

$$y = b_1 F_1 + u_y, \quad (2)$$

where y is an observed score, b_1 is a regression coefficient, F_1 is an unobserved common factor score, and u_y is an unobserved unique factor. In order to identify the scale for the factors F_1 and F_2 , one path leading from each is fixed to a numeric value of 1. All of the covariances between the variables in path diagrams such as those in Figures 1 and 3 can be calculated using rules of path analysis.

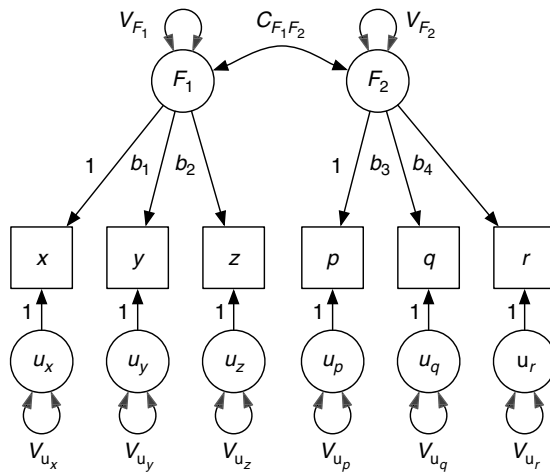


Figure 3 Simple structure confirmatory factor model expressed as a path diagram

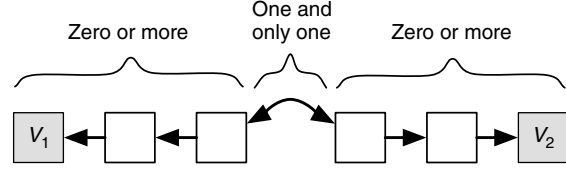


Figure 4 Schematic of the rule for forming a bridge between two variables: exactly one double-headed arrow with zero or more single-headed arrows pointing away from each end towards the selected variable(s)

Path Analysis

The predicted covariance (or correlation) between any two variables v_1 and v_2 in a path model is the sum of all *bridges* between the two variables that satisfy the form shown in Figure 4 [7]. Each bridge contains one and only one double-headed arrow. From each end of the double-headed arrow leads a sequence of zero or more single-headed arrows pointing toward the variable of interest at each end of the bridge. All of the regression coefficients from the single-headed arrows in the bridge as well as the one variance or covariance from the double-headed arrow are multiplied together to form the component of covariance associated with that bridge. The sum of these components of covariance from all bridges between any two selected variables v_1 and v_2 equals the total covariance between v_1 and v_2 .

If a variable v is at both ends of the bridge, then each bridge beginning and ending at v calculates a component of the variance v . The sum of all bridges between a selected variable v and itself calculates the total variance of v implied by the model.

As an example, consider the covariance between x and y predicted by the bivariate regression model shown in Figure 2. There are two bridges between x and y . First, if the double-headed arrow is the variance V_x , then there is a length zero sequence of single-headed arrows from one end of V_x and pointing to x and a length one sequence single-headed arrows leading from the other end of V_x and pointing to y . This bridge is illustrated in Figure 5-a and leads to a covariance component of $V_x b_1$, the product of the coefficients used to form the bridge. Second, if the double-headed arrow is the covariance C_{xz} between x and z , then there is a length zero sequence of single-headed arrows from one end of C_{xz} leading to x and a length one sequence of single-headed arrows leading from the other end of C_{xz} to y . This bridge

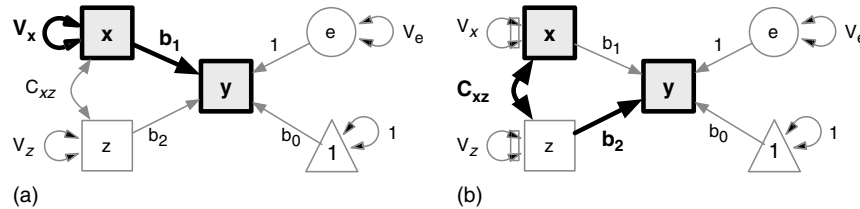


Figure 5 Two bridges between the variables x and y

is illustrated in Figure 5-b and results in a product $C_{xz}b_2$. Thus the total covariance C_{xy} is the sum of the direct effect V_xb_1 and the background effect $C_{xz}b_2$

$$C_{xy} = b_1V_x + b_2C_{xz} \quad (3)$$

Path analysis is especially useful in gaining a deeper understanding of the covariance relations implied by a specified structural model. For instance, when a theory specifies mediating variables, one may work out all of the background covariances implied by the theory. Concepts such as suppression effects or the relationship between measurement interval and cross-lag effects can be clearly explained using a path analytic approach but may seem more difficult to understand from merely studying the equivalent algebra.

References

- [1] Arbuckle, J.L. (1997). *Amos User's Guide*, Version 3.6 SPSS, Chicago.
- [2] Boker, S.M., McArdle, J.J. & Neale, M.C. (2002). An algorithm for the hierarchical organization of path diagrams and calculation of components of covariance between variables, *Structural Equation Modeling* **9**(2), 174–194.
- [3] Duncan, O.D. (1966). Path analysis: sociological examples, *The American Journal of Sociology* **72**(1), 1–16.
- [4] Goldberger, A.S. (1971). Econometrics and psychometrics: a survey of communalities, *Econometrica* **36**(6), 841–868.
- [5] Jöreskog, K.G. (1973). A general method for estimating a linear structural equation system, in *Structural Equation Models in the Social Sciences*, A.S. Goldberger & O.D. Duncan, eds, Seminar, New York, pp. 85–112.
- [6] McArdle, J.J. & Aber, M.S. (1990). Patterns of change within latent variable structural equation modeling, in *New Statistical Methods in Developmental Research*, A. von Eye ed., Academic Press, New York, pp. 151–224.
- [7] McArdle, J.J. & Boker, S.M. (1990). *Rampath*, Lawrence Erlbaum, Hillsdale.
- [8] McArdle, J.J. & McDonald, R.P. (1984). Some algebraic properties of the reticular action model for moment structures, *The British Journal of Mathematical and Statistical Psychology* **87**, 234–251.
- [9] Neale, M.C., Boker, S.M., Xie, G. & Maes, H.H. (1999). *Mx: Statistical Modeling*, Box 126 MCV, 23298: 5th Edition, Department of Psychiatry, Richmond.
- [10] Wolfe, L.M. (1999). Sewall Wright on the method of path coefficients: an annotated bibliography, *Structural Equation Modeling* **6**(3), 280–291.
- [11] Wolfe, L.M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: an annotated bibliography, *Structural Equation Modeling* **10**(1), 1–34.
- [12] Wright, S. (1918). On the nature of size factors, *The Annals of Mathematical Statistics* **3**, 367–374.
- [13] Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs, *Proceedings of the National Academy of Sciences* **6**, 320–332.
- [14] Wright, S. (1934). The method of path coefficients, *The Annals of Mathematical Statistics* **5**, 161–215.

(See also **Linear Statistical Models for Causation: A Critical Review; Structural Equation Modeling: Checking Substantive Plausibility; Structural Equation Modeling: Nontraditional Alternatives**)

STEVEN M. BOKER AND JOHN J. MCARDLE

Pattern Recognition

LUDMILA I. KUNCHEVA AND CHRISTOPHER J. WHITAKER

Volume 3, pp. 1532–1535

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pattern Recognition

Pattern recognition deals with classification problems that we would like to delegate to a machine, for example, scanning for abnormalities in smear test samples, identifying a person by voice and a face image for security purposes, detecting fraudulent credit card transactions, and so on. Each object (test sample, person, transaction) is described by a set of p features and can be thought of as a point in some p -dimensional feature space.

A classifier is a formula, algorithm or technique that outputs a class label for any collection of values of the p features submitted to its input. For designing a classifier, also called **discriminant analysis**, we use a *labeled data set*, $\mathbf{Z}$, of n objects, where each object is described by its feature values and true class label.

The fundamental idea used in statistical pattern recognition is Bayes decision theory [2]. The c classes are treated as random entities that occur with prior probabilities $P(\omega_i)$, $i = 1, \dots, c$. The posterior probability of being in class ω_i for an observed data point $\mathbf{x}$ is calculated using Bayes rule

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)}, \quad (1)$$

where $p(\mathbf{x}|\omega_i)$ is the class-conditional probability density function (pdf) of $\mathbf{x}$, given class ω_i (see **Bayesian Statistics; Bayesian Belief Networks**). According to the Bayes rule, the class with the largest posterior probability is selected as the label of $\mathbf{x}$. Ties are broken randomly. The Bayes rule guarantees the minimum misclassification rate. Sometimes the misclassifications cost differently for different classes. Then we can use a *loss matrix* $\Lambda = [\lambda_{ij}]$, where λ_{ij} is a measure of the loss incurred if we assign class label ω_i when the true label is ω_j . The *minimum risk classifier* assigns $\mathbf{x}$ to the class with the minimum expected risk

$$R_{\mathbf{x}}(\omega_i) = \sum_{j=1}^c \lambda_{ij} P(\omega_j|\mathbf{x}). \quad (2)$$

In general, the classifier output can be interpreted as a set of c degrees of support, one for each class (discriminant scores obtained through discriminant functions). We label $\mathbf{x}$ in the class with the largest support.

In practice, the prior probabilities and the class-conditional pdfs are not known. The pdfs can be estimated from the data using either a parametric or non-parametric approach. Parametric classifiers assume the form of the probability distributions and then estimate the parameters from $\mathbf{Z}$. The linear and quadratic discriminant classifiers, which assume multivariate normal distributions as $p(\mathbf{x}|\omega_i)$, are commonly used (see **Discriminant Analysis**). Nonparametric classifier models include the *k-nearest neighbor classifier* (k -nn) and *kernel classifiers* (e.g., Parzen, support vector machines (SVM)). The k -nn classifier assigns $\mathbf{x}$ to the class most represented among the closest k neighbors of $\mathbf{x}$.

Instead of trying to estimate the pdfs and applying Bayes' rule, some classifier models directly look for the best discrimination boundary between the classes, for example, **classification and regression trees**, and **neural networks**.

Figure 1 shows a two-dimensional data set with two banana-shaped classes. The dots represent the observed data points, and members of the classes are denoted by their color. Four classifiers are built, each one splitting the feature space into two classification regions. The boundary between the two regions is denoted by the white line. The linear discriminant classifier results in a linear boundary, while the quadratic discriminant classifier results in a quadratic boundary. While both these models are simple, for this difficult data set, neither is adequate. Here, the greater flexibility afforded by the 1-nn and neural network models result in more accurate but complicated boundaries. Also, the class regions found by 1-nn and the neural network may not be connected sets of points, which is not possible for the linear and quadratic discriminant classifiers.

Classifier Training and Testing

In most real life problems we do not have a ready made classification algorithm. Take for example, a classifier that recognizes expressions of various feelings from face images. We can only provide a rough guidance in a linguistic format and pick out features that we believe are relevant for the task. The classifier has to be trained by using a set of labeled examples. The training depends on the classifier model. The nearest neighbor classifier (1-nn) does not require any training; we can classify a new data point

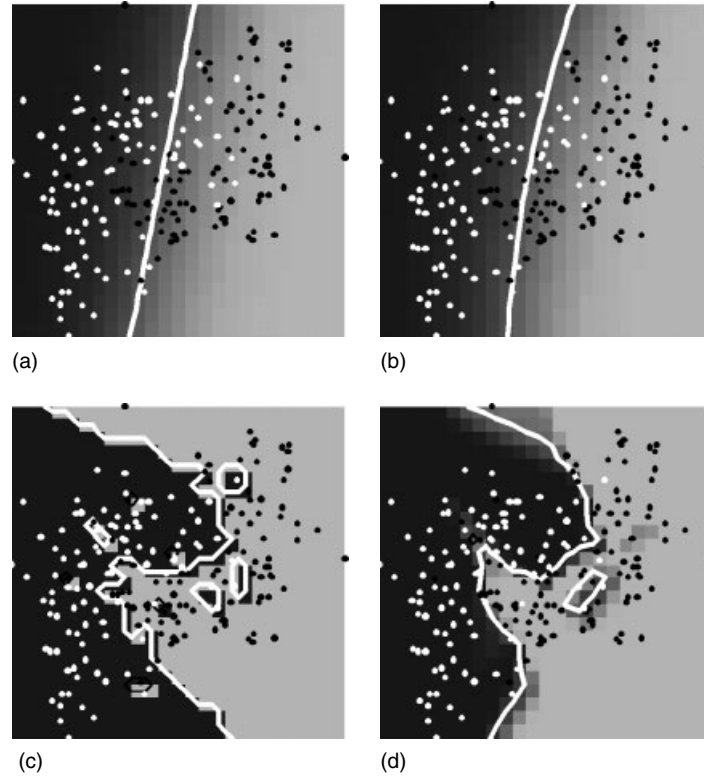


Figure 1 Classification regions for two banana-shaped classes found by four classifiers (a) Linear, (b) Quadratic, (c) 1-nn, (d) Neural Network

right away by finding its nearest neighbor in $\mathbf{Z}$. On the other hand, the lack of a suitable training algorithm for neural networks was the cause for their dormant period between 1940s and 1980s (the error back-propagation algorithm revitalized their development). When trained, some classifiers can provide us with an interpretable decision strategy (e.g., tree models and k -nn) whereas other classifiers behave as black boxes (e.g., neural networks). Even when we can verify the logic of the decision making, the ultimate judge of the classifier performance is the classification error. Estimating the **misclassification rate** of our classifier is done through the training protocol. Part of the data set, $\mathbf{Z}$, is used for training and the remaining part is left for testing. The most popular training/testing protocol is **cross-validation**. $\mathbf{Z}$ is divided into K approximately equal parts, one is left for testing, and the remaining $K - 1$ are pooled as the training set. This process is repeated K times (K -fold cross-validation) leaving aside a different part each time.

The error of the classifier is the averaged testing error across the K testing parts.

Variable Selection

Not all features are important for the classification task. Classifiers may perform better with fewer features. This is a paradox from an information-theoretic point of view. Its explanation lies in the fact that the classifiers that we use and the parameter estimates that we calculate are imperfect; therefore, some of the supposed information is actually noise to our model. *Feature selection* reduces the original feature set to a subset without adversely affecting the classification performance. *Feature extraction*, on the other hand, is a dimensionality reduction approach whereby all initial features are used and a small amount of new features are calculated from them (e.g., **principal component analysis**, **projection pursuit**, **multidimensional scaling**).

There are two major questions in feature selection: what criterion should we use to evaluate the subsets? and how do we search among all possible subset-candidates? Since the final goal is to have an accurate classifier, the most natural choice of a criterion is the minimum error of the classifier built on the subset-candidate. Methods using a direct estimate of the error are called *wrapper methods*. Even with modern computational technology, training a classifier and estimating its error for each examined subset of features might be prohibitive. An alternative class of feature selection methods where the criterion is indirectly related to the error are called *filter methods*. Here the criterion used is a measure of discrimination between the classes, for example, the **Mahalanobis distance** between the class centroids.

For large p , checking all possible subsets is often not feasible. There are various search algorithms, the simplest of which are the *sequential forward selection* (SFS) and the *sequential backward selection* (SBS). In SFS, we start with the single best feature (according to the chosen criterion) and add one feature at a time. The second feature to enter the selected subset will be the feature that makes the best pair with the feature already selected. The third feature is chosen so that it makes the best triple containing the already selected two features, and so on. In SBS, we start with the whole set of features and remove the single feature which gives the best remaining subset of $p - 1$ features. Next we remove the feature that results in the best remaining subset of $p - 2$ features, and so on. SFS and SBS, albeit simple, have been found to be surprisingly robust and accurate. A modification of these is the *floating search* feature selection algorithm, which leads to better results at the expense of an expanded search space. Feature selection is an art rather than science as it relies on heuristics, intuition, and domain knowledge. Among many others, *genetic algorithms* have been applied for feature selection with various degree of success.

Cluster Analysis

In some problems, the class labels are not defined in advance. Then, the problem is to find a class structure in the data set, if there is any. The number of clusters is usually not specified in advance, which makes the problem even more difficult. If we guess wrongly, we may impose a structure onto a data set that does not

have one or may fail to discover an existing structure. Cluster analysis procedures can be roughly grouped into *hierarchical* and *iterative optimization methods* (see **Hierarchical Clustering; k-means Analysis**).

Classifier Ensembles

Instead of using a single classifier, we may combine the outputs of several classifiers in an attempt to reach a more accurate or reliable solution. At this stage, there are a large number of methods, directions, and paradigms in designing classifier ensembles but there is no agreed taxonomy for this relatively young area of research.

Fusion and Selection. In classifier fusion, we assume that all classifiers are ‘experts’ across the whole feature space, and therefore their votes are equally important for any $\mathbf{x}$. In classifier selection, first ‘an oracle’ or meta-classifier decides whose *region of competence* $\mathbf{x}$ is in, and then the class label of the nominated classifier is taken as the ensemble decision.

Decision Optimization and Coverage Optimization. Decision optimization refers to ensemble construction methods that are primarily concerned with the combination rule assuming that the classifiers in the ensembles are given. Coverage optimization looks at building the individual classifiers assuming a fixed classification rule.

Five simple combination rules (combiners) are illustrated below. Suppose that there are 3 classes and 7 classifiers whose outputs for a particular $\mathbf{x}$ (discriminant scores) are organized in a 7 by 3 matrix (called a *decision profile* for $\mathbf{x}$) as given in Table 1. The overall support for each class is calculated by applying a simple operation to the discriminant scores for that class only. The majority vote operates on the label outputs of the seven classifiers.

Each possible class label occurs as the final ensemble label in the five shaded cells. This shows the flexibility that we have in choosing the combination rule for the particular problem.

We can consider the classifier outputs as new features, disregard their context as discriminant scores, and use these features to build a classifier. We can thus build hierarchies of classifiers (*stacked generalization*). There are many combination methods

Table 1 An ensemble of seven classifiers, $D_1, \dots, D_7$: the decision profile for $\mathbf{x}$ and five combination rules

	Support for ω_1	Support for ω_2	Support for ω_3	Label output
D_1	0.24	0.44	0.56	ω_3
D_2	0.17	0.13	0.59	ω_3
D_3	0.22	0.32	0.86	ω_3
D_4	0.17	0.40	0.49	ω_3
D_5	0.27	0.77	0.45	ω_2
D_6	0.51	0.90	0.06	ω_2
D_7	0.29	0.46	0.03	ω_2
Minimum	0.17	0.13	0.03	ω_1
Maximum	0.51	0.90	0.86	ω_2
Average	0.27	0.49	0.43	ω_2
Product	0.0001	0.0023	0.0001	ω_2
Majority	—	—	—	ω_3

proposed in the literature that involve various degrees of training.

Within the coverage optimization group are **bagging**, **random forests** and **boosting**. Bagging takes L random samples (with replacement) from $\mathbf{Z}$ and builds one classifier on each sample. The ensemble decision is made by the majority vote. The success of bagging has been explained by its ability to reduce the variance of the classification error of a single classifier model. Random forests are defined as a variant of bagging such that the production of the individual classifiers depends on a random parameter and independent sampling. A popular version of random forests is an ensemble where each classifier is built upon a random subsets of features, sampled with replacement from the initial feature set.

A boosting algorithm named *AdaBoost* has been found to be even more successful than bagging. Instead of drawing random bootstrap samples, AdaBoost designs the ensemble members one at a time, based on the performance of the previous member. A set of weights is maintained across the data points in $\mathbf{Z}$. In the resampling version of AdaBoost, the weights are taken as probabilities. Each training sample is drawn using the weights. A classifier is built on

the sampled training set, and the weights are modified according to its performance. Points that have been correctly recognized get smaller weights and points that have been misclassified get larger weights. Thus difficult to classify objects will have more chance to be picked in the subsequent training sets. The procedure stops at a predefined number of classifiers (e.g., 50). The votes are combined by a weighted majority vote, where the classifier weight depends on its error rate. AdaBoost has been proved to reduce the training error to zero. In addition, when the training error does reach zero, which for most classifier models means that they have been overtrained and the testing error might be arbitrarily high, AdaBoost ‘miraculously’ keeps reducing the testing error further. This phenomenon has been explained by the *margin theory* but with no claim about a global convergence of the algorithm. AdaBoost has been found to be more sensitive to noise in the data than bagging. Nevertheless, AdaBoost has been declared by Leo Breiman to be the ‘most accurate available off-the-shelf classifier’ [1].

Pattern recognition is related to artificial intelligence and machine learning. There is renewed interest in this topic as it underpins applications in modern domains such as **data mining**, document classification, financial forecasting, organization and retrieval of multimedia databases, microarray data analysis (see **Microarrays**), and many more [3].

References

- [1] Breiman, L. (1998). Arcing classifiers, *The Annals of Statistics* **26**(3), 801–849.
- [2] Duda, R.O., Hart, P.E. & Stork, D.G. (2001). *Pattern Classification*, 2nd Edition, John Wiley & Sons, New York.
- [3] Jain, A.K., Duin, R.P.W. & Mao, I. (2000). Statistical pattern recognition: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1), 4–37.

LUDMILA I. KUNCHEVA AND CHRISTOPHER
J. WHITAKER

Pearson, Egon Sharpe

BRIAN S. EVERITT

Volume 3, pp. 1536–1536

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pearson, Egon Sharpe

Born: August 11, 1895, in Hampstead, London.

Died: June 12, 1980, in Midhurst, Sussex.

Egon Pearson was born in Hampstead, London in 1895, the only son of **Karl Pearson**, the generally acknowledged founder of biometrics and one of the principal architects of the modern theory of mathematical statistics. The younger Pearson began to read mathematics at Trinity College, Cambridge in 1914, but his undergraduate studies were interrupted first by a severe bout of influenza and then by war work. It was not until 1920 that Pearson received his B.A. after taking the Military Special Examination in 1919, which had been set up to cope with those who had had their studies disrupted by the war.

In 1921, Pearson took up a position as lecturer in his father's Department of Applied Statistics at University College. But it was not until five years later in 1926 that Karl Pearson allowed his son to begin lecturing in the University and only then because his own health prevented him from teaching. During these five years, Egon Pearson did manage to produce a stream of high-quality research publications on statistics and also became an assistant editor of *Biometrika*. In 1925, **Jerzy Neyman** joined University College as a Rockefeller Research Fellow and befriended the rather introverted Egon. The result was a joint research project on the problems of hypothesis testing that eventually led to a general approach to statistical and scientific problems known as **Neyman–Pearson inference**. Egon Pearson also collaborated with **W. S. Gossett**, whose ideas played a major part in Pearson's own discussions with

Neyman, and in drawing his attention to the important topic of robustness.

In 1933, Karl Pearson retired from the Galton Chair of Statistics and, despite his protests, his Department was split into two separate parts, one (now the Department of Eugenics) went to **R. A. Fisher**, the other, a Department of Applied Statistics, went to Egon Pearson. Working in the same institute as Fisher, a man who had attacked his father aggressively, was not always a comfortable experience for Egon Pearson.

In 1936, Karl Pearson died and his son became managing editor of *Biometrika*, a position he continued to hold until 1966. During the Second World War, Egon Pearson undertook war work on the statistical analysis of the fragmentation of shells hitting aircraft, work for which he was awarded a CBE in 1946. It was during this period that he became interested in the use of statistical methods in quality control.

In 1955 Pearson was awarded the Gold Medal of the Royal Statistical Society and served as its President from 1955 to 1956. His presidential address was on the use of geometry in statistics. In 1966, Egon Pearson was eventually elected Fellow of the Royal Society. In the 1970s, Pearson worked on the production of an annotated version of his father's lectures on the early history of statistics, which was published in 1979 [1] just before his death in 1980.

Reference

- [1] Pearson, E.S., ed., (1979). *The History of Statistics in the Seventeenth and Eighteenth Centuries*, Macmillan Publishers, New York.

BRIAN S. EVERITT

Pearson, Karl

MICHAEL COWLES

Volume 3, pp. 1536–1537

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pearson, Karl

Born: March 27, 1857, in London, UK.

Died: April 27, 1936, in Surrey, UK.

In the dozen or so years that spanned the turn of the nineteenth into the twentieth century, Karl Pearson constructed the modern discipline of statistics. His goal was to make biology and heredity quantitative disciplines and to establish a scientific footing for a eugenic social philosophy.

Pearson entered Kings College, Cambridge, in 1876 and graduated in mathematics third in his year (third wrangler) in 1879. In the early 1880s, Carl with a 'C' gave way to Karl with a 'K'. There is no evidence for the claim that the change was due to his espousal of Socialism and in honor of Marx. The spelling of his name had varied since childhood. After Cambridge, he studied in Germany and then began a search for a job. There was a brief consideration of a career in law and he was called to the Bar, after his second attempt at the Bar examinations, in 1882.

In 1880, Pearson published, under the pen name 'Loki', *The New Werther*, followed in 1882 by *The Trinity, a 19th Century Passion Play* – works that showed that he was not a poet! During the first half of the 1880s, he gave lectures on ethics, free thought, socialism, and German history and culture. He was the inspiration and a founder of a club formed to discuss women's issues and through it met his wife, Maria Sharpe.

In 1884, he was appointed Goldsmid Professor of Applied Mathematics and Mechanics at University College, London. Pearson's meeting his friend and colleague Walter Weldon steered Pearson away from his work in applied mathematics toward a new discipline of *biometrics*. Later, he became the first professor of National Eugenics. He was the editor of *Biometrika*, a journal founded with the assistance of **Francis Galton**, for the whole of his life.

A dispute over the degrees of freedom for chi-square is one of the more bizarre controversies that Pearson spent much fruitless time on. It had been pointed out to Pearson that the *df* to be applied when the expected frequencies in a contingency table were constrained by the marginal totals were not to be calculated in the same way as they are when

the expected frequencies were known *a priori* as in the goodness of fit test (*see Goodness of Fit for Categorical Variables*). Raymond Pearl also raised problems about the test, compounding his heresy by doing it in the context of Mendelian theory, a view that Pearson opposed. The *df* point was also raised by **Fisher** in 1916. Pearson attempted to argue that the aim was to find the lowest chi-square for a given grouping rather than its *p* level, thus, to be most charitable, clouding the issue, or, to be much less charitable, misunderstanding his own test.

The well-known deviation score formula for the correlation coefficient was shown by Pearson to be the best in a 1896 paper, which includes a section on the mathematical foundations of correlation [2, 3]. Here, he acknowledges the work of Bravais and of Edgeworth. Twenty-five years later, he repudiated these statements, emphasizing his own work and that of Galton. Even in this, his seminal contribution, a row with Fisher over the probability distribution of *r* ensued. Pearson and his colleagues embarked on the laborious task of calculating the ordinates of the frequency curves for *r* for various *df*. Fisher arrived at a much simpler method using his *r* to *t* (or *r* to *z*) transformation, but Pearson could not or would not accept it.

Pearson's most important contributions were produced before Weldon's untimely death in 1906. His book *The Grammar of Science* [1] has not received the attention it deserves. His contribution to statistics and the generous help that he gave many students has to be and will continue to be acknowledged, but his work was weakened by controversy.

References

- [1] Pearson, K. (1892). *The Grammar of Science*, Walter Scott, London.
- [2] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. regression, heredity and panmixia, *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318.
- [3] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine* 5th series, **50**, 157–175.

MICHAEL COWLES

Pearson Product Moment Correlation

DIANA KORNBROT

Volume 3, pp. 1537–1539

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pearson Product Moment Correlation

Pearson's product moment correlation, r , describes the strength of the linear relation, between two metric (interval or ratio) variables (see **Scales of Measurement**) such as height and weight. A high magnitude of r indicates that the squared distance of the points from the best fitting straight line on a **scatterplot** (plot of Y v X or X v Y) is small. The statistic r is a measure of association (see **Measures of Association**) and does NOT imply causality, in either direction. For valid interpretation of r , the variables should have a bivariate normal distribution (see **Catalogue of Probability Density Functions**), and a linear relation, as in Panels A and B of Figure 1.

The slopes of linear regressions of standardized Y scores on standardized X scores, and of standardized X scores on standardized Y scores generate r (see **Standardized Regression Coefficients**). Steeper

slopes (on a standardized scale) correspond to larger magnitudes of r . (A variable is standardized by a linear transformation to a new variable with mean zero and standard deviation 1).

r^2 gives the proportion of variance (variability) accounted for by the linear relation (see **R-squared, Adjusted R-squared**). An r of 0.3, as frequently found between personality measures, only accounts for 10% of the variability.

Figure 1 provides scatterplots of some possible relations between X and Y , and the associated values of r and r^2 . Panel (a) shows a moderate positive linear correlation, while Panel (b) shows a strong negative linear correlation. Panel (c) shows no correlation. In Panel (d), r is also near zero. The relationship is strong, but nonlinear, as might occur between stress and performance. Panel (e) shows a negative nonlinear relation, as might occur if Y is light intensity needed to see and X is time in the dark. The relationship is negative (descending), but nonlinear. Panel (f) has same data as Panel (c), with an added **outlier** at (29,29). r is dramatically changed, from -0.009 to 0.56 .

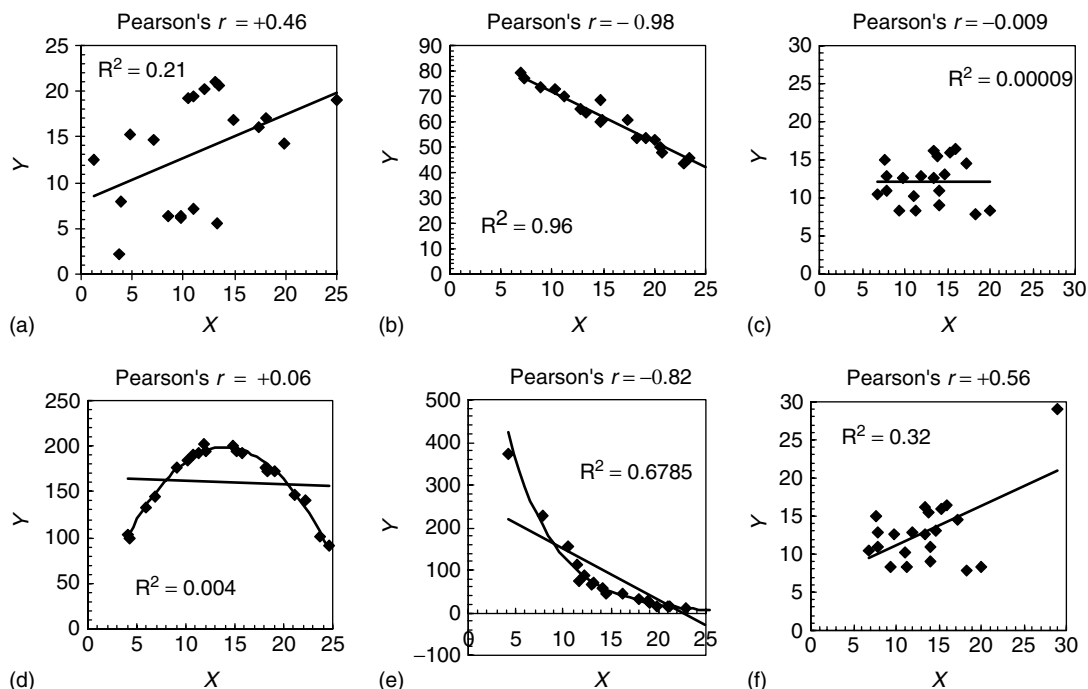


Figure 1 Scatterplots of Y against X for a variety of relationships and correlation values. The correlation coefficient, r , should *never* be interpreted without a scatterplot

2 Pearson Product Moment Correlation

Calculation

The population values of r for N pairs of points (X, Y) is given by:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}. \quad (1)$$

Equation 1, applied to a sample, gives a biased estimate of the population value. The adjusted correlation, r_{adj} , gives an unbiased estimate [1] (*see R-squared, Adjusted R-squared*).

$$r_{\text{adj}} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}}$$

or

$$r_{\text{adj}} = r \sqrt{1 - \left(\frac{1}{N - 2}\right) \left(\frac{1 - r^2}{r^2}\right)}. \quad (2)$$

The value of r overestimates r_{adj} , often with large bias (e.g., 75% overestimate for $r_{\text{adj}} = 0.10$ with $N = 50$). So, the value r_{adj} *should* be reported (although reporting r is all too common).

Confidence Limits and Hypothesis Testing

The sampling distributions of r only approaches normality for very large N . Equation 3 gives a

statistic that is t -distributed with $N - 2$ degrees of freedom when population r is zero [1]. It should be used to test the null hypothesis that there is no association.

$$t = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}} \quad (3)$$

It might be argued that r_{adj} should be used in (3). However, standard statistical packages (e.g., SPSS, JMP-IN) use r .

Confidence intervals and hypotheses about values of r other than 0 can be obtained from the Fisher transformation, which gives normally distributed statistic, z_r , with variance $1/(N - 3)$ [1, 2].

$$z_r = \frac{1}{2} \ln \left[\frac{1 + r}{1 - r} \right]. \quad (4)$$

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.
- [2] Sheskin, D.J. (2000). *Handbook of Parametric and Non-parametric Statistical Procedures*, 2nd Edition, Chapman & Hall, London.

DIANA KORNROT

Percentiles

DAVID CLARK-CARTER

Volume 3, pp. 1539–1540

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Percentiles

By definition, percentiles are **quantiles** that divide the set of numbers into 100 equal parts. As an example, the 10th percentile is the number that has 10% of the set below it, while the 90th percentile is the number that has 90% of the set below it. Moreover, the 25th percentile is the first **quartile** (Q_1), the 50th percentile is the second quartile (the **median**, Q_2), and the 75th percentile is the third quartile (Q_3).

The general method for finding percentiles is the same as that for quantiles. However, suppose that we have the reading ages for a class of eight-year-old children, as shown in column 2 of Table 1, and we want to ascertain each child's percentile position or rank. One way of doing this is to find the cumulative frequency for each child's reading score, then convert these to cumulative proportions by dividing each by the sample size and then convert these proportions into cumulative percentages, as shown in columns 3–5 of Table 1. From the table, we can see that the first three children are in the 20th percentile for reading age. If we had the norms for the test, we would be able to find each child's percentile point for the population. We note too that the median reading age (50th percentile) for the children is between 101 and 103 months.

Certain percentiles such as the 25th, 50th, and 75th (and in some cases the 10th and 90th) are represented on **box plots** [1].

Table 1 Reading ages in months of a class of eight-year-old children with the associated cumulative frequencies, proportions, and percentages

Child	Reading age (in months)	Cumulative frequency	Cumulative proportion	Cumulative percentage
1	84	1	0.07	6.67
2	87	2	0.13	13.33
3	90	3	0.20	20.00
4	94	4.5	0.30	30.00
5	94	4.5	0.30	30.00
6	100	6	0.40	40.00
7	101	7	0.47	46.67
8	103	8	0.53	53.33
9	104	9.5	0.63	63.33
10	104	9.5	0.63	63.33
11	106	11	0.73	73.33
12	107	12	0.80	80.00
13	108	13	0.87	86.67
14	111	14	0.93	93.33
15	112	15	1.00	100.00

Reference

- [1] Cleveland, W.S. (1985). *The Elements of Graphing Data*, Wadsworth, Monterey, California.

DAVID CLARK-CARTER

Permutation Based Inference

CLIFFORD E. LUNNEBORG

Volume 3, pp. 1540–1541

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Permutation Based Inference

Permutation-based inference had its origins in the 1930s in the seminal efforts of **R. A. Fisher** [2] and **E. J. G. Pitman** [6, 7]. Fisher described an alternative to the matched-pairs t Test that made no assumption about the form of the distribution sampled while Pitman titled his articles extending permutation inference to the two-sample, multiple-sample, and correlation contexts as ‘significance tests which may be applied to samples from any population.’

The idea behind permutation tests is best introduced by way of an example. Let

y : (18, 21, 25, 30, 31, 32)

be a random sample of size six from a distribution, Y , of body-mass-indices (BMI) for women in the freshman class at Cedar Grove High School and

z : (19, 20, 22, 23, 28, 35)

be a second random sample drawn from the BMI distribution, Z , for men in that class.

The null hypothesis is that the two distributions, Y and Z are identical. This is the same null hypothesis underlying the independent-samples t Test, but without the assumption that Y and Z are normally distributed. Under this null hypothesis, the observations in the two samples are said to be *exchangeable* with one another [3]. For example, the BMI of 18 in the women’s sample is just as likely to have turned up in the sample from the men’s distribution, while the BMI of 35 in the men’s sample is just as likely to have turned up in the sample from the women’s distribution; that is, under the null hypothesis, a women’s sample of

y : (21, 25, 30, 31, 32, 35)

and a men’s sample of

z : (18, 19, 20, 22, 23, 28)

are just as likely as the samples actually observed.

These, of course, are not the only arrangements of the observed BMIs that would be possible under the null hypothesis. Every permutation of the 12 indices, 6 credited to the women’s distribution and 6

to the men’s distribution, can be attained by a series of pairwise exchanges between the two samples, and, hence, has the same chance of having been observed. For our example, there are $[12!/(6! \times 6!)] = 924$ arrangements of the observed BMIs, each arrangement equally likely under the null hypothesis.

A permutation test results when we are able to say whether the observed arrangement of the data points is unlikely under the null hypothesis. To pursue the analogy with the two-sample t Test, we ask whether a difference in sample means as large as that observed,

$$\text{Mean}(y) - \text{Mean}(z) = 1.66667, \quad (1)$$

is unlikely under the null hypothesis. The answer depends on where the value of 1.66667 falls in the *reference distribution* made up of the differences in mean computed for each of the 924 permutations of the BMIs. In fact, 307 of the 924 mean differences are greater than or equal to 1.6667 and another 307 are smaller than or equal to -1.6667 . Taking the alternative hypothesis to be nondirectional, we can assign a P value of $(614/924) = 0.6645$ to the test. There is no evidence here that the two samples could not have been drawn from identical BMI distributions.

The permutation-test decision to reject or not reject the null hypothesis depends only upon a distribution generated by the data at hand. It does not rely upon unverifiable assumptions about the entire population. If the observations are exchangeable, the resulting P value will be exact. Moreover, permutation tests can be applied to data derived from finite as well as infinite populations [6].

The *reference distribution* of a permutation test statistic is not quite the same thing as the **sampling distribution** we associate with, say, the normal-theory t Test. The latter is based on all possible pairs of independent samples, six observations each from identical Y and Z distributions, while the former considers only those pairs of samples arising from the permutation of the observed data points. Because of this rooting in the observed data, the permutation test is said to be a *conditional test*.

The P value in our example was computed more-or-less instantaneously by the **perm.test** function from an **R** package for exact rank tests (www.R-project.org). In the 1930s, of course, enumerating the 924 possible permutations and tabulating the resulting test statistics would have been a prodigious if not prohibitive computational task. In fact,

2 Permutation Based Inference

the thrust of [2, 6, 7] was that the permutation test P value could be *approximated* by the P value associated with a parametric test. Fisher [2] summarized his advice to use the parametric approximation this way: ‘(although) the statistician does not carry out this very tedious process, his conclusions have no justification beyond the fact they could have been arrived at by this very elementary method.’

Today, we have computing facilities, almost universally available, that are fast and inexpensive enough to make it no longer necessary or desirable to substitute an approximate test for an exact one.

Permutation tests offer more than distribution-free alternatives to established parametric tests. As noted by Bradley [1], the real strength of permutation-based inference lies in the ability to choose a test statistic best suited to the problem at hand, rather than be limited by the ability of precalculated tables. An excellent example lies in the analysis of k -samples, where permutation-based tests are available for both ordered and unordered categories as well as for a variety of loss functions [3–5].

Permutation-based inference has been applied successfully to correlation (*see* **Pearson Product Moment Correlation**), **contingency tables**, k -sample comparisons, **multivariate analysis**, with censored,

missing, or **repeated measures** data, and to situations where parametric tests do not exist.

The author gratefully acknowledges the assistance of Phillip Good in the preparation of this article.

References

- [1] Bradley, J.V. (1968). *Distribution-Free Statistical Tests*, Prentice Hall.
- [2] Fisher, R.A. (1935). *Design of Experiments*, Hafner, New York.
- [3] Good, P. (2004). *Permutation Tests*, 3rd Edition, Springer, New York.
- [4] Mielke, P.W. & Berry, K.J. (2001). *Permutation Methods – A Distance Function Approach*, Springer, New York.
- [5] Pesarin, F. (2001). *Multivariate Permutation Tests*, Wiley & Sons, New York.
- [6] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, Parts I and II, *Journal of the Royal Statistical Society, Supplement* **4**, 119–130, 225–232.
- [7] Pitman, E.J.G. (1938). Significance tests which may be applied to samples from any population. Part III. The analysis of variance test, *Biometrika* **29**, 322–335.

CLIFFORD E. LUNNEBORG

Person Misfit

ROB R. MEIJER

Volume 3, pp. 1541–1547

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Person Misfit

Since the beginning of psychological and educational standardized testing, inaccuracy of measurement has received widespread attention. In this overview, we discuss research methods for determining the fit of individual item score patterns to a test model. In the past two decades, important contributions to assessing individual test performance arose from item response theory (IRT) (*see Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data*); these contributions are summarized as person-fit research. However, because person-fit research has also been conducted without IRT modeling, approaches outside the IRT framework are also discussed. This review is restricted to dichotomous (0, 1) item scores. For a comprehensive review of person-fit, readers are referred to [9] and [10].

As a measure of a person's ability level, the total score (or the trait level estimate) may be inadequate. For example, a person may guess some of the correct answers to multiple-choice items, thus raising his/her total score on the test by luck and not by ability, or an examinee not familiar with the test format may due to this unfamiliarity obtain a lower score than expected on the basis of his/her ability level [23]. Inaccurate measurement of the trait level may also be caused by sleeping behavior (inaccurately answering the first questions in a test as a result of, for example, problems of getting started), cheating behavior (copying the correct answers of another examinee), and plodding behavior (working very slowly and methodically, and, as a result, generating item score patterns that are too good to be true given the stochastic nature of a person's response behavior as assumed by most IRT models).

It is important to realize that not all types of aberrant behavior affect individual test scores. For example, a person may guess the correct answers of some of the items but also guess wrong on some of the other items, and, as the result of the stochastic nature of guessing, this process may not result in substantially different test scores under most IRT models to be discussed below. Whether aberrant behavior leads to misfitting item score patterns depends on numerous factors such as the type and the amount of aberrant behavior.

Furthermore, it may be noted that all methods discussed can be used to detect misfitting item score patterns, but that several of these methods do not allow the recovery of the mechanism that created the deviant item score patterns. Other methods explicitly test against specific violations of a test model assumption, or against particular types of deviant item score patterns. The latter group of methods, therefore, may facilitate the interpretation of misfitting item score patterns.

Person-fit Methods Based on Group Characteristics

Most person-fit statistics compare an individual's observed and expected item scores across the items from a test. The expected item scores are determined on the basis of an IRT model or on the basis of the observed item means in the sample. In this section, group-based statistics are considered. In the next section, IRT-based person-fit statistics are considered.

Let n persons take a test consisting of k items, and let π_g denote the proportion-correct score on item g that can be estimated from the sample by $\hat{\pi}_g = n_g/n$, where n_g is the number of 1 scores. Furthermore, let the items be ordered and numbered according to decreasing proportion-correct score (increasing item difficulty): $\pi_1 > \pi_2 > \dots > \pi_k$, and let the realization of a dichotomous (0,1) item score be denoted by $X_g = x_g$ ($g = 1, \dots, k$). Examinees are indexed i , with $i = 1, \dots, n$.

Most person-fit statistics are a count of certain score patterns for item pairs (to be discussed shortly), and compare this count with the expectation under the deterministic Guttman [4] model. Let θ be the latent trait (*see Latent Variable*) known from IRT (to be introduced below), and let δ be the location parameter, which is a value on the θ scale. $P_g(\theta)$ is the conditional probability of giving a correct answer to item g . The Guttman model is defined by

$$\theta < \delta_g \Leftrightarrow P_g(\theta) = 0; \quad (1)$$

and

$$\theta \geq \delta_g \Leftrightarrow P_g(\theta) = 1. \quad (2)$$

The Guttman model thus excludes a correct answer on a relatively difficult item h and an incorrect answer on an easier item g by the same examinee:

2 Person Misfit

$X_h = 1$ and $X_g = 0$, for all $g < h$. Item score combinations (0, 1) are called ‘errors’ or ‘inversions’. Item score patterns (1, 0), (0, 0), and (1, 1) are permitted, and known as ‘Guttman patterns’ or ‘conformal’ patterns.

Person-fit statistics that are based on group characteristics compare an individual’s item score pattern with the other item score patterns in the sample. Rules-of-thumb have been proposed for some statistics, but these rules-of-thumb are not based on sampling distributions, and, therefore, difficult to interpret.

Although group-based statistics may be sensitive to misfitting item score patterns, a drawback is that their null distributions are unknown and, as a result, it cannot be decided on the basis of significance probabilities when a score pattern is unlikely given a nominal Type I error rate. In general, let t be the observed value of a person-fit statistic T . Then, the significance probability or probability of exceedance is defined as the probability under the sampling distribution that the value of the test statistic is smaller than the observed value: $p^* = P(T \leq t)$, or larger than the observed value: $p^* = P(T \geq t)$, depending on whether low or high values of the statistic indicate aberrant item score patterns. Although it may be argued that this is not a serious problem as long as one is only interested in the use of a person-fit statistic as a descriptive measure, a more serious problem is that the distribution of the numerical values of most group-based statistics is dependent on the total score [2]. This dependence implies that when one critical value is used across total scores, the probability of classifying a score pattern as aberrant is a function of the total score, which is undesirable. To summarize, it can be concluded that the use of group-based statistics has been explorative, and, with the increasing interest in IRT modeling, interest in group-based person-fit has gradually declined.

Person-fit Measures Based on Item Response Theory

In IRT, the probability of obtaining a correct answer on item g ($g = 1, \dots, k$) is a function of the latent trait value (θ) and characteristics of the item such as the location δ [21]. This conditional probability $P_g(\theta)$ is the item response function (IRF). Further, we define the vector with item score random

variables $\mathbf{X} = (X_1, \dots, X_k)$, and a realization $\mathbf{x} = (x_1, \dots, x_k)$. $P_g(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1-, 2-, 3-PLM). The 3-PLM is defined as

$$P_g(\theta) = \gamma_g + \frac{(1 - \gamma_g) \exp[\alpha_g(\theta - \delta_g)]}{1 + \exp[\alpha_g(\theta - \delta_g)]}, \quad (3)$$

where γ_g is the lower asymptote (γ_g is the probability of a 1 score for low-ability examinees (that is, $\theta \rightarrow -\infty$)); α_g is the slope parameter (or item discrimination parameter); and δ_g is the item location parameter. The 2-PLM can be obtained by fixing $\gamma_g = 0$ for all items; and the 1-PLM, or **Rasch model**, can be obtained by additionally fixing $\alpha_g = 1$ for all items.

A major advantage of IRT models is that the goodness-of-fit of a model to empirical data can be investigated. Compared to group-based person-fit statistics, this provides the opportunity of evaluating the fit of item score patterns to an IRT model. To investigate the goodness-of-fit of item score patterns, several IRT-based person-fit statistics have been proposed.

Let $w_g(\theta)$ be a suitable function for weighting the item scores and adapting person-fit scale scores, respectively. Following [18], a general form in which most person-fit statistics can be expressed is

$$V = \sum_{g=1}^k [X_g - P_g(\theta)] w_g(\theta). \quad (4)$$

Examples of these kinds of statistics are given in [19] and [23].

Likelihood-based Statistics

Most studies, to be discussed below, have been conducted using some suitable function of the loglikelihood function

$$l = \sum_{g=1}^k \{X_g \ln P_g(\theta) + (1 - X_g) \ln [1 - P_g(\theta)]\}. \quad (5)$$

This statistic, first proposed by Levine and Rubin [7], was further developed and applied in a series of articles by Drasgow et al. [2]. Two problems exist when using l as a fit statistic. The first problem is that l is not standardized, implying that

the classification of an item score pattern as normal or aberrant depends on θ . The second problem is that for classifying an item score pattern as aberrant, a distribution of the statistic under the null hypothesis of fitting item scores is needed, and for l , this null distribution is unknown. Solutions proposed for these two problems are the following.

To overcome the problem of dependence on θ and the problem of unknown sampling distribution, Drasgow et al. [2] proposed a standardized version l_z of l , which was less confounded with θ , and which was purported to be asymptotically standard normally distributed; l_z is defined as

$$l_z = \frac{l - E(l)}{[\text{var}(l)]^{1/2}}, \quad (6)$$

where $E(l)$ and $\text{var}(l)$ denote the expectation and the variance of l , respectively. These quantities are given by

$$E(l) = \sum_{g=1}^k \{P_g(\theta) \ln[P_g(\theta)] \times [1 - P_g(\theta)] \ln[1 - P_g(\theta)]\}, \quad (7)$$

and

$$\text{var}(l) = \sum_{g=1}^k P_g(\theta)[1 - P_g(\theta)] \left[\ln \frac{P_g(\theta)}{1 - P_g(\theta)} \right]^2. \quad (8)$$

Molenaar and Hoijsink [11] argued that l_z is only standard normally distributed when the true θ values are used, but a problem arises in practice when θ is replaced by the maximum likelihood estimate $\hat{\theta}$. Using an estimate and not the true θ will have an effect on the distribution of a person-fit statistic. These studies showed that when **maximum likelihood estimates** $\hat{\theta}$ were used, the variance of l_z was smaller than expected under the standard normal distribution using the true θ , particularly for tests up to moderate length (say, 50 items or fewer). As a result, the empirical Type I error was smaller than the nominal Type I error. Molenaar and Hoijsink [11] used the statistic $M = -\sum_{g=1}^k \delta_g X_g$ and proposed three approximations to the distribution of M under the Rasch model.

Snijders [18] derived the asymptotic **sampling distribution** for a group of person-fit statistics that

have the form given in (4), and for which the maximum likelihood estimate $\hat{\theta}$ was used instead of θ . It can easily be shown that $l - E(l)$ can be written in the form of (5) choosing

$$w_g(\theta) = \ln \left[\frac{P_g(\theta)}{1 - P_g(\theta)} \right] \quad (9)$$

Snijders [18] derived expressions for the first two moments of the distribution: $E[V(\hat{\theta})]$ and $\text{var}[V(\hat{\theta})]$, and performed a simulation study using maximum likelihood estimation for θ . The results showed that the approximation was satisfactory for $\alpha = 0.05$ and $\alpha = 0.10$, but that the empirical Type I error was higher than the nominal Type I error for smaller values of α .

Optimal Person-fit Statistics

Levine and Drasgow [6] proposed a likelihood ratio statistic, which provided the most powerful test for the null hypothesis that an item score pattern is normal versus the alternative hypothesis that it is aberrant. The researcher in advance has to specify a model for normal behavior (e.g., the 1-, 2-, or 3-PLM) and a model that specifies a particular type of aberrant behavior (e.g., a model in which violations of local independence are specified).

Klauer [5] investigated aberrant item score patterns by testing a null model of normal response behavior (Rasch model) against an alternative model of aberrant response behavior. Writing the Rasch model as a member of the exponential family,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \mu(\theta)h(\mathbf{x}) \exp[\theta R(\mathbf{x})], \quad (10)$$

where

$$\mu(\theta) = \prod_{g=1}^k [1 + \exp(\theta - \delta_g)]^{-1},$$

$$h(\mathbf{x}) = \exp \left(- \sum x_g \delta_g \right), \quad (11)$$

and $R(\mathbf{x}) = \text{number-correct score}$, Klauer [5] modeled aberrant response behavior using the two-parameter exponential family, and introducing an extra person parameter η , as

$$P(\mathbf{X} = \mathbf{x}|\theta, \eta) = \mu(\theta, \eta)h(\mathbf{x}) \exp[\eta T(\mathbf{x}) + \theta R(\mathbf{x})], \quad (12)$$

where $T(\mathbf{x})$ depends on the particular alternative model considered. Using the exponential family of models, a uniformly most powerful test can be used for testing $H_0 : \eta = \eta_0$ against $H_1 : \eta \neq \eta_0$. Let a test be subdivided into two subtests A_1 and A_2 . Then, as an example of η , $\eta = \theta_1 - \theta_2$ was considered, where θ_1 is an individual's ability on subtest A_1 , and θ_2 is an individual's ability on subtest A_2 . Under the Rasch model, it is expected that θ is invariant across subtests and, thus, $H_0 : \eta = 0$ can be tested against $H_1 : \eta \neq 0$. For this type of aberrant behavior, $T(\mathbf{x})$ is the number-correct score on either one of the subtests. Klauer [5] also tested H_0 of equal item discrimination parameters for all persons against person-specific item discrimination and H_0 of local independence against violations against local independence. Results showed that the **power** of these tests depended on the type and the severeness of the violations. Violations against noninvariant ability ($H_0 : \eta = 0$) were found to be the most difficult to detect.

What is interesting in both [5] and [6] is that model violations are specified in advance and that tests are proposed to investigate these model violations. This is different from the approach followed in most person-fit studies where the alternative hypothesis simply says that the null hypothesis is not true. An obvious problem is, which alternative models to specify? A possibility is to specify a number of plausible alternative models and then successively test model-conform item score patterns against these alternative models. Another option is to first investigate which model violations are most detrimental to the use of the test envisaged, and then test against the most serious violations.

The Person-response Function

Trabin and Weiss [20] proposed to use the person response function (PRF) to identify aberrant item score patterns. At a fixed θ value, the PRF specifies the probability of a correct response as a function of the item location δ . In IRT, the IRF often is assumed to be a nondecreasing function of θ , whereas the PRF is assumed to be a *nonincreasing* function of δ . To construct an observed PRF, Trabin and Weiss [20] ordered items to increasing $\hat{\delta}$ values and then formed subtests of items by grouping items according to $\hat{\delta}$ values. For fixed $\hat{\theta}$, the observed PRF was constructed by determining, in each subtest, the mean probability of a correct response. The expected

PRF was constructed by estimating, according to the 3-PLM, in each subtest, the mean probability of a correct response. A large difference between the expected and observed PRFs was interpreted as an indication of nonfitting responses for that examinee. Sijtsma and Meijer ([17]; see also [3]) and Reise [14] further refined person-fit methodology based on the person-response function.

Person-fit Research in Computer Adaptive Testing

Bradlow et al. [1] and van Krimpen-Stoop and Meijer [22] defined person-fit statistics that make use of the property of computer-based testing that a fitting item score pattern will consist of an alternation of correct and incorrect responses, especially at the end of the test when $\hat{\theta}$ comes closer to θ . Therefore, a string of consecutive correct or incorrect answers may be the result of aberrant response behavior. Sums of consecutive negative or positive residuals [$X_g - P_g(\theta)$] can be investigated using a cumulative sum procedure [12].

Usefulness of IRT-based Person-fit Statistics

Several studies have addressed the usefulness of IRT-based person-fit statistics. In most studies, simulated data were used, and in some studies, empirical data [15] were used. The following topics were addressed:

1. detection rate of fit statistics and comparing fit statistics with respect to several criteria such as distributional characteristics and relation to the total score;
2. influence of item, test, and person characteristics on the detection rate;
3. applicability of person-fit statistics to detect particular types of misfitting item score patterns; and
4. relation between misfitting score patterns and the validity of test scores.

On the basis of this research, the following may be concluded. For many person-fit statistics for short tests and tests of moderate length (say, 10–60 items), due to the use of $\hat{\theta}$ rather than θ for most statistics, the nominal Type I error rate is not in agreement with the empirical Type I error rate. In general, sound statistical methods have been derived for the Rasch

model, but because this model is rather restrictive to empirical data, the use of these statistics also is restricted.

Furthermore, it may be wise to first investigate possible threats to the fit of individual item score patterns before using a particular person-fit statistic. For example, if violations against local independence are expected, one of the methods proposed in [5] may be used instead of a general statistic such as the M statistic proposed in [11]. Not only are tests against a specific alternative more powerful than general statistics, also the type of deviance is easier to interpret.

A drawback of some person-fit statistics is that only deviations against the model are tested. This may result in interpretation problems. For example, item score patterns not fitting the Rasch model may be described more appropriately by means of the 3-PLM. If the Rasch model does not fit the data, other explanations are possible. Because in practice it is often difficult, if not impossible, to substantially distinguish different types of item score patterns and/or to obtain additional information using background variables, a more fruitful strategy may be to test against specific alternatives [8].

Almost all statistics are of the form given in (4) but the weights are different. The question then is, which statistic should be used? From the literature, it can be concluded that the use of a statistic depends on what kind of model is used. Using the Rasch model, the theory presented by Molenaar and Hoijtink and their statistic M is a good choice. Statistic M should be preferred over statistics like l_z because the critical values for M are more accurate than those of l_z . With respect to the 2-PLM and the 3-PLM, all statistics proposed suffer from the problem that the standard normal distribution is inaccurate when $\hat{\theta}$ is used instead of θ . This seriously reduces the applicability of these statistics. The theory recently proposed in [18] may help the practitioner to obtain the correct critical values.

Conclusions

The aim of person-fit measurement is to detect item score patterns that are improbable given an IRT model or given the other patterns in a sample. The first requirement, thus, is that person-fit statistics are sensitive to misfitting item score patterns. After

having reviewed the studies using simulated data, it can be concluded that detection rates are highly dependent on (a) the type of aberrant response behavior, (b) the θ value, and (c) the test length. When item score patterns do not fit an IRT model, high detection rates can be obtained in particular for extreme θ s, even when Type I errors are low (e.g., 0.01). The reason is that for extreme θ s, deviations from the expected item score patterns tend to be larger than for moderate θ s. As a result of this pattern misfit, the bias in $\hat{\theta}$ tends to be larger for extreme θ s than for moderate θ s. The general finding that detection rates for moderate θ tend to be lower than for extreme θ s, thus, is not such a bad result and certainly puts the disappointment some authors [13] expressed about low detection rates for moderate θ s in perspective.

Relatively few studies have investigated the usefulness of person-fit statistics for analyzing empirical data. The few studies that exist have found some evidence that groups of persons with *a priori* known characteristics, such as test-takers lacking motivation, may produce deviant item score patterns that are unlikely given the model. However, again, it depends on the degree of aberrance of response behavior how useful person-fit statistics really are. We agree with some authors [15] that more empirical research is needed. Whether person-fit statistics can help the researcher in practice depends on the context in which research takes place.

Smith [16] mentioned four actions that could be taken when an item score pattern is classified as aberrant: (a) Instead of reporting one ability estimate for an examinee, several ability estimates can be reported on the basis of subtests that are in agreement with the model; (b) modify the item score pattern (for example, eliminate the unreached items at the end) and reestimate θ ; (c) do not report the ability estimate and retest a person; or (d) decide that the error is small enough for the impact on the ability to be marginal. This decision can be based on comparing the error introduced by measurement disturbance and the standard error associated with each ability estimate. Which of these actions is taken very much depends on the context in which testing takes place. The usefulness of person-fit statistics, thus, also depends heavily on the application for which it is intended.

References

- [1] Bradlow, E.T., Weiss, R.E. & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive testing, *Journal of the American Statistical Association* **93**, 910–919.
- [2] Drasgow, F., Levine, M.V. & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices, *British Journal of Mathematical and Statistical Psychology* **38**, 67–86.
- [3] Emons, W.H.M., Meijer, R.R. & Sijtsma, K. (2004). Testing hypothesis about the person-response function in person-fit analysis, *Multivariate Behavioral Research* **39**, 1–35.
- [4] Guttman, L. (1950). The basis for scalogram analysis, in *Measurement and Prediction*, S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star & J.A. Claussen, eds, Princeton University Press, Princeton, pp. 60–90.
- [5] Klauer, K.C. (1995). The assessment of person fit, in *Rasch Models, Foundations, Recent Developments, and Applications*, G.H. Fischer & I.W. Molenaar, eds, Springer-Verlag, New York, pp. 97–110.
- [6] Levine, M.V. & Drasgow, F. (1988). Optimal appropriateness measurement, *Psychometrika* **53**, 161–176.
- [7] Levine, M.V. & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores, *Journal of Educational Statistics* **4**, 269–290.
- [8] Meijer, R.R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics, *Psychological Methods* **8**, 72–87.
- [9] Meijer, R.R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: a review and new developments, *Applied Measurement in Education* **8**, 261–272.
- [10] Meijer, R.R. & Sijtsma, K. (2001). Methodology review: evaluating person fit, *Applied Psychological Measurement* **25**, 107–135.
- [11] Molenaar, I.W. & Hoijsink, H. (1990). The many null distributions of person fit indices, *Psychometrika* **55**, 75–106.
- [12] Page, E.S. (1954). Continuous inspection schemes, *Biometrika* **41**, 100–115.
- [13] Reise, S.P. (1995). Scoring method and the detection of person misfit in a personality assessment context, *Applied Psychological Measurement* **19**, 213–229.
- [14] Reise, S.P. (2000). Using multilevel logistic regression to evaluate person fit in IRT models, *Multivariate Behavioral Research* **35**, 543–568.
- [15] Reise, S.P. & Waller, N.G. (1993). Traitendness and the assessment of response pattern scalability, *Journal of Personality and Social Psychology* **65**, 143–151.
- [16] Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators, *Educational and Psychological Measurement* **45**, 433–444.
- [17] Sijtsma, K. & Meijer, R.R. (2001). The person response function as a tool in person-fit research, *Psychometrika* **66**, 191–208.
- [18] Snijders, T.A.B. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter, *Psychometrika* **66**, 331–342.
- [19] Tatsuka, K.K. (1984). Caution indices based on item response theory, *Psychometrika* **49**, 95–110.
- [20] Trabin, T.E. & Weiss, D.J. (1983). The person response curve: fit of individuals to item response theory models, in *New Horizons in Testing*, D.J. Weiss, ed., Academic Press, New York.
- [21] van der Linden, W.J. & Hambleton, R.K., eds (1997). *Handbook of Modern Item Response Theory*, Springer Verlag, New York.
- [22] van Krimpen-Stoop, E.M.L.A. & Meijer, R.R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques, in *Computerized Adaptive Testing: Theory and Practice*, W.J. van der Linden & C.A.W. Glas, eds, Kluwer Academic Publishers, Boston.
- [23] Wright, B.D. & Stone, M.H. (1979). *Best Test Design*, Rasch measurement. Mesa Press, Chicago.

(See also **Hierarchical Item Response Theory Modeling; Maximum Likelihood Item Response Theory Estimation**)

ROB R. MEIJER

Pie Chart

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

Volume 3, pp. 1547–1548

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pie Chart

Pie charts are different from most of the common types of graphs and charts. They do not plot data against axes; instead they plot the data in a circle. They are most suited to displaying frequency data of a single categorical variable. Each slice of the pie represents a different category, and the size of the slice represents the proportion of values in the sample within that category. Figure 1 shows an example from a class of undergraduate students who had to think aloud about when they would complete an essay. Their thoughts were categorized, and the resulting frequencies are shown in the figure. Each slice of the pie is labeled.

As with all graphs, it is important to include all the information that is necessary to understand the data being presented. For Figure 1, the percentage and/or number of cases could be included. In some circumstances, it may be beneficial to have more than one pie chart to represent the data. The diameter of each chart can be varied to provide an extra dimension of information. For instance, if we had a measure of the number of days it took students to complete the essay we could display this in separate pie charts where the diameter of the chart represented completion time.

Most commentators argue that pie charts are inferior for displaying data compared with other

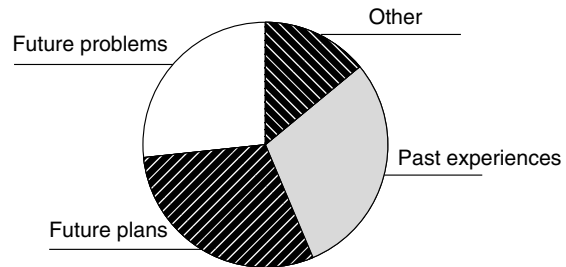


Figure 1 Distribution of students' reported thoughts when estimating the completion time of an academic assignment

methods such as tables and **bar charts** [1]. This is because people have difficulty comparing the size of different angles and therefore difficulty knowing the relative magnitudes of the different slices. Pie charts are common in management and business documents but not in scientific documents. It is usually preferable to use a table or a bar chart.

Reference

- [1] Cleveland, W.S. & McGill, R. (1987). Graphical perception: the visual decoding of quantitative material when on graphical displays of data, *Journal of the Royal Statistical Society. Series A* **150**, 192–229.

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

Pitman Test

CLIFFORD E. LUNNEBORG

Volume 3, pp. 1548–1550

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Pitman Test

The Pitman test [7] is sometimes identified as the Fisher–Pitman test [10] in the continuing but suspect belief that the Tasmanian mathematician **E. J. G. Pitman** (1897–1993) merely extended or modified an idea of **R. A. Fisher’s** [3]. While Fisher may have foreseen the population-sampling version of the test, the far more useful randomization-based version clearly is Pitman’s work [2].

In what follows, I adopt the practice of Edgington [2] in distinguishing between a permutation test (*see* **Permutation Based Inference**) and a **randomization test**. A permutation test requires random samples from one or more populations, each population of sufficient size that the sample observations can be regarded as independently and identically distributed (i.i.d.). A randomization test requires only that the members of a nonrandom sample be randomized among treatment groups. The Pitman test comes in both flavors.

The Two-sample Pitman Permutation Test

Consider the results of drawing random samples, each of size four, from the (large) fifth-grade school populations in Cedar Grove and Forest Park. For the students sampled from the Cedar Grove population, we record arithmetic test scores of (25, 30, 27, 15), and for those sampled from Forest Park (28, 30, 32, 18).

The null hypothesis of the two-sample test is that the two population distributions of arithmetic scores are identical. Thus, the Pitman test is a distribution-free alternative (*see* **Distribution-free Inference, an Overview**) to the two-sample t Test. The null hypothesis for the latter test is that the two distributions are not only identical but follow a normal law.

The logic of the two-sample Pitman test is this. Under the null hypothesis, any four of the eight observed values could have been drawn from the Cedar Grove population and the remaining four from Forest Park. That is, we are as likely to have obtained, say, (25, 27, 28, 32) from Cedar Grove and (30, 15, 30, 18) from Forest Park as the values actually sampled. In fact, under the null hypothesis, all possible permutations of the eight observed scores,

four from Cedar Grove and four from Forest Park, are equally likely. There are $M = 8!/(4!4!) = 70$ permutations, and our random sampling gave us one of these equally likely outcomes.

By comparing the test statistic computed for the observed random samples with the 69 other possible values, we can decide whether our outcome is consonant with the null hypothesis or not.

What test statistic we compute and how we compare its value against this permutation reference distribution depends, of course, on our scientific hypothesis. The most common use of the Pitman two-sample test is as a location test, to test for a difference in population means. The test may be one-tailed or two-tailed, depending on whether our alternative hypothesis specifies a direction to the difference in population means. A natural choice of test statistic would be the difference in sample means.

For our data, we have as the difference in sample means, Forest Park minus Cedar Grove, $s = [(28 + 30 + 32 + 18)/4] - [(25 + 30 + 27 + 15)/4] = 27.0 - 24.25 = 2.75$. Is this a large difference, relative to what we would expect under the null hypothesis?

Assume our alternative hypothesis is nondirectional (*see* **Directed Alternatives in Testing**). We want to know the probability, under the null hypothesis, of observing a mean difference of $+2.75$ or greater, or a mean difference of -2.75 or smaller. How many of the 70 possible permutations of the data produce mean differences in these two tails?

In fact, 36 of the 70 permutations of the data yield mean differences with absolute values of 2.75 or greater. So the resulting P value is $36/70 = 0.514$. There is no evidence in these sparse data that the null hypothesis might not be correct.

For samples of equal size, the permutation test reference distribution is symmetric. The two-tailed P value is double the one-tailed P value.

The Pitman test is exact (*see* **Exact Methods for Categorical Data**) in the sense that if we reject the null hypothesis whenever the obtained P value is no greater than α , we are assured that the probability of a Type I error will not exceed α . For small samples, however, the test is conservative. The probability of a Type I error is smaller than α . For our example, if we set $\alpha = 0.05$, we would reject the null hypothesis if our observed statistic was either the smallest or the largest value in the reference distribution, with

an obtained P value of $2/70 = 0.0286$. We could not reject the null hypothesis if our observed statistic was one of the two largest or two smallest, for, then the obtained P value would be $4/70 = 0.0571$, greater than α .

Software for the Pitman test is available in statistical programs such as StatXact (www.cytel.com), SC (www.mole-soft.demon.co.uk), and R (www.R-project.org) (as part of an add-on package for R) (see **Software for Statistical Analyses**).

The Pitman permutation test is discussed further in [1], [4], [9], [10]. In some references, the test is known as the raw scores permutation test to distinguish it from more recently developed permutation tests requiring transformation of the observations to ranks or to normal scores.

The Two-group Pitman Randomization Test

In view of the scarcity of random samples, the population sampling permutation test has limited utility to behavioral scientists. Typically, we must work with available cases: patient volunteers, students enrolled in Psychology 101, the animals housed in the local vivarium. By randomizing these available cases among treatments, however, we not only practice good experimental design, we also enable statistical inference that does not depend upon population sampling.

To illustrate the Pitman test in the randomization-inference (see **Randomization**) context, let us reattribute the arithmetic test scores used in the earlier section. Assume now we have eight fifth-grade students, not a random sample from any population. We randomly divide the eight into two groups of four. Students in the first of these treatment groups are given a practice exam one week in advance of the arithmetic test. Those in the second group receive this practice exam the afternoon before the test. Here, again, are the arithmetic test scores: for the Week-ahead treatment (28, 30, 32, 18), and for the Day-before treatment (25, 30, 27, 15).

Is there an advantage to the Week-ahead treatment? The observed mean difference, Week-ahead minus Day-before, is $+2.75$. Is that worthy of note?

The randomization test null hypothesis is that any one of the students in this study would have earned exactly the same score on the test if he or she had

been randomized to the other treatment. For example, Larry was randomized to the Week-ahead treatment and earned a test score of 28. The null hypothesis is that he would have earned a 28 if he had been randomized to the Day-before treatment also.

There are, again, 70 different ways in which the eight students could have been randomized four apiece to the two treatments. And, under the null hypothesis, we can compute the mean difference for each of those randomizations.

The alternative hypothesis is that, across these eight students, there is a tendency for the test score of a student to be higher with Week-ahead availability of the practice exam than with Day-before availability.

The mechanics of the randomization test are identical to those of the permutation test, and we would find that the probability of a mean difference of $+2.75$ or greater under the null hypothesis, the P value for our one-tailed test, to be $18/70 = 0.2571$, providing no consequential evidence against the null hypothesis.

The randomization test of Pitman is discussed more fully in [2], [5], [6]. The software for the permutation test can be used, of course, for the randomization test as well.

The permutation test requires random samples from well-defined populations. In the rare event that we do have random samples, we can extend our statistical inferences to those populations sampled. The statistical inferences based on randomization are limited to the set of cases involved in the randomization. While this may seem a severe limitation, randomization of cases among treatments is a gold standard for causal inference, linking differential response as unambiguously as possible to differential treatment [8]. Much research is intended to demonstrate just such a linkage.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [2] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [3] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [4] Good, P. (2000). *Permutation Tests*, 2nd Edition, Springer-Verlag, New York.
- [5] Lunneborg, C.E. (2000). *Data Analysis by Resampling*, Duxbury, Pacific Grove.

- [6] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.
- [7] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, *Royal Statistical Society Supplement* **4**, 119–130.
- [8] Rubin, D.B. (1991). Practical applications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213–1234.
- [9] Sprent, P. (1998). *Data Driven Statistical Methods*, Chapman & Hall, London.
- [10] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Edition, Chapman & Hall/CRC, Boca Raton.

CLIFFORD E. LUNNEBORG

Placebo Effect

PATRICK ONGHENA

Volume 3, pp. 1550–1552

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Placebo Effect

The placebo effect is narrowly defined as the beneficial effect associated with the use of medication that has no intrinsic pharmacological effect (e.g., sugar pills or saline injections). More generally, it refers to the beneficial effect associated with all kinds of interventions that do not include the presumed active ingredients (e.g., sham procedures, mock surgery, attention placebo, or pseudo therapeutic encounters and symbols, such as a white coat) [5, 14, 15]. If the effect is harmful, then it is called a *nocebo effect* [2]. Both the placebo and the nocebo effects can be operationally defined as the difference between the results of a placebo/nocebo condition and a no-treatment condition in a randomized trial [6].

The use of placebos is commonplace in clinical testing in which the randomized double-blind placebo-controlled trial represents the gold standard of clinical evidence (*see Clinical Trials and Intervention Studies*). Ironically, this widespread use of placebos in controlled clinical trials has led to more confusion than understanding of the placebo effect because those trials usually do not include a no-treatment condition. The two most popular fallacies regarding the interpretation of placebo effects in controlled clinical trials are (a) that all improvement in the placebo condition is due to the placebo effect, and (b) that the placebo effect is additive to the treatment effect.

The fallacious nature of the first statement becomes evident if one takes other possible explanations of improvement into account: spontaneous recovery, the natural history of an illness, patient, or investigator bias, and **regression artifacts**; to name the most important ones [4, 6, 10]. In the absence of a no-treatment condition, both clinicians and researchers have a tendency to mistakenly attribute all improvement to the administration of the placebo itself. However, comparison of an active therapy with an inactive therapy offers control for the active therapy effect but does not separate the placebo effect from other confounding effects; hence the need for a no-treatment control condition to assess and fully understand the placebo effect (see operational definition above).

Also, the idea of an additive placebo effect, which can be subtracted from the effect in the treatment condition to arrive at a true treatment effect,

is misguided. There is ample evidence that most often the placebo effect interacts with the treatment effect, operating recursively, and synergistically in the course of the treatment [4, 11].

Because of these methodological issues, and because of the different ways in which a placebo can be defined (narrow or more comprehensive), it should not come as a surprise that the size of the placebo effect and the conditions for its appearance remain controversial. A frequently cited figure, derived from an influential classical paper by Beecher [3], is that about one-third of all patients in clinical trials improve due to the placebo effect. However, in a recent systematic review of randomized trials comparing a placebo condition with a no-treatment condition, little evidence was found that placebos have clinically significant effects, except perhaps for small effects in the treatment of pain [7, 8]. Recently, functional magnetic resonance imaging experiments have convincingly shown that the human brain processes pain differently when participants receive or anticipate a placebo [18], but the clinical importance of placebo effects might be smaller than, and not as general as, once thought.

Parallel to this discussion on **effect sizes**, there is also an interesting debate regarding the explanation and possible working mechanisms of the placebo effect [9, 17]. The two leading models are the classical conditioning model and the response expectancy model.

According to a classical conditioning scheme, the pharmacological properties of the medication (or the active ingredients of the intervention) act as an unconditioned stimulus (US) that elicits an unconditioned biological or behavioral response (UR). By frequent pairings of this US with a neutral stimulus (NS), like a tablet, a capsule, or an injection needle, this NS becomes intrinsically associated with the unconditioned stimulus (US), and becomes what is called a *conditioned stimulus* (CS). This means that, if the CS (e.g., a tablet) is presented in the absence of the US (e.g., the active ingredients), this CS also elicits the same biological or behavioral response, now called the *conditioned response* (CR). In other words: the CS acts as a placebo [1, 16].

However, several authors have remarked that there are problems with this scheme [4, 9, 13]. Although classical conditioning as an explanatory mechanism seems plausible for some placebo effects, there appears to exist a variety of other placebo

effects for which conditioning can only be part of the story or in which other mechanisms are, at least partly, responsible. For instance, placebo effects are reported in the absence of previous exposure to the active ingredients of the intervention or after a single encounter, and medication placebos can sometimes produce effects opposite to the effects of the active drug [16]. These phenomena suggest that anticipation and other cognitive factors play a crucial role in the placebo response. This is in line with the response expectancy model, which states that bodily changes occur to the extent that the subject expects them to. According to this model, expectancies act as a mediator between the conditioning trials and the placebo effect, and, therefore, verbal information can strengthen or weaken a given placebo effect [9, 13, 17].

Whatever the possible working mechanisms or the presumed size of the placebo effect, from a methodological perspective, it is recommended to distinguish the placebo effect from other validity threats, like history, experimenter bias, and regression, and to restrict its use to a particular type of **reactivity** effect (or artifact, depending on the context) (see **Quasi-experimental Designs; External Validity; Internal Validity**). From this perspective, the placebo effect is considered as a specific threat to the treatment construct validity because it arises when participants or patients interpret the treatment setting in a way that makes the actual intervention different from the intervention as planned by the researcher. In fact, the placebo effect, and especially the response expectancy model, reminds us that it is hard to assess the effects of external agents without taking into account the way in which these external agents are interpreted by the participants. To emphasize this subjective factor, Moerman [11, 12] proposed to rename (and reframe) the placebo effect as a ‘meaning response’, which he defined as ‘the physiological or psychological effects of meaning in the treatment of illness’ [11, p. 14].

References

- [1] Ader, R. (1997). Processes underlying placebo effects: the preeminence of conditioning, *Pain Forum* **6**, 56–58.
- [2] Barskey, A.J., Saintfort, R., Rogerts, M.P. & Borus, J.F. (2002). Nonspecific medication side effects and the nocebo phenomenon, *Journal of the American Medical Association* **287**, 622–627.
- [3] Beecher, H.K. (1955). The powerful placebo, *Journal of the American Medical Association* **159**, 1602–1606.
- [4] Engel, L.W., Guess, H.A., Kleinman, A. & Kusek, J.W. eds (2002). *The Science of the Placebo: Toward an Interdisciplinary Research Agenda*, BMJ Books, London.
- [5] Harrington, A. ed. (1999). *The Placebo Effect: An Interdisciplinary Exploration*, Harvard University Press, Cambridge.
- [6] Hróbjartsson, A. (2002). What are the main methodological problems in the estimation of placebo effects? *Journal of Clinical Epidemiology* **55**, 430–435.
- [7] Hróbjartsson, A. & Gøtzsche, P.C. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment, *New England Journal of Medicine* **344**, 1594–1602.
- [8] Hróbjartsson, A. & Gøtzsche, P.C. (2004). Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment, *New England Journal of Medicine* **344**, 1594–1602.
- [9] Kirsch, I. (2004). Conditioning, expectancy, and the placebo effect: comment on Stewart-Williams and Podd (2004), *Psychological Bulletin* **130**, 341–343.
- [10] McDonald, C.J., Mazzuca, S.A. & McCabe, G.P. (1983). How much of the placebo ‘effect’ is really statistical regression? *Statistics in Medicine* **2**, 417–427.
- [11] Moerman, D.E. (2002). *Medicine, Meaning, and the “Placebo Effect”*, Cambridge University Press, London.
- [12] Moerman, D.E. & Jonas, W.B. (2002). Deconstructing the placebo effect and finding the meaning response, *Annals of Internal Medicine* **136**, 471–476.
- [13] Montgomery, G.H. & Kirsch, I. (1997). Classical conditioning and the placebo effect, *Pain* **72**, 107–113.
- [14] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
- [15] Shapiro, A. & Shapiro, E. (1997). *The Powerful Placebo: From Ancient Priest to Modern Physician*, Johns Hopkins University Press, Baltimore.
- [16] Siegel, S. (2002). Explanatory mechanisms of the placebo effect: Pavlovian conditioning, in *The Science of the Placebo: Toward an Interdisciplinary Research Agenda*, L.W. Engel, H.A. Guess, A. Kleinman & J.W. Kusek, eds, BMJ Books, London, pp. 133–157.
- [17] Stewart-Williams, S. & Podd, J. (2004). The placebo effect: dissolving the expectancy versus conditioning debate, *Psychological Bulletin* **130**, 324–340.
- [18] Wager, T.D., Rilling, J.K., Smith, E.E., Sokolik, A., Casey, K.L., Davidson, R.J., Kosslyn, S.M., Rose, R.M. & Cohen, J.D. (2004). Placebo-induced changes in fMRI in the anticipation and experience of pain, *Science* **303**, 1162–1167.

PATRICK ONGHENA

Point Biserial Correlation

DIANA KORNBROT

Volume 3, pp. 1552–1553

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Point Biserial Correlation

The point biserial correlation, r_{pb} , is the value of **Pearson's product moment correlation** when one of the variables is dichotomous, taking on only two possible values coded 0 and 1, and the other variable is metric (interval or ratio). For example, the dichotomous variable might be political party, with left coded 0 and right coded 1, and the metric variable might be income. The metric variable should be approximately normally distributed for both groups.

Calculation

The value of r_{pb} can be calculated directly from (1) [2].

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{s}_y} \sqrt{\frac{N_1 N_0}{N(N-1)}}, \quad (1)$$

where $\bar{Y}_0$ and $\bar{Y}_1$ are means of the metric observations coded 0 and 1 respectively; N_0 and N_1 are number of observations coded 0 and 1 respectively; N is total number of observations, $N_0 + N_1$; and $\bar{s}_y$ is standard deviation of all the metric observations

$$\bar{s}_y = \frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N-1}. \quad (2)$$

Equation (1) is generated by using the standard equation for the Pearson's product moment correlation, r , with one of the dichotomous variables coded 0 and the other coded 1. Consequently, r_{pb} can easily be obtained from standard statistical packages as the value or Pearson's r when one of the variables only takes on values of 0 or 1.

Interpretation of r_{pb} as an Effect Size

The point biserial correlation, r_{pb} , may be interpreted as an **effect size** for the difference in means between two groups. In fact, r_{pb}^2 is the proportion of variance accounted for by the difference between the means of the two groups. The Cohen's d effect size (defined as the difference between means divided by the pooled

standard deviation and also known as Glass's g (*see Effect Size Measures*)), can be obtained from r_{pb} via (3) [1]

$$d = \sqrt{\frac{N(N-2)r_{pb}^2}{N_1 N_0 (1 - r_{pb}^2)}} \quad (3)$$

Hypothesis Testing and Power for r_{pb}

A value of r_{pb} that is significantly different from zero is completely equivalent to a significant difference in means between the two groups. Thus, an independent groups t Test with $(N-2)$ degrees of freedom, df , may be used to test whether r_{pb} is nonzero (or equivalently, a one-way **analysis of variance** (ANOVA) with two levels). The relation between the t -statistic for comparing two independent groups and r_{pb} is given by (4) [1, 2]

$$t = \sqrt{N-2} \frac{r_{pb}}{\sqrt{1 - r_{pb}^2}} \quad (4)$$

Power to detect a nonzero r_{pb} will also be the same as that for an independent groups t Test. For example, a total sample size, $N = 52$, is needed for a power of 0.80 (80% probability of detection) for a 'large' effect size, corresponding to $r_{pb} = 0.38$. Similarly, a total sample size, $N = 52$, is needed for a power of 0.80 for a 'medium' effect size, corresponding to $r_{pb} = 0.24$.

Summary

The point biserial correlation is a useful measure of effect size, that is, statistical magnitude, of the difference in means between two groups. It is based on Pearson's product moment correlation.

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.
- [2] Sheskin, D.J. (2000). *Handbook of Parametric and Non-parametric Statistical Procedures*, 2nd Edition, Chapman & Hall, London.

DIANA KORNROT

Polychoric Correlation

SCOTT L. HERSHBERGER

Volume 3, pp. 1554–1555

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Polychoric Correlation

Introduction

The *polychoric correlation* is used to correlate two ordinal variables, X and Y , when it is assumed that underlying each is a continuous, normally distributed latent variable [2, 5]. At least one of the ordinal variables has more than two categories. When both variables are ordinal, but the assumption of a latent, normally distributed variable is *not* made, we obtain **Spearman's rank order correlation (Spearman's rho)** [4]. The **tetrachoric correlation** is a specific form of the polychoric correlation when the polychoric correlation is computed between two binary variables [1]. The *polyserial correlation* is another variant of the polychoric correlation, in which one of the variables is ordinal and the other is continuous [2]. Polychoric correlations between two ordinal variables rest on the interpretation of the ordinal variables as discrete versions of latent continuous 'ideally measured' variables, and the assumption that these corresponding latent variables are bivariate-normally distributed (*see Catalogue of Probability Density Functions*). In a bivariate-normal distribution, the distribution of Y is normal at fixed values of X , and the distribution of Y is normal at fixed values of Y . Two variables that are bivariate-normally distributed must each be normally distributed. The calculation of the polychoric correlation involves corrections that approximate what the **Pearson product-moment correlation** would have been if the data had been continuous.

Statistical Definition

Estimating the polychoric correlation assumes that an observed ordinal variable (Y) with C categories is related to an underlying continuous latent variable (Y^*) having $C - 1$ thresholds (τ_c):

$$Y = c, \tau_c < Y^* \leq \tau_{c+1}, \quad (1)$$

in which $\tau_0 = -\infty$ and $\tau_C = +\infty$. Since Y^* is assumed to be normally distributed, the probability that $Y^* \leq \tau_c$ is

$$p(Y^* \leq \tau_c) = \Phi\left(\frac{\tau_c - \mu}{\sigma}\right), \quad (2)$$

where Φ is the cumulative normal probability density function, and Y^* is a **latent variable** with mean μ and standard deviation σ . Thus, $p(Y^* \leq \tau_c)$ is the cumulative probability that latent variable Y^* is less than or equal to the c th threshold τ_c .

It also follows that

$$p(Y^* > \tau_c) = 1 - \Phi\left(\frac{\tau_c - \mu}{\sigma}\right). \quad (3)$$

If the observed variable Y had only two categories, defining the two probabilities $p(Y^* \leq \tau_c)$ and $p(Y^* > \tau_c)$ would suffice: only a single threshold would be required. However, when the number of categories is greater than two, more than one threshold is required. In general then, the probability that the observed ordinal variable Y is equal to a given category c is

$$p(Y = c) = p(Y^* \leq \tau_{c+1}) - p(Y^* \leq \tau_c). \quad (4)$$

For example, if the ordinal variable has three categories, two thresholds must be defined:

$$\begin{aligned} \sum_{c=1}^3 p(Y_c) &= p(Y = c_1) + p(Y = c_2) + p(Y = c_3) \\ &= p(Y_1^* \leq \tau_1) + p(\tau_1 < Y_2^* \leq \tau_2) + p(\tau_2 < Y_3^*) \\ &= \Phi\left(\frac{\tau_1 - \mu}{\sigma}\right) + \left\{ \Phi\left(\frac{\tau_2 - \mu}{\sigma}\right) \right. \\ &\quad \left. - \Phi\left(\frac{\tau_1 - \mu}{\sigma}\right) \right\} + \Phi\left(\frac{\tau_2 - \mu}{\sigma}\right). \end{aligned} \quad (5)$$

The thresholds define the cutpoints or transitions from one category of an ordinal variable to another on a normal distribution. For a given μ and σ of Y^* and the two thresholds τ_1 and τ_2 , the probability that latent variable Y^* is in category 1 is the area under the normal curve $-\infty$ and τ_1 ; the probability that Y^* is in category two is in the area bounded by τ_1 and τ_2 ; and the probability that Y^* is in category three is the area between τ_2 and $+\infty$. The polychoric correlation may also be thought of as the correlation between the normal scores Y_i^* , Y_j^* as estimated from the category scores Y_i , Y_j . These normal scores are based on the threshold values. In essence, the calculation of the polychoric correlation involves correcting errors arising from category group errors, and approximates what the Pearson product-moment correlation would have been if the data had been continuous.

2 Polychoric Correlation

The form of the distribution of the latent Y^* variables – bivariate normality (*see Catalogue of Probability Density Functions*) – has to be assumed so as to specify the likelihood function of the polychoric correlation. For example, the program PRELIS 2 [3] estimates the polychoric correlation by limited information maximum likelihood (*see Maximum Likelihood Estimation*). In PRELIS 2, the polychoric correlation is estimated in two steps. First, the thresholds are estimated from the cumulative frequencies of observed Y category scores and the inverse of the standard normal probability density function, assuming $\mu = 0$ and $\sigma = 1$. Second, the polychoric correlation is then estimated by restricted maximum likelihood conditional on the threshold values.

Example The table below, (Table 1) adapted from Drasgow [2], summarizes the number of lambs born to 227 ewes over two years:

Spearman's rho between the two years is .29, while the Pearson product-moment correlation is .32. We used two methods based on maximum likelihood estimation to calculate the polychoric correlation. The first method, joint maximum likelihood, estimates the polychoric correlation and the thresholds at the same time. The second method, limited information maximum likelihood estimation, first estimates the thresholds from the observed category frequencies, and then estimates the polychoric correlation conditional on these thresholds.

Table 1 Number of Lambs Born to 227 Ewes in 1951 and 1952. Please also note that the row 1952 should be 1951.

Lambs born in 1952	Lambs born in 1952			
	None	1	2	Total
None	58	52	1	111
1	26	58	3	87
2	8	12	9	29
Total	92	122	13	227

The parameter estimates from joint maximum likelihood estimation are

$$\begin{aligned}
 r &= 0.419 \\
 \tau_{1,1952} &= -.242 \\
 \tau_{2,1952} &= 1.594 \\
 \tau_{1,1953} &= -.030 \\
 \tau_{2,1953} &= 1.133.
 \end{aligned} \tag{6}$$

The parameter estimates from limited information maximum likelihood estimation are

$$\begin{aligned}
 r &= 0.420 \\
 \tau_{1,1952} &= -.240 \\
 \tau_{2,1952} &= 1.578 \\
 \tau_{1,1953} &= -.028 \\
 \tau_{2,1953} &= 1.137.
 \end{aligned} \tag{7}$$

Thus, the results using either maximum likelihood approach are nearly identical.

References

- [1] Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*, Wadsworth, Monterey.
- [2] Drasgow, F. (1988). Polychoric and polyserial correlations, in *Encyclopedia of the Statistical Sciences*, Vol. 7, L. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 69–74.
- [3] Jöreskog, K.G. & Sörbom, D. (1996). *PRELIS user's manual, version 2*, Scientific Software International, Chicago.
- [4] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, McGraw-Hill, New York.
- [5] Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient, *Psychometrika*, **44**, 443–460.

(See also **Structural Equation Modeling: Software**)

SCOTT L. HERSHBERGER

Polynomial Model

L. JANE GOLDSMITH AND VANCE W. BERGER

Volume 3, pp. 1555–1557

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Polynomial Model

A polynomial model is a form of linear model in which the predictor variables (the x 's) appear as powers or cross-products in the model equation. The highest power (or sum of powers for x 's appearing in a cross-product term) is called the *order* of the model. The simplest polynomial model is the single predictor model of order 1:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

where Y is the dependent variable, x is the predictor variable, and ε is an error term. The errors associated with each observation of Y are assumed to be independent and identically distributed according to the Gaussian law with mean 0 and unknown variance σ^2 ; that is, according to commonly used notation, $\varepsilon \sim N(0, \sigma^2)$. Thus, a first-order polynomial model with one predictor is recognizable as the simple linear regression model. See [1] and [2]. With k predictors, a first-order polynomial model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad (2)$$

the **multiple linear regression** model. Of course, parameter estimates can be obtained and hypotheses can be tested using linear model least-squares theory (see **Least Squares Estimation**). A model of order two with one predictor variable is

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \quad (3)$$

This is the well-known formula for a parabola that opens up if $\beta_2 > 0$ and opens down if $\beta_2 < 0$. Figure 1 demonstrates a second-order polynomial model $\beta_2 < 0$. The line represents the underlying second-order polynomial and the dots represent observations made according to the $\varepsilon \sim N(0, \sigma^2)$ law, with $\sigma^2 = 0.01$.

We see that, according to this model, Y increases to a maximum value at $x = 1$ and then decreases. A curve like this can be used to model blood concentrations of a drug component from the time of administration ($x = 0$), for example. Concentration of this chemical increases as the medicine is metabolized, then falls off through excretion from the body. At $x = 3$, blood concentration is expected to be 0. It may seem odd to refer to a second-order polynomial, or even a higher-order polynomial, as a *linear*

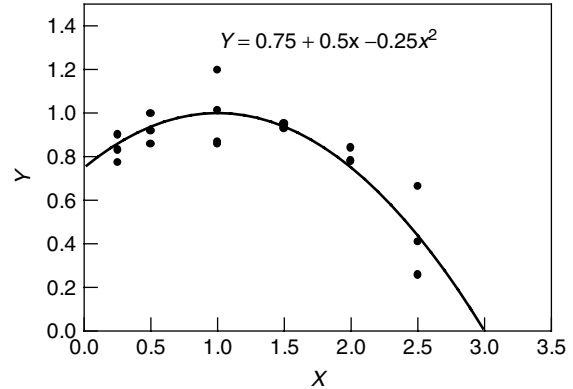


Figure 1 An example of a polynomial curve

model, but the model is linear in the coefficients to be estimated, even if it is not linear in the predictor x . Of course, given the predictor x , one can create another predictor, x^2 , which can be viewed not as the square of the first predictor but rather as a separate predictor. It could even be given a different name, such as w . Now the linear model is, in fact, linear in the predictors as well as in the coefficients. Either way, standard linear regression methods apply. Some authors refer to polynomial models as *curvilinear models*, thus acknowledging the curvature and the linearity that apply. A second-order model with k predictors is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \cdots + \beta_{kk} x_k^2 + \beta_{12} x_1 x_2 + \cdots + \beta_{ij} x_i x_j + \cdots + \beta_{k-1k} x_{k-1} x_k + \varepsilon \quad (4)$$

The notation convention in this second-order multivariate model is that β_i is the coefficient of x_i and β_{ij} is the coefficient of $x_i x_j$. Thus, in general, β_{jj} is the coefficient of x_j^2 . A second-order polynomial model with k predictor variables will contain $(k+1)(k+2)/2$ terms if all terms are retained. Often a simple model (say, one without cross-product terms) is desired, but terms should not be omitted without careful analysis of scientific and statistical evidence. The second-order polynomial model also admits the use of linear model inferential techniques for parameter estimation and hypothesis testing; that is, the β_i 's and β_{ij} 's can be estimated and tested using the usual least-squares regression theory. Model fit

2 Polynomial Model

can be evaluated through residual analysis by plotting residuals versus the $\hat{Y}_i$'s (the predicted values).

By extending the notation convention given above for second-order multivariate polynomial models in a completely natural way, third- and higher-order models can be specified. Third-order models have some use in research: cubic equations can model a dependent variable that rises and falls and rises again (or, with a change of sign, falls, then rises, then falls again). Higher-than-third-order models have rarely been used, however, as their complicated structures have not been found useful in modeling natural phenomena. All polynomial models have the advantage that inference and analysis methods of linear least-squares regression can be applied if the assumptions

of independent observations and normality of error terms are met.

References

- [1] Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition, John Wiley & Sons, New York.
- [2] Ratkowsky, D.A. (1990). *Handbook of Nonlinear Regression Models*, Marcel Dekker, New York.

L. JANE GOLDSMITH AND VANCE W. BERGER

Population Stratification

EDWIN J.C.G. VAN DEN OORD

Volume 3, pp. 1557–1557

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Population Stratification

A commonly used method to detect disease mutations is to test whether **genotype** or marker allele frequencies differ between groups of cases and controls. These frequencies will differ if the marker allele is the disease mutation or if it is very closely located to the disease mutation on the chromosome. A problem is that case-control status and marker allele frequencies can also be associated because of population stratification. Population stratification refers to subgroups in the sample, for example, ethnic groups, that differ from each other with respect to the marker allele frequency as well as the disease prevalence. For example, assume that in population 1 the prevalence of disease X is higher than in population 2. Because of different demographic and population genetic histories, the frequency of marker allele A could also be higher in population 1 than in population 2. If the underlying population structure would not be taken into account, an association will be observed in the total pooled sample. That is, the cases are more likely to come from population 1 because it has a higher disease prevalence, and the cases are also more likely to have allele A because it has a higher frequency in population 1. Allele A is, however, not a disease mutation nor does it lie close to a disease mutation on the chromosome. Because it does not contain information that would help to identify the location of a disease mutation, the association is said to be false or spurious.

If the subgroups in the population can be identified, such spurious findings can be easily avoided

by performing the tests within each subgroup. In the case of unobserved heterogeneity, one approach to avoid false positive findings due to population stratification is to use family-based tests. For example, one could count the number of times heterozygous parents transmit allele A rather than allele a to an affected child. Because all family members belong to the same population strata, a significant result cannot be the result of population stratification. A disadvantage of these family-based tests is that it may be impractical to collect DNA from multiple family members. Another disadvantage is that some families will not be informative (e.g., a homozygous parent will always transmit allele A), resulting in a reduction of sample size. Because of these disadvantages, alternative procedures such as ‘genomic control’ have been proposed that claim to detect and control for population stratification without the need to genotype family members. The basic idea is that if many unlinked markers are examined, the presence of population stratification should cause differences in relatively large subset of the markers.

An important question involves the extent population stratification is a threat for **case-control studies**. The answer to this question seems to fluctuate over time. The fact is that there are currently not many examples of false positives as a result of population stratification where the subgroups could not be identified easily.

(See also **Allelic Association**)

EDWIN J.C.G. VAN DEN OORD

Power

DAVID C. HOWELL

Volume 3, pp. 1558–1564

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Power

Power is traditionally defined as the probability of correctly rejecting a false null hypothesis. As such it is equal to $1 - \beta$, where β is the probability of a Type II error – failing to reject a false null hypothesis. Power is a function of the sample size (N), the effect size (ES) (see **Effect Size Measures**), the probability of a Type I error (α), and the specific **experimental design**.

The purpose of the article is to review and elaborate important issues dealing with statistical power, rather than to become engaged in the technical issues of direct calculation. Additional material on calculation can be found in the entry **power analysis for categorical methods**, and in work by Cohen [5–8].

Following many years in which behavioral scientists paid little attention to power [4, 20], in the last 15 or 20 years there has been an increase in the demand that investigators consider the power of their studies. Funding agencies, institutional review boards, and even some journals expect to see that the investigator made a serious effort to estimate the power of a study before conducting the research. Agencies do not want to fund work that does not have a reasonable chance of arriving at informative conclusions, and review boards do not want to see animals or patients wasted needlessly – either because the study has too little power or because it has far more power than is thought necessary.

It is an open question whether such requirements are having the effect of increasing the true power of experiments. Investigators can increase *reported* power by arbitrarily increasing the effect size they claim to seek until the power calculation for the number of animals they planned to run anyway reaches some adequate level. Senn [21] put it this way: ‘*Clinically* relevant difference: used in the theory of clinical trials as opposed to *cynically* relevant difference which is used in the practice (p. 174, italics added)’.

Four important variables influence the power of an experiment. These are the size of the effect, the sample size, the probability of a Type I error (α), and the experimental design.

Effect Size (ES)

The basic concepts in the definition of power go back to the development of statistical theory by **Neyman**

and **Pearson** [17, 18]. Whereas **Fisher** [11] recognized only a null hypothesis (H_0) that could be retained or rejected, Neyman and Pearson argued for the existence of an alternative hypothesis (H_1) in addition to the null. Effect size is a measure of the degree to which parameters expected under the alternative hypothesis differ from parameters expected under the null hypothesis (see **Effect Size Measures**). Different disciplines, and even different subfields within the behavioral sciences, conceptualize effect sizes differently. In medicine, where the units of measurement are more universal and meaningful (e.g., change in mmHg for blood pressure, and proportion of treated patients showing an improvement), it is sensible to express the size of an effect in terms of raw score units. We can look for a change of 10 mmHg in systolic blood pressure as a clinically meaningful difference. In psychology, where the units of measurement are often measures like the number of items recalled from a list, or the number of seconds a newborn will stare at an image, the units of measurement have no consensual meaning. As a result we often find it difficult to specify the magnitude of the difference between two group means that would be meaningful. For this reason, psychologists often scale the difference by the size of the standard deviation; it is more meaningful to speak of a quarter of a standard deviation change in fixation length than to speak of a 2 sec change in fixation.

As an illustration of standardized effect size consider the situation in which we wish to compare two group means. Our null hypothesis is usually that the corresponding population means are equal, with $\mu_1 - \mu_2 = 0$. Under the alternative hypothesis, for a two-tailed test, $\mu_1 - \mu_2 \neq 0$. Thus, our measure of effect size should reflect the anticipated difference in the population means under H_1 . However, the size of the anticipated mean difference needs to be considered in relation to the population variance, and so we will define $ES = (\mu_1 - \mu_2)/\sigma$, where σ is taken as the (common) size of the population standard deviation. Thus, the effect size is scaled by, or standardized by, the population standard deviation. The effect size measures considered in this article are generally standardized. Thus, they are scale free, and do not depend on the original units of measurement.

Small, Medium, and Large Effects

Cohen [4] initially put forth several rules of thumb concerning what he considered small, medium, and

large values of ES . These values vary with the statistical procedure. For example, when comparing the means of two independent groups, Cohen defined ‘small’ as $ES = 0.20$, ‘medium’ as $ES = 0.50$, and ‘large’ as $ES = 0.80$.

Lenth [14] challenged Cohen’s provision of small, medium, and large estimates of effect size, as well as his choice of standardized measures. Lenth argued persuasively that the experimenter needs to think very carefully about both the numerator and denominator for ES , and should not simply look at their ratio. Cohen would have been one of the first to side with Lenth’s emphasis on the importance of thinking clearly about the difference to be expected in population means, the value of the population correlation, and so on. Cohen put forth those definitions tentatively in the beginning, though he did give them more weight later. The calculation of ES is serious business, and should not be passed over lightly. The issue of standardization is more an issue of reporting effects than of power estimation, because all approaches to power involve standardizing measures at some stage.

Sample Size (N)

After the effect size, the most important variable controlling power is the size of our sample(s). It is important both because it has such a strong influence on the power of a test, and because it is the variable most easily manipulated by the investigator. It is difficult, though not always impossible, to manipulate the size of your effect or the degree of variability with the groups, but it is relatively easy to alter the sample size.

Probability of a Type I Error (α)

It should be apparent that the power of an experiment depends on the cutoff for α . The more willing we are to risk making a Type I error, the more null hypotheses, true and false, we will reject. And as the number of rejected false null hypotheses increases, so does the power of our study.

Statisticians, and particularly those in the behavioral sciences, have generally recognized two traditional levels of α : 0.05 and 0.01. While α can be set at other values, these are the two that we traditionally employ, and these are the two that we are likely to

continue to employ, at least in the near future. But if these are the only levels we feel confident in using, partly because we fear criticism from colleagues and review panels, then we do not have much to play with. We can calculate the power associated with each level of α and make our choice accordingly. Because, all other things equal, power is greater with $\alpha = 0.05$ than with $\alpha = 0.01$, there is a tendency to settle on $\alpha = 0.05$.

When we discuss what Buchner, Erdfelder, and Faul [3] call ‘compromise power’, we will come to a situation where there is the possibility of specifying the relative seriousness of Type I and Type II errors and accepting α at values greater than 0.05. This is an option that deserves to be taken seriously, especially for exploratory studies.

Experimental Design

We often don’t think about the relevance of the experimental design to determination of power, but changes in the design can have a major impact. The calculation of power assumes a specific design, but it is up to us to specify that design. For example, all other things equal, we will generally have more power in a design involving paired observations than in one with independent groups. When all assumptions are met, standard parametric procedures are usually more powerful than their corresponding distribution-free alternatives, though that power difference is often slight. (See Noether [19] for a treatment of power of some common nonparametric tests.) For a third example, McClelland [15] presents data showing how power can vary considerably depending on how the experiment is designed. Consider an experiment that varies the dosage of a drug with levels of 0.5, 1, 1.5, 2, and 2.5 μg . If observations were distributed evenly over those dosages, it would take a total of $2N$ observations to have the same power as an experiment in which only N participants were equally divided between the highest and lowest dosages, with no observations assigned to the other three levels. (Of course, any nonlinear trend in response across the five dosage levels would be untestable.)

McClelland and Judd [16] illustrate similar kinds of effects for the detection of interaction in factorial designs. Their papers, and one by Irwin and McClelland [13] are important for those contemplating optimal experimental designs.

The Calculation of Power

The most common approach to power calculation is the approach taken by Cohen [5]. Cohen developed his procedures in conjunction with appropriate statistical tables, which helps to explain the nature of the approach. However, the same issues arise if you are using statistical software to carry out your calculations. Again, this section describes the general approach and does not focus on the specifics.

An important principle in calculating power and sample size requirements is the separation of our measure of the size of an effect (ES), which does not have anything to do with the size of the sample, from the sample size itself. This allows us to work with the anticipated effect size to calculate the power associated with a given sample size, or to reverse the process and calculate the sample size required for the desired level of power.

Cohen's approach first involves the calculation of the effect size for the particular experimental design. Following that we define a value (here called δ) which is some function of the sample size. For example, when testing the difference between two population means, $\delta = ES\sqrt{n/2}$, where n is equal to the number of observations in any one sample. Analogous calculations are involved for other statistical tests. What we have done here is to calculate ES independent of sample size, and then combine ES and sample size to derive power.

Once we have δ , the next step simply involves entering the appropriate table or statistical software with δ and α , and reading off the corresponding level of power.

If, instead of calculating power for a given sample size, we want to calculate the sample size required for a given level of power, we can simply use the table backwards. We find the value of δ that is required for the desired power, and then use δ and ES to solve for n .

A Priori, Retrospective, and Compromise Power

In general the discussion above has focused on a *a priori* power, which is the power that we would calculate before the experiment is conducted. It is based on reasonable estimates of means, variances, correlations, proportions, and so on, that we believe represent the parameters for our population or populations. This is

what we generally think of when we consider statistical power.

In recent years, there has been an increased interest in what is often called *retrospective (or post hoc) power*. For our purposes retrospective power will be defined as power that is calculated after an experiment has been completed on the basis of the results of that experiment. For example, retrospective power asks the question 'If the values of the population means and variances were equal to the values found in this experiment, what would be the resulting power?'

One reason why we might calculate retrospective power is to help in the design of future research. Suppose that we have just completed an experiment and want to replicate it, perhaps with a different sample size and a demographically different pool of participants. We can take the results that we just obtained, treat them as an accurate reflection of the population means and standard deviations, and use those values to calculate ES . We can then use that ES to make power estimates. This use of retrospective power, which is, in effect, the *a priori* power of our next experiment, is relatively noncontroversial. Many statistical packages, including SAS and SPSS, will make these calculations for you (see **Software for Statistical Analyses**).

What is more controversial, however, is to use retrospective power calculations as an explanation of the obtained results. A common suggestion in the literature claims that if the study was not significant, but had high retrospective power, that result speaks to the acceptance of the null hypothesis. This view hinges on the argument that if you had high power, you would have been very likely to reject a false null, and thus nonsignificance indicates that the null is either true or nearly so. But as Hoenig and Heisey [12] point out, there is a false premise here. It is not possible to fail to reject the null and yet have high retrospective power. In fact, a result with p exactly equal to 0.05 will have a retrospective power of essentially 0.50, and that retrospective power will decrease for $p > 0.05$.

The argument is sometimes made that retrospective power tells you more than you can learn from the obtained P value. This argument is a derivative of the one in the previous paragraph. However, it is easy to show that for a given effect size and sample size, there is a 1:1 relationship between p and retrospective power. One can be derived from the

other. Thus, retrospective power offers no additional information.

As Hoenig and Heisey [12] argue, rather than focus our energies on calculating retrospective power to try to learn more about what our results have to reveal, we are better off putting that effort into calculating confidence limits on the parameter(s) or the effect size. If, for example, we had a t Test on two independent groups with $t(48) = 1.90$, $p = 0.063$, we would fail to reject the null hypothesis. When we calculate retrospective power we find it to be 0.46. When we calculate the 95% confidence interval on $\mu_1 - \mu_2$ we find $-1.10 < \mu_1 - \mu_2 < 39.1$. The confidence interval tells us more about what we are studying than does the fact that power is only 0.46. (Even had the difference been slightly greater, and thus significant, the confidence interval shows that we still do not have a very good idea of the magnitude of the difference between the population means.)

Thomas [24] considered an alternative approach to retrospective power that uses the obtained standard deviation, but ignores the obtained mean difference. You can then choose what you consider to be a minimally acceptable level of power, and solve for the smallest effect size that is detectable at that power. We then know if our study, or better yet future studies, had or have a respectable chance of detecting a behaviorally meaningful result.

Retrospective power can be a useful tool when evaluating studies in the literature, as in a meta-analysis, or planning future work. But retrospective power is not a useful tool for excusing our own nonsignificant results.

A third, and less common, approach to power analysis is called *compromise power*. For *a priori* power we compute ES and δ , and then usually determine power when α is either 0.05 or 0.01. There is nothing that says that we need to use those values of α , but they are the ones traditionally used. However, another way to determine power would be to specify alternative values for α by considering the relative importance we assign to Type I and Type II errors. If Type II errors are particularly serious, such as in early exploratory research, we might be willing to let α rise to keep down β .

One way to approach compromise power is to define an acceptable ratio of the seriousness of Type II and Type I errors. This ratio is β/α . If the two types of error were equally serious, this ratio would be 1. (Our traditional approach often sets α at 0.05 and β at

0.20, for a ratio of 4.) Given the appropriate software we could then base our power calculations on our desire to have the appropriate β/α ratio, rather than fixing α at 0.05 and specifying the desired power.

There is no practical way that we can calculate compromise power using standard statistical tables, but the program G*Power, to be discussed shortly, will allow us to perform the necessary calculation. For example, suppose that we contemplated a standard experiment with $n = 25$ participants in each of two groups, and an expected effect size of 0.50. Using a traditional cutoff of $\alpha = 0.05$, we would have power equal to 0.41, or $\beta = 0.59$. In this case, the probability of making a Type II error is approximately 12 times as great as the probability of making a Type I error. But we may think that ratio is not consistent of our view of the seriousness of these two types of errors, and that a ratio of 2:1 would be more appropriate. We can thus set the β/α ratio at 2. For a two group experiment with 25 participants in each group, where we expected $ES = 0.50$, a medium effect size, then G*Power shows that letting $\alpha = 0.17$ will give us a power of 0.65. That is an abnormally high risk of a Type I error, but it is more in line with our subjective belief in the seriousness of the two kinds of errors, and may well be worth the price in exploratory research. This is especially true if our sample size is fixed at some value by the nature of the research, so that we can't vary N to increase our power.

Power Analysis Software

There is a wealth of statistical software for power analysis available on the World Wide Web, and even a casual search using www.google.com will find it. Some of this software is quite expensive, and some of it is free, but 'free' should not be taken to mean 'inferior'. Much of it is quite good. Surprisingly most of the available software was written a number of years ago, as is also the case with reviews of that software.

I will only mention a few pieces of software here, because the rest is easy to find on the web. To my mind the best piece of software for general use is G*Power [10], which can be downloaded for free at <http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>. This software comes with an excellent manual and is available for both Macintosh and PC computers. It

will handle tests on correlations, proportions, means (using both t and F), and contingency tables.

You can also perform excellent power analyses from nQuery Advisor (see <http://www.statsol.ie/nquery/nquery.htm>). This is certainly not free software, but it handles a wide array of experimental designs and does an excellent job of allowing you to try out different options and plot your results. It will also allow you to calculate sample size required to produce a confidence interval of specified width.

A very nice set of Java applets have been written by Russell Lenth, and are available at <http://www.stat.uiowa.edu/~rlenth/Power/>. An advantage of using Java applets for calculation of power is that you only have to open a web page to do so.

A very good piece of software is a program called *DataSim* by Bradley [2]. It is currently available for Macintosh systems, though there is an older DOS version. It is not very expensive and will carry out many analyses in addition to power calculations. *DataSim* was originally designed as a data simulation program, at which it excels, but it does a very nice job of doing power calculations.

For those who use SAS for data analysis, O'Brien and Muller have written a power module, which is available at <http://www.bio.ri.ccf.org/Power/>. This module is easy to include in a SAS program.

A good review of software for power calculations can be found at <http://www.zoology.ubc.ca/~krebs/power.html>. This review is old, but in this case that does not seem to be much of a drawback.

Confidence Limits on Effect Size

Earlier in this article I briefly discussed the advantage of setting confidence limits on the parameter(s) being estimated, and argued that this was an important way of understanding our results. It is also possible to set confidence limits on effect sizes.

Smithson [22] has argued for the use of confidence limits on (standardized) effect size, on the grounds that confidence limits on unstandardized effects do not allow us to make comparisons across studies using different variables. On the other hand, confidence intervals on standardized effect sizes, which

are dimensionless, can more easily be compared. While one could argue with the requirement of comparability across variables, there are times when it is useful.

The increased emphasis on confidence limits has caused some [1] to suggest that sample size calculations should focus solely on the width of the resulting confidence interval, and should disregard the consideration of power for statistical tests. Daly [9], on the other hand, has emphasized that we need to consider both confidence interval width and the minimum meaningful effect size (such as a difference in means) when it comes to setting our sample sizes. Otherwise we risk seriously underestimating the sample size required for satisfactory power. For example, suppose that the difference in means between two treatments is expected to be 5 points, and we calculate the sample size needed for a 95% confidence interval to extend just under 5 points on either side of the mean. If we obtain such an interval and the observed difference was 5, the confidence interval would not include 0 and we would reject our null hypothesis. But a difference of 5 points is an estimate, and sampling error will cause our actual data to be larger or smaller than that. Half the time the obtained interval will come out to be wider than that, and half the time it will be somewhat narrower. For the half the time that it is wider, we would fail to reject the null hypothesis of no difference. So in setting the sample size to yield a mean width of the confidence interval with predefined precision, our experiment has a power coefficient of only 0.50, which we would likely find unacceptable.

It is possible to set confidence limits on power as well as on effect sizes, making use of the fact that the sample variance is a random variable. However, our estimates of power, especially *a priori* power, are already sufficiently tentative that the effort is not likely to be rewarding. You would most likely be further ahead putting the effort into calculating, and perhaps plotting, power for a range of reasonable values of the effect sizes and/or sample sizes.

Smithson [23] has argued that one should calculate both power and a confidence interval on the effect size. He notes that the confidence interval does not narrow in width as power increases because of an increased *ES*, and it is important to anticipate the width of our confidence interval. We might then wish to increase the sample size, even in the presence of high power, so as to decrease

6 Power

Table 1 Effect size formulae for various standard tests (column 2) and sample sizes required for power = 0.80 with $\alpha = 0.05$ for several levels of effect size (columns 3–7). (The formulae were taken from Cohen [5], and the estimated sample sizes were calculated using G*Power (where appropriate) or tables from Cohen [5])

Test	Effect size formula	ES				
		0.2	0.4	0.6	0.8	1.0
Difference between two independent means ($H_0: \mu_1 - \mu_2 = 0$)	$ES = \frac{\mu_1 - \mu_2}{\sigma_e}$	394	100	45	26	17
Difference between means of paired samples ($H_0: \mu_1 - \mu_2 = \mu_D = 0$)	$ES = \frac{\mu_1 - \mu_2}{\sigma_D} = \frac{\mu_D}{\sigma_D}$	394	100	45	26	17
One way analysis of variance ($H_0: \mu_1 = \mu_2 = \dots = \mu_k = 0$) (sample sizes/group for $df_{bet} = 2$)	$ES = \sqrt{\frac{\sum(\mu_j - \mu)^2/k}{\sigma_e^2}}$	82	22	10	6	–
Product moment correlation	$ES = \rho$	91	44	17	7	–
Comparing two independent proportions ($H_0: P_1 - P_2 = 0$); $\phi = 2\arcsin \sqrt{P}$	$ES = \phi_1 - \phi_2$	392	98	44	25	16

Test	Effect Size	ES				
		0.05	0.10	0.20	0.30	0.40
Multiple correlation (Sample size for 2 predictors)	$ES = \frac{R^2}{1 - R^2}$	196	100	52	36	28
Multiple partial correlation	$ES = \frac{R_{01.2}^2}{1 - R_{01.2}^2}$	196	100	52	36	28

Test	Effect Size	ES				
		0.05	0.10	0.15	0.20	0.25
Single proportion against $P = 0.50$ ($H_0: P = 0.50$)	$ES = P_1 - 0.50$	783	194	85	49	30
						12
Goodness of fit chi-square (sample sizes for $df = 3$)	$ES = \sqrt{\sum_{i=1}^c \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	331	68	30	<25	<25
Contingency table chi-square (sample size for $df = 1$)	$ES = \sqrt{\sum_{i=1}^c \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	196	44	<25	<25	<25

P_1, P_2 = population proportions under H_1 .

P_0 = population proportion under H_0 .

c = number of cells in one way or contingency table.

μ_1, μ_2, μ_j = population means.

σ_e = population standard deviation.

σ_D = standard deviation of difference scores.

k = number of groups.

the confidence interval on the size of the resulting effect.

Calculating Power for Specific Experimental Designs

Once we have an estimate of our effect size, the calculation of power is straightforward. As discussed above, there are a number of statistical programs available for computing power. In addition, hand calculations using standard tables (e.g., Cohen [5]) are relatively simple and straightforward. Regardless of whether computations are done by software or by hand, we can choose to calculate the power for a given sample size, or the sample size required to produce a desired level of power. Whether we are using software or standard tables, we enter the effect size and the sample size, and then read off the available power. Alternatively we could enter the effect size and the desired power and read off the required sample size. Table 1, in addition to defining the effect size for each of many different designs, also contains the sample sizes required for power = 0.80 for a representative set of effect sizes.

References

- [1] Beal, S.L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means, *Biometrics* **45**, 969–977.
- [2] Bradley, D.R. (1997). DATASIM: a general purpose data simulator, in *Technology and Teaching*, L. Lloyd, ed., Information Today, Medford, pp. 93–117.
- [3] Buchner, A., Erdfelder, E. & Faul, F. (1997). *How to Use G*Power* [WWW document], URL : http://www.psych.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html
- [4] Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review, *Journal of Abnormal and Social Psychology* **65**, 145–153.
- [5] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.
- [6] Cohen, J. (1990). Things I have learned (so far), *American Psychologists* **45**, 1304–1312.
- [7] Cohen, J. (1992a). A power primer, *Psychological Bulletin* **112**, 155–159.
- [8] Cohen, J. (1992b). Statistical power analysis, *Current Directions in Psychological Science* **1**, 98–101.
- [9] Daly, L.E. (2000). Confidence intervals and sample size, in *Statistics with Confidence*, D.G. Altman, D. Machin, T.N. Bryant & M.J. Gardner, eds, BMJBooks, London.
- [10] Erdfelder, E., Faul, F. & Buchner, A. (1996). G*Power: a general power analysis program, *Behavior Research Methods, Instruments, & Computers* **28**, 1–11.
- [11] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.
- [12] Hoenig, J.M. & Heisey, D.M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis, *American Statistician* **55**, 19–24.
- [13] Irwin, J.R. & McClelland, G.H. (2003). Negative consequences of dichotomizing continuous predictor variables, *Journal of Marketing Research* **40**, 366–371.
- [14] Lenth, R.V. (2001). Some practical guidelines for effective sample size determination, *American Statistician* **55**, 187–193.
- [15] McClelland, G.H. (1997). Optimal design in psychological research, *Psychological Methods* **2**, 3–19.
- [16] McClelland, G.H. & Judd, C.M. (1993). Statistical difficulties of detecting interactions and moderator effects, *Psychological Bulletin* **114**, 376–390.
- [17] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference, *Biometrika* **20a**, 175–240.
- [18] Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Transactions of the Royal Society of London. Series A* **231**, 289–337.
- [19] Noether, G.E. (1987). Sample size determination for some common nonparametric tests, *Journal of the American Statistical Association* **82**, 645–647.
- [20] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?, *Psychological Bulletin* **105**, 309–316.
- [21] Senn, S. (1997). *Statistical Issues of Drug Development*, Wiley, Chichester.
- [22] Smithson, M. (2000). *Statistics with Confidence*, Sage Publications, London.
- [23] Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: the importance of noncentral distributions in computing intervals, *Educational and Psychological Measurement* **61**, 605–632.
- [24] Thomas, L. (1997). Retrospective power analysis, *Conservation Biology* **11**, 276–280.

DAVID C. HOWELL

Power Analysis for Categorical Methods

EDGAR ERDFELDER, FRANZ FAUL AND AXEL BUCHNER

Volume 3, pp. 1565–1570

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Power Analysis for Categorical Methods

In testing statistical hypotheses, two types of error can be made: (a) a type-1 error, that is, a valid null hypothesis (H_0) is falsely rejected, and (b) a type-2 error, that is, an invalid H_0 is falsely retained when the alternative hypothesis H_1 is in fact true. Obviously, a reasonable decision between H_0 and H_1 requires both the type-1 error probability α and the type-2 error probability β to be sufficiently small. The **power** of a statistical test is the complement of the type-2 error probability, $1 - \beta$, and it represents the probability of correctly rejecting H_0 when H_1 holds [4]. As described in this entry, researchers can assess or control the power of statistical tests for categorical data by making use of the following types of power analysis [6]:

1. *Post hoc power analysis*: Compute the power of the test for a given sample size N , an alpha-level α , and an effect of size w , that is, a prespecified deviation w of the population parameters under H_1 from those specified by H_0 (see [4], Chapter 7);
2. *A-priori power analysis*: Compute the sample size N that is necessary to obtain a power level $1 - \beta$, given an effect of size w and a type-1 error level α .
3. *Compromise power analysis*: Compute the critical value, and the associated α and $1 - \beta$ probabilities, of the test statistic defining the boundary between decisions for H_1 and H_0 , given a sample of size N , an effect of size w and a desired ratio $q = \beta/\alpha$ of the error probabilities reflecting the relative seriousness of the two error types.
4. *Sensitivity analysis*: Compute the effect size w that can be detected given a power level $1 - \beta$, a specific α level, and a sample of size N .

Statistical Hypotheses for Categorical Data

Statistical tests for categorical data often refer to a sample of N independent observations drawn randomly from a population where each observation can be assigned to one and only one of J observation categories $C_1, C_2, \dots, C_J$. These categories may correspond to the J possible values of a single discrete

variable (e.g., religion) or to the J cells of a bivariate or multivariate contingency table. For example, for a 2 by 2 table of two dichotomous variables (*see Two by Two Contingency Tables*) A (say, gender) and B (say, interest in soccer) the number of categories would be $J = 2 \times 2 = 4$. Two types of statistical hypotheses need to be distinguished:

1. *Simple null hypotheses* assume that the observed frequencies $y_1, y_2, \dots, y_J$ of the J categories are a random sample drawn from a multinomial distribution with category probabilities $p(C_j) = \pi_{oj}$, for $j = 1, \dots, J$. That is, each category C_j has a certain probability π_{oj} exactly specified by H_0 . An example would be a uniform distribution across the categories such that $\pi_{oj} = 1/J$ for all j . Thus, for $J = 3$ religion categories (say, C_1 : Catholics, C_2 : Protestants, C_3 : other) the uniform distribution hypothesis predicts $H_0: \pi_{o1} = \pi_{o2} = \pi_{o3} = 1/3$. In contrast, the alternative hypothesis H_1 would predict that at least two category probabilities differ from $1/3$. Power analyses for simple null hypotheses are described in the section “Simple Null Hypotheses”.
2. *Composite null hypotheses* are also based on a multinomial distribution model for the observed frequencies but do not specify exact category probabilities. Instead, the probabilities $p(C_j)$ of the categories C_j , $j = 1, \dots, J$, are assumed to be functions of S unknown real-valued parameters $\theta_1, \theta_2, \dots, \theta_s$ that need to be estimated from the data. Hence, $p(C_j) = f_j(\theta_1, \theta_2, \dots, \theta_s)$, where the model equations f_j define a function mapping the parameter space Ω (the set of all possible parameter vectors $\theta = (\theta_1, \theta_2, \dots, \theta_s)$) onto a subset $\prod_o$ of the set $\prod$ of all possible category probability vectors. Examples include parameterized multinomial models playing a prominent role in behavioral research such as **log-linear models** [1], processing-tree models [2], and signal-detection models (*see Signal Detection Theory*) [7]. Composite null hypotheses posit that the model defined by the model equations is valid, that is, that the vector $\pi = (p(C_1), p(C_2), \dots, p(C_J))$ of category probabilities in the underlying population is consistent with the model equations. Put more formally, $H_0: \pi \in \prod_o$ and, correspondingly, $H_1: \pi \notin \prod_o$. Power analyses for composite null hypotheses

2 Power Analysis for Categorical Methods

are described in the section “Composite Null Hypotheses”.

The section on “Joint Multinomial Models” is devoted to the frequent situation that $K > 1$ samples of N_k observations in sample $k, k = 1, \dots, K$, are drawn randomly and independently from K different populations, with a multinomial distribution model holding for each sample. Fortunately, the power analysis procedures described in sections “Simple Null Hypotheses” and “Composite Null Hypotheses” easily generalize to such joint multinomial models both for simple and composite null hypotheses.

Simple Null Hypotheses

Any test statistic PD_λ of the power divergence family can be used for testing simple null hypotheses for multinomial models [8]:

$$\begin{aligned}
 PD_\lambda &= \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^J y_j \left[\left(\frac{y_j}{N\pi_{oj}} \right)^\lambda - 1 \right], \\
 &\text{for any real - valued } \lambda \notin \{-1, 0\} \\
 PD_{\lambda=-1} &= \lim_{\lambda \rightarrow -1} \left(\frac{2}{\lambda(\lambda+1)} \right. \\
 &\quad \times \left. \sum_{j=1}^J y_j \left[\left(\frac{y_j}{N\pi_{oj}} \right)^\lambda - 1 \right] \right) \\
 PD_{\lambda=0} &= \lim_{\lambda \rightarrow 0} \left(\frac{2}{\lambda(\lambda+1)} \right. \\
 &\quad \times \left. \sum_{j=1}^J y_j \left[\left(\frac{y_j}{N\pi_{oj}} \right)^\lambda - 1 \right] \right) \\
 &= 2 \sum_{j=1}^J y_j \ln \left(\frac{y_j}{N\pi_{oj}} \right) \tag{1}
 \end{aligned}$$

As specified above, y_j denotes the observed frequency in category j , N the total sample size, and π_{oj} the probability assigned to category j by H_0 so that $N\pi_{oj}$ is the expected frequency under H_0 . Read and Cressie [8] have shown that, if H_0 holds and Birch’s [3] regularity conditions are met, PD_λ is asymptotically chi square distributed with $df = J - 1$ for each

fixed value of λ . Both the well-known Pearson chi-square statistic (*see Contingency Tables*)

$$X^2 = \sum_{j=1}^J \frac{(y_j - N\pi_{oj})^2}{N\pi_{oj}} = PD_{\lambda=1} \tag{2}$$

and the often used log-likelihood-ratio chi-square statistic

$$G^2 = 2 \sum_{j=1}^J y_j \ln \left(\frac{y_j}{N\pi_{oj}} \right) = PD_{\lambda=0} \tag{3}$$

are special cases of the PD_λ family. However, Read and Cressie [8] argued that these statistics may not be optimal. The Cressie-Read statistic PD_λ with $\lambda = 2/3$, a ‘compromise’ between G^2 and X^2 , performs amazingly well in many situations when compared to G^2 , X^2 and other PD_λ statistics.

If H_0 is violated and H_1 holds with a category probability vector $\pi_1 = (\pi_{11}, \pi_{12}, \dots, \pi_{1J})$, $\pi_1 \neq \pi_0$, then the distribution of any PD_λ statistic can be approximated by a noncentral chi-square distribution with $df = J - 1$ and noncentrality parameter

$$\begin{aligned}
 \gamma_\lambda &= \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^J N\pi_{1j} \left[\left(\frac{N\pi_{1j}}{N\pi_{oj}} \right)^\lambda - 1 \right], \\
 &\text{for any real - valued } \lambda \notin \{-1, 0\}, \tag{4}
 \end{aligned}$$

with $\gamma_{\lambda=-1}$ and $\gamma_{\lambda=0}$ defined as limits as in (1) [8]. Thus, the noncentrality parameter γ_λ is just PD_λ , with the expected category frequencies under H_1 , $N\pi_{1j}$, replacing the observed frequencies y_j in (1). Alternatively, one can first compute an effect size index w_λ (*see Effect Size Measures*) measuring the deviation of the H_1 parameters from the H_0 parameters uncontaminated by the sample size N :

$$\begin{aligned}
 w_\lambda &= \sqrt{\frac{2}{\lambda(\lambda+1)} \sum_{j=1}^J \pi_{1j} \left[\left(\frac{\pi_{1j}}{\pi_{oj}} \right)^\lambda - 1 \right]}, \\
 &\text{for any real - valued } \lambda \notin \{-1, 0\}. \tag{5}
 \end{aligned}$$

In a second step, we derive the noncentrality parameter γ_λ from w_λ and the sample size N as follows:

$$\gamma_\lambda = Nw_\lambda^2. \tag{6}$$

The effect size index for $\lambda = 1$, $w_{\lambda=1}$, is equivalent to Cohen’s [4] effect size parameter w for

Pearson's chi-square test:

$$\mathbf{w} = \sqrt{\sum_{j=1}^J \frac{(\pi_{1j} - \pi_{0j})^2}{\pi_{0j}}} = w_{\lambda=1} \quad (7)$$

Cohen ([4], Chapter 7) suggested to interpret effects of sizes $\mathbf{w} = 0.1$, $\mathbf{w} = 0.3$, and $\mathbf{w} = 0.5$ as 'small', 'medium', and 'large', respectively, and it seems reasonable to generalize these effect size conventions to other measures of the w_λ family. However, it has been our experience that these labels may be misleading in some applications because they do not always correspond to intuitions about small, medium, and large effects if expressed in terms of parameter differences between H_1 and H_0 . Therefore, we prefer assessing the power of a test in terms of the model parameters under H_1 , not in terms of the $\mathbf{w}$ effect size conventions.

As an illustrative example, let us assume that we want to test the equiprobability hypothesis for $J = 3$ religion types ($H_0: \pi_{01} = \pi_{02} = \pi_{03} = 1/3$) against the alternative hypothesis $H_1: \pi_{01} = \pi_{02} = 1/4; \pi_{03} = 1/2$ using Pearson's X^2 statistic (i.e., $PD_{\lambda=1}$) with $df = 3 - 1 = 2$. If we have $N = 100$ observations and select $\alpha = .05$, what is the power of this test? One way of answering this question would be to compute $\mathbf{w}$ for these parameters as defined in (7) and then use (6) to derive the noncentrality parameter $\gamma_{\lambda=1}$. The power of the test is simply the area under this noncentral chi-square distribution to the right of the critical value corresponding to $\alpha = .05$. The power table 7.3.16 of Cohen ([4], p. 235) can be used for approximating this power value. More convenient and flexible are PC-based power analysis tools such as GPOWER [6]¹. In GPOWER, select 'Chi-square Test' and 'Post hoc' type of power analysis. Next, click on 'calc effect size' (Macintosh version: calc w) to enter the category probabilities under H_0 and H_1 . 'Calc&Copy' provides us with the effect size $\mathbf{w} = 0.3536$ and copies it to the main window of GPOWER. We also need to enter $\alpha = .05$, $N = 100$, and $df = 2$ before a click on 'calculate' yields the post hoc power value $1 - \beta = .8963$ for this situation.

Assume that we are not satisfied with this state of affairs because the two error probabilities are not exactly balanced ($\alpha = .05$, $\beta = 1 - .8963 = .1037$). To remedy this problem, we select the 'compromise' type of power analysis in GPOWER and enter

$q = \text{beta}/\alpha = 1$ as the desired ratio of error probabilities. Other things being equal, we obtain a critical value of 5.1791 as an optimal decision criterion between H_0 and H_1 , corresponding to exactly balanced error probabilities of $\alpha = \beta = .0751$.

If we are still not satisfied and want to make sure that $\alpha = \beta = .05$ for $\mathbf{w} = 0.3536$, we can proceed to an 'A priori' type of power analysis in GPOWER, which yields $N = 124$ as the necessary sample size required by the input parameters.

Composite Null Hypotheses

As an example of a composite H_0 , consider the log-linear model without interaction term for the 2 by 2 contingency table of two dichotomous variables A (gender) and B (interest in soccer):

$$\ln(e_{0ij}) = u + u_{Ai} + u_{Bj}, \text{ with} \quad (8)$$

$$\sum_{i=1}^2 u_{Ai} = \sum_{j=1}^2 u_{Bj} = 0$$

The logarithms of the expected frequencies for cell (i, j) , $i = 1, \dots, I$, $j = 1, \dots, J$, are explained in terms of a grand mean u , a main effect of the i -th level of variable A , and a main effect of the j -th level of variable B . None of these parameters is specified *a priori*, so that we now have a composite H_0 rather than a simple H_0 . Nevertheless, any Read-Cressie statistic

$$PD_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J y_{ij} \left[\left(\frac{y_{ij}}{\hat{e}_{0ij}} \right)^\lambda - 1 \right], \quad (9)$$

for any real - valued $\lambda \notin \{-1, 0\}$,

with $PD_{\lambda=-1}$ and $PD_{\lambda=0}$ defined as limits as in (1), is still asymptotically chi-square distributed under this composite H_0 , albeit with $df = IJ - 1 - S$, where S is the number of free parameters estimated from the data [8]. In our example, we have $IJ = 2 \times 2$ cells, and we estimate $S = 2$ free parameters u_{A1} and u_{B1} ; all other parameters are determined by N and the two identifiability constraints $\sum_{i=1}^2 u_{Ai} = \sum_{j=1}^2 u_{Bj} = 0$. Hence, we have $df = 4 - 1 - 2 = 1$. As before, y_{ij} denotes the observed frequency in cell (i, j) , and $\hat{e}_{0ij}$ is the corresponding expected cell frequency under the model (8), with the u parameters being estimated

4 Power Analysis for Categorical Methods

from the observed frequencies by minimizing the PD_λ statistic chosen by the researcher [8].

If the H_0 model is false and the saturated **log-linear model**

$$\ln(e_{1ij}) = u + u_{Ai} + u_{Bj} + u_{AiBj},$$

$$\sum_{i=1}^2 u_{Ai} = \sum_{j=1}^2 u_{Bj} = \sum_{i=1}^2 u_{AiBj} = \sum_{j=1}^2 u_{AiBj} = 0, \quad (10)$$

holds with interaction terms $u_{AiBj} \neq 0$, then PD_λ as defined in (9) is approximately noncentrally chi-square distributed with $df = IJ - 1 - S$ and noncentrality parameter

$$\gamma_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J \mathbf{e}_{1ij} \left[\left(\frac{\mathbf{e}_{1ij}}{\hat{\mathbf{e}}_{0ij}} \right)^\lambda - 1 \right],$$

for any real - valued $\lambda \notin \{-1, 0\}$. (11)

As before, $\gamma_{\lambda=-1}$ and $\gamma_{\lambda=0}$ are defined as limits. In (11), $\mathbf{e}_{1ij}$ denotes the expected frequency in cell (i, j) under the H_1 model (10) and $\hat{\mathbf{e}}_{0ij}$ denotes the corresponding expected frequency under the H_0 model (8) estimated from u parameters that minimize the noncentrality parameter γ_λ . Thus, to obtain the correct noncentrality parameter for tests of composite null hypotheses, simply set up an artificial data set containing the expected cell frequencies under H_1 as ‘data’. Then fit the null model to these ‘data’ and compute PD_λ . The PD_λ ‘statistic’ for this artificial data set corresponds to the noncentrality parameter γ_λ required for computing the power of the PD_λ test under H_1 . Note that this rule works for any family of parameterized multinomial models (e.g., logit models, ogive models, processing-tree models),

not just for log-linear models designed for two-dimensional contingency tables.

As an illustrative example, we will refer to the chi-square test of association for the 2 by 2 table (cf. [4]). Let us assume that the additive null model (8) is wrong, and the alternative model (10) holds with parameters $u = 3.0095$ and $u_{A1} = u_{B1} = u_{A1B1} = 0.5$. By computing the natural logarithms of the expected cell frequencies according to (10) and taking the antilogarithms, we obtain the expected cell frequencies $\mathbf{e}_{1ij}$ shown on the left side of Table 1. Fitting the null model (8) without interaction term to these expected frequencies using the G^2 ‘statistic’ ($= PD_{\lambda=0}$) gives us the expected frequencies under the null model shown in the right part of Table 1.

What would be the power of the G^2 test for this H_1 in case of a sample size of $N = 120$ and a type-1 error probability $\alpha = .05$? In the first step, we calculate the effect size $w_{\lambda=0}$ measuring the discrepancy between the H_0 and the H_1 probabilities corresponding to the expected frequencies in Table 1 according to (5). We obtain the rather small effect size $w_{\lambda=0} = 0.1379$ which is close to Cohen’s effect size measure $w = 0.1503$ for the same H_1 . Using (6), we compute the noncentrality parameter $\gamma_{\lambda=0} = 120 \times 0.1379^2 = 2.282$. The area to the right of the critical value corresponding to $\alpha = .05$ under the noncentral chi-square distribution with $df = 4 - 1 - 2 = 1$ is the power of the G^2 test for the composite H_0 (i.e., model (8)). We can compute it very easily by using GPOWER’s ‘Chi-square test’ option together with the ‘Post hoc’ type of power analysis. Selecting $\alpha = .05$, $w = 0.1379$, $N = 120$, $df = 1$, and clicking on ‘calculate’ provides us with the very low power value $1 - \beta = .3269$. A ‘compromise’ power analysis based on the error probability ratio $q = \beta/\alpha = 1$ results in disappointing values of $\alpha = \beta = .3068$

Table 1 Expected frequencies for a 2 by 2 contingency table under the saturated log-linear model with parameters $u = 3.0095$, $u_{A1} = u_{B1} = u_{A1B1} = 0.5$ (H_1 ; left side) and under a log-linear model without interaction term (H_0 ; right side)

a) Expected frequencies under the saturated model (H_1)				b) Expected frequencies under the additive model (H_0)			
A B:	Interest in soccer		Σ	A B:	Interest in soccer		Σ
Gender	Yes	No		Gender	Yes	No	
Male	90.878	12.298	103.176	Male	88.705	14.471	103.176
Female	12.298	4.525	16.832	Female	14.471	2.361	16.832
Σ	103.176	16.832	120	Σ	103.176	16.832	120

for the same set of parameters. Because both error probabilities are unacceptably large, we proceed to an ‘*a priori*’ power analysis in GPOWER to learn that we need $N = 684$ to detect an effect of size $w_{\lambda=0} = 0.1379$ with a $df = 1$ likelihood-ratio test at $\alpha = .05$ and a power of $1 - \beta = .95$.

Joint Multinomial Models

So far we have discussed the situation of a single sample of N observations drawn randomly from one multinomial distribution (see **Catalogue of Probability Density Functions**). In behavioral research, we are often faced with the problem to compare distributions of discrete variables for $K > 1$ populations (e.g., experimental and control groups, target-present versus target-absent trials, trials with liberal and conservative pay-off schedules, etc.). Fortunately, the power analysis procedures for simple and composite null hypotheses easily generalize to $K > 1$ samples drawn randomly and independently from K multinomial distributions (i.e., joint multinomial models). Again, we can use any PD_λ statistic:

$$PD_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{k=1}^K \sum_{j=1}^{J_k} y_{kj} \left[\left(\frac{y_{kj}}{N_k \pi_{okj}} \right)^\lambda - 1 \right], \quad (12)$$

where the special cases $PD_{\lambda=-1}$ and $PD_{\lambda=0}$ are defined as limits. In (12), y_{kj} denotes the observed frequency in the j -th category, $j = 1, \dots, J_k$, of the k -th sample, $k = 1, \dots, K$, and N_k is the size of the k -th sample. In case of simple null hypotheses requiring no parameter estimation, π_{okj} is the probability of the j -th category in population k hypothesized by H_0 . In case of composite null hypotheses, simply replace $N_k \pi_{okj}$ by the expected frequencies under H_0 that are closest to the observed frequencies in terms of PD_λ . Under H_0 and Birch’s [3] regularity conditions for joint multinomial models, any PD_λ statistic is asymptotically chi-square distributed with

$$df = \sum_{k=1}^K (J_k - 1) - S, \quad (13)$$

where S is the number of free parameters estimated from the data. By contrast, if H_0 is actually false and H_1 holds with expected frequencies $\mathbf{e}_{1kj}$ in the j -th

category of population k , the same statistic is approximately noncentral chi-square distributed with the same degrees of freedom and noncentrality parameter

$$\gamma_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum_{k=1}^K \sum_{j=1}^{J_k} \mathbf{e}_{1kj} \left[\left(\frac{\mathbf{e}_{1kj}}{\mathbf{e}_{0kj}} \right)^\lambda - 1 \right]. \quad (14)$$

In (14), $\mathbf{e}_{1kj}$ and $\mathbf{e}_{0kj}$ denote the expected frequencies under H_1 and H_0 , respectively. In case of composite null hypotheses requiring parameter estimation, the $\mathbf{e}_{0kj}$ must be calculated from H_0 parameters that minimize γ_λ .

Researchers preferring to perform power analyses in terms of standardized effect size measures need separate effect size measures for each of the K populations. A suitable index for the k -th population (inspired by [5], Footnote 1) is

$$w_{\lambda(k)} = \sqrt{\frac{2}{\lambda(\lambda + 1)} \sum_{j=1}^{J_k} \frac{\mathbf{e}_{1kj}}{N_k} \left[\left(\frac{\mathbf{e}_{1kj}}{\mathbf{e}_{0kj}} \right)^\lambda - 1 \right]}, \quad (15)$$

which relates to the total effect size across the K populations as follows:

$$w_\lambda = \sqrt{\sum_{k=1}^K \frac{N_k}{N} (w_{\lambda(k)})^2}. \quad (16)$$

If we use (6) to compute the noncentrality parameter from w_λ , we obtain the same value γ_λ as from (14). The power of the PD_λ test is simply the area to the right of the critical value determined by α under the noncentral chi-square distribution with df given by (13) and noncentrality parameter given by (14). GPOWER can be used to compute this power value very conveniently.

Acknowledgment

The work on this entry has been supported by grants from the TransCoop Program of the Alexander von Humboldt Foundation and the Otto Selz Institute, University of Mannheim.

Note

1. The most recent versions of GPOWER can be downloaded free of charge for several computer platforms at <http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.
- [2] Batchelder, W.H. & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling, *Psychonomic Bulletin & Review* **6**, 57–86.
- [3] Birch, M.W. (1964). A new proof of the Pearson-Fisher theorem, *Annals of Mathematical Statistics* **35**, 817–824.
- [4] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Hillsdale.
- [5] Erdfelder, E. & Bredenkamp, J. (1998). Recognition of script-typical versus script-atypical information: effects of cognitive elaboration, *Memory & Cognition* **26**, 922–938.
- [6] Erdfelder, E., Faul, F. & Buchner, A. (1996). GPOWER: a general power analysis program, *Behavior Research Methods, Instruments, & Computers* **28**, 1–11.
- [7] Macmillan, N.A. & Creelman, C.D. (1991). *Detection Theory: A User's Guide*, Cambridge University Press, Cambridge.
- [8] Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.

EDGAR ERDFELDER, FRANZ FAUL AND
AXEL BUCHNER

Power and Sample Size in Multilevel Linear Models

TOM A.B. SNIJDERS

Volume 3, pp. 1570–1573

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Power and Sample Size in Multilevel Linear Models

Power of statistical tests generally depends on sample size and other design aspects; on effect size or, more generally, parameter values; and on the level of significance. In multilevel models (see **Linear Multilevel Models**), however, there is a sample size for each level, defined as the total number of units observed for this level. For example, in a three-level study of pupils nested in classrooms nested in schools, there might be observations on 60 schools, a total of 150 classrooms, and a total of 3300 pupils. On average, in the data, each classroom then has 22 pupils, and each school contains 2.5 classrooms. What are the relevant sample sizes for power issues? If the researcher has the freedom to choose the sample sizes for a planned study, what are sensible guidelines?

Power depends on the parameter being tested, and power considerations are different depending on whether the researcher focuses on, for example, testing a regression coefficient, a variance parameter, or is interested in the size of means of particular groups. In most studies, regression coefficients are of primary interest, and this article focuses on such coefficients. The cited literature gives methods to determine power and required sample sizes also for estimating parameters in the random part of the model.

A primary qualitative issue is that, for testing the effect of a level-one variable, the level-one sample size (in the example, 3300) is of main importance; for testing the effect of a level-two variable it is the level-two sample size (150 in the example); and so on. The average cluster sizes (in the example, 22 at level two and 2.5 at level three) are not very important for the power of such tests. This implies that the sample size at the highest level is the main limiting characteristic of the design. Almost always, it will be more informative to have a sample of 60 schools with 3300 pupils than one of 30 schools also with 3300 pupils. A sample of 600 schools with a total of 3300 pupils would even be a lot better with respect to power, in spite of the low average number of students (5.5) sampled per school, but in practice, such a study would of course be much more expensive. A second qualitative issue is that for testing fixed regression

coefficients, small cluster sizes are not a problem. The low average number of 2.5 classrooms per school has in itself no negative consequences for the power of testing regression coefficients. What is limited by this low average cluster size, is the power for testing random slope variances at the school level, that is, between-school variances of effects of classroom- or pupil-level variables; and the reliability of estimating those characteristics of individual schools, calculated from classroom variables, that differ strongly between classes. (It may be recalled that the latter characteristics of individual units will be estimated in the multilevel methodology by posterior means, also called *empirical Bayes estimates*; see **Random Effects in Multivariate Linear Models: Prediction**.)

When quantitative insight is required in power for testing regression coefficients, it often is convenient to consider power as a consequence of the standard error of estimation. Suppose we wish to test the null hypothesis that a regression coefficient γ is 0, and for this coefficient we have an estimate $\hat{\gamma}$, which is approximately normally distributed, with standard error $\text{s.e.}(\hat{\gamma})$. The t -ratio $\hat{\gamma}/\text{s.e.}(\hat{\gamma})$ can be tested using a t distribution; if the sample size is large enough, a standard normal distribution can be used. The power will be high if the true value (effect size) of γ is large and if the standard error is small; and a higher level of significance (larger permitted probability of a type I error) will lead to a higher power. This is expressed by the following formula, which holds for a one-sided test, and where the significance level is indicated by α and the type II error probability, which is equal to 1 minus power, by β :

$$\frac{\gamma}{\text{s.e.}(\hat{\gamma})} \approx z_{1-\alpha} + z_{1-\beta}, \quad (1)$$

where $z_{1-\alpha}$ and $z_{1-\beta}$ are the critical points of the standard normal distribution. For example, to obtain at significance level $\alpha = 0.05$ a fairly high power of at least $1 - \beta = 0.80$, we have $z_{1-\alpha} = 1.645$ and $z_{1-\beta} = 0.84$, so that the ratio of true parameter value to standard error should be at least $1.645 + 0.84 \approx 2.5$.

For two-sided tests, the approximate formula is

$$\frac{\gamma}{\text{s.e.}(\hat{\gamma})} \approx z_{1-(\alpha/2)} + z_{1-\beta}, \quad (2)$$

but in the two-sided case this formula holds only if the power is not too small (or, equivalently, the effect size is not too small), for example, $1 - \beta \geq 0.3$.

2 Power and Sample Size in Multilevel Linear Models

For some basic cases, explicit formulas for the estimation variances (i.e., squared standard errors) are given below. This gives a basic knowledge of, and feeling for, the efficiency of multilevel designs. These formulas can be used to compute required sample sizes. The formulas also underpin the qualitative issues mentioned above. A helpful concept for developing this feeling is the *design effect*, which indicates how the particular design chosen – in our case, the multilevel design – affects the standard error of the parameters. It is defined as

$$deff = \frac{\text{squared standard error under this design}}{\text{squared standard error under standard design}}, \quad (3)$$

where the ‘standard design’ is defined as a design using a simple random sample with the same total sample size at level one. (The determination of sample sizes under simple random sample designs is treated in the article in **Sample Size and Power Calculation**) If *deff* is greater than 1, the multilevel design is less efficient than a simple random sample design (with the same sample size); if it is less than 1, the multilevel design is more efficient. Since squared standard errors are inversely proportional to sample sizes, the required sample size for a multilevel design will be given by the sample size that would be required for a simple random sample design, multiplied by the design effect.

The formulas for the basic cases are given here (also see [11]) for two-level designs, where the cluster size is assumed to be constant, and denoted by n . These are good approximations also when the cluster sizes are variable but not too widely different. The number of level-two units is denoted m , so the total sample size at level one is mn . In all models mentioned below, the level-one residual variance is denoted $\text{var}(R_{ij}) = \sigma^2$ and the level-two residual variance by $\text{var}(U_{0j}) = \tau^2$.

1. For estimating a population mean μ in the model

$$Y_{ij} = \mu + U_{0j} + R_{ij}, \quad (4)$$

the estimation variance is

$$\text{var}(\hat{\mu}) = \frac{n\tau^2 + \sigma^2}{mn}. \quad (5)$$

The design effect is $deff = 1 + (n-1)\rho_1 \geq 1$, where ρ_1 is the intraclass correlation, defined by $\rho_1 = \tau^2/(\sigma^2 + \tau^2)$.

2. For estimating the regression coefficient γ_1 of a level-one variable X_1 in the model

$$Y_{ij} = \gamma_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \cdots + \gamma_p X_{pij} + U_{0j} + R_{ij}, \quad (6)$$

where it is assumed that X_1 does not have a random slope and has zero between-group variation, that is, a constant group mean, and that it is uncorrelated with any other explanatory variables X_k ($k \geq 2$), the estimation variance is

$$\text{var}(\hat{\gamma}_1) = \frac{\sigma^2}{mns_{X1}^2}, \quad (7)$$

where the within-group variance s_{X1}^2 of X_1 also is assumed to be constant. The design effect here is $deff = 1 - \rho_1 \leq 1$.

3. For estimating the effect of a level-one variable X_1 under the same assumptions except that X_1 now does have a random slope,

$$Y_{ij} = \gamma_0 + (\gamma_1 + U_{1j})X_{1ij} + \gamma_2 X_{2ij} + \cdots + \gamma_p X_{pij} + U_{0j} + R_{ij}, \quad (8)$$

where the random slope variance is τ_1^2 , the estimation variance of the fixed effect is

$$\text{var}(\hat{\gamma}_1) = \frac{n\tau_1^2 s_{X1}^2 + \sigma^2}{mns_{X1}^2}, \quad (9)$$

with design effect

$$deff = \frac{n\tau_1^2 s_{X1}^2 + \sigma^2}{\tau_1^2 s_{X1}^2 + \tau^2 + \sigma^2}, \quad (10)$$

which can be greater than or less than 1.

4. For estimating the regression coefficient of a level-two variable X_1 in the model

$$Y_{ij} = \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2ij} + \cdots + \gamma_p X_{pij} + U_{0j} + R_{ij}, \quad (11)$$

where the variance of X_1 is s_{X1}^2 , and X_1 is uncorrelated with any other variables X_k ($k \geq 2$),

the estimation variance is

$$\text{var}(\hat{\gamma}_1) = \frac{n\tau^2 + \sigma^2}{mns_{X_1}^2}, \quad (12)$$

and the design effect is $deff = 1 + (n - 1)\rho_1 \geq 1$.

This illustrates that multilevel designs sometimes are more, and sometimes less, efficient than simple random sample designs. In case 2, a level-one variable without between-group variation, the multilevel design is always more efficient. This efficiency of within-subject designs is a well-known phenomenon. For estimating a population mean (case 1) or the effect of a level-two variable (case 4), on the other hand, the multilevel design is always less efficient, and more seriously so as the cluster size and the **intraclass correlation** increase.

In cases 2 and 3, the same type of regression coefficient is being estimated, but in case 3, variable X_1 has a random slope, unlike in case 2. The difference in $deff$ between these cases shows that the details of the multilevel dependence structure matters for these standard errors, and hence for the required sample sizes. It must be noted that what matters here is not how the researcher specifies the multilevel model, but the true model. If in reality there is a positive random slope variance but the researcher specifies a random intercept model without a random slope, then the true estimation variance still will be given by the formula above of case 3, but the standard error will be misleadingly estimated by the formula of case 2, usually a lower value.

For more general cases, where there are several correlated explanatory variables, some of them having random slopes, such clear formulas are not available. Sometimes, a very rough estimate of required sample sizes can still be made on the basis of these formulas. For more generality, however, the program **PinT** (see below) can be used to calculate standard errors in rather general two-level designs.

A general procedure for estimating power and standard errors for any parameters in arbitrary designs is by Monte Carlo simulation of the model and the estimates. This is described in [4, Chapter 10].

For further reading, general treatments can be found in [1, 2], [4, Chapter 10], [11], and [13, Chapter 10]. More specific sample size and design issues, often focusing on two-group designs, are treated in [3, 5, 6, 7, 8, 9, 10].

Computer Programs

ACluster calculates required sample sizes for various types of cluster randomized designs, not only for continuous but also for binary and time-to-event outcomes, as described in [1].

See website www.update-software.com/Acluster.

OD (Optimal Design) calculates power and optimal sample sizes for testing treatment effects and variance components in multisite and cluster randomized trials with balanced two-group designs, and in repeated measurement designs, according to [8, 9, 10].

See website <http://www.ssicentral.com/other/hlmod.htm>.

PinT (Power in Two-level designs) calculates standard errors of regression coefficients in two-level designs, according to [12]. With extensive manual.

See website <http://stat.gamma.rug.nl/snijders/multilevel.htm>.

RMAS2 calculates the sample size for a two-group repeated measures design, allowing for attrition, according to [3].

See website <http://tigger.uic.edu/~hedeker/works.html>.

References

- [1] Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*, Arnold Publishing, London.
- [2] Hayes, R.J. & Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials, *International Journal of Epidemiology* **28**, 319–326.
- [3] Hedeker, D., Gibbons, R.D. & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups, *Journal of Educational and Behavioral Statistics* **24**, 70–93.
- [4] Hox, J.J. (2002). *Multilevel Analysis, Techniques and Applications*, Lawrence Erlbaum Associates, Mahwah.
- [5] Moerbeek, M., van Breukelen, G.J.P. & Berger, M.P.F. (2000). Design issues for experiments in multilevel populations, *Journal of Educational and Behavioral Statistics* **25**, 271–284.
- [6] Moerbeek, M., van Breukelen, G.J.P. & Berger, M.P.F. (2001a). Optimal experimental designs for multilevel logistic models, *The Statistician* **50**, 17–30.
- [7] Moerbeek, M., van Breukelen, G.J.P. & Berger, M.P.F. (2001b). Optimal experimental designs for multilevel

4 Power and Sample Size in Multilevel Linear Models

- models with covariates, *Communications in Statistics, Theory and Methods* **30**, 2683–2697.
- [8] Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials, *Psychological Methods* **2**, 173–185.
- [9] Raudenbush, S.W. & Liu, X.-F. (2000). Statistical power and optimal design for multisite randomized trials, *Psychological Methods* **5**, 199–213.
- [10] Raudenbush, S.W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change, *Psychological Methods* **6**, 387–401.
- [11] Snijders, T.A.B. (2001). Sampling, in *Multilevel Modelling of Health Statistics*, Chap. 11, A. Leyland & H. Goldstein, eds, Wiley, Chichester, pp. 159–174.
- [12] Snijders, T.A.B. & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research, *Journal of Educational Statistics* **18**, 237–259.
- [13] Snijders, T.A.B. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, London.

TOM A.B. SNIJDERS

Prediction Analysis of Cross-Classifications

KATHRYN A. SZABAT

Volume 3, pp. 1573–1579

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Prediction Analysis of Cross-Classifications

Prediction analysis [4–6, 8, 11–13, 15] is an approach to the analysis of **cross-classified**, **cross-sectional**, or longitudinal (*see Longitudinal Data Analysis*) data. It is a method for predicting events based on specification of predicted relations among qualitative variables and includes techniques for both stating and evaluating those event predictions. Event predictions are based on hypothesized relations among predictor (independent) and criterion (dependent) variables. Event predictions predict for each case what state of a criterion a variable follows from a given state of a predictor variable. States are categories of qualitative variables. Hypotheses can be directed at relations among variables at a given point in time (as in cross-sectional designs) as well as at the extent to which those relations change over time (as in longitudinal designs). Prediction analysis is performed on sets of predictions. It allows for the evaluation of the overall set of hypothesized event predictions as well as the evaluation of each hypothesized event prediction in the set.

Examples

Table 1 reproduces a data set presented by Trautner, Geravi, and Nemeth [9] in their study of appearance-reality (AR) distinction and development of gender constancy (GC) understanding in children. The data can be used to evaluate the prediction that AR distinction is a necessary prerequisite of the understanding of GC. Performance (1 = failed, 2 = passed) on both an appearance-reality distinction task (ART) and a gender constancy test (GCT) was recorded for each child in the study. The prediction can be expressed as *performance on the ART predicts performance on the GCT*. Trautner et al. report that there was a strong association between ability to distinguish appearance from reality and GC understanding (Cohen kappa = 0.76, $p < 0.0001$).

Table 2 reproduces a data set presented by Casey [1] in her study of selective attention abilities. The data can be used to evaluate the prediction of agreement between levels of two tasks: a mirror-image task and a shape-detail task. According to

Table 1 Performance on two tests: appearance–reality distinction Task (ART) and gender constancy test (GCT). Reproduced from Trautner, H.M., Geravi, J. & Nemeth, R. (2003). Appearance-reality distinction and development of gender constancy understanding in children, *International Journal of Behavioral Development* 27(3), 275–283. Reprinted by permission of the International Society for the Study of Behavioral Development (<http://www.tandf.co.uk/journals>), Taylor & Francis, & Judit Gervai

X = ART				
Y = GCT	Failed	Failed	Passed	
		31	5	36
	Passed	4	35	39
		35	40	75

Table 2 Level on two tasks: mirror image (MI) and shape–detail (SD). Reproduced from Casey, M.B. (1986). Individual differences in selective attention among pre-readers: a key to mirror-image confusions, *Development Psychology* 22, 58–66. Copyright © 1986 by the American Psychological Association. Reprinted with permission of the APA and M.B. Casey

X = MI Level				
Y = SD Level		1	2	3
	1	9	1	0
	2	3	7	2
	3	1	4	1
		13	12	3

Kinsbourne’s maturational theory, the shape-detail problem would be predicted to show equivalent level effects with the mirror-image problem. Casey reports that findings support Kinsbourne’s selective attention theory.

Prediction Analysis: The Hildebrand et al. Approach

Prediction analysis of cross-classifications was formally introduced by Hildebrand, Laing, and Rosenthal [4–6]. Recognizing the limitations of classical cross-classification analysis techniques, Hildebrand et al. developed a prediction analysis approach as an alternative method. The prediction analysis approach is concerned with predicting a type of relation and

2 Prediction Analysis of Cross-Classifications

then measuring the success of that prediction. The method involves *prediction logic*, expressing predictions that relate qualitative variables and the *del* measure used to analyze data from the perspective of the prediction of interest. The prediction analysis method can be applied to one-to-one (one predictor state to one criterion state) and one-to-many (one predictor state to many criterion states) predictions about variables, to bivariate and multivariate settings, to predictions stated *a priori* (before the data are analyzed) and selected *ex post* (after the data are analyzed), to single observations (degree-1 predictions) and observation pairs (degree-2 predictions), and to pure (absolute predictions) and mixed (actuarial predictions) strategies. Thus, the prediction analysis approach should cover a wide range of research possibilities.

Here, we present prediction analysis for bivariate, pure, and degree-1 predictions stated *a priori*. For comprehensive, detailed coverage of the Hildebrand et al. approach, readers are referred to the Hildebrand, Laing, and Rosenthal book, *Prediction Analysis of Cross-classifications*.

Specification of Prediction Analysis Hypotheses

The prediction analysis approach starts with the specification of event predictions about qualitative variables. The language for stating such predictions, called *prediction logic*, is defined by a formal analogy to elementary logic. A prediction logic proposition is stated precisely by specifying its domain and the set of error events it identifies. The domain of a prediction logic proposition includes all events in the cross-classification of the predictor and criterion qualitative variables. The set of error events contains events that are not consistent with a prediction.

von Eye and Brandtstadter [11–13, 15] describe how to use statement calculus, from which prediction logic is derived, when formulating hypotheses in prediction analysis. Prediction hypotheses within a set of predictions are stated in terms of ‘if-then’ variable relationships. Predictor states are linked to criterion states using an implication operator ($\rightarrow$), often read ‘implies’ or ‘predicts.’ Multiple predictors or multiple criteria are linked using the conjunction operator ($\wedge$), to be read ‘and,’ or the adjunction operator ($\vee$), to be read ‘or.’ These logical operators can be applied to both dichotomous (two state) and polytomous (more than two state) variables.

A set of predictions, H , contains one or more hypothesized predictions. Each prediction within a set is called an *elementary prediction*. Elementary predictions can be made to link (a) a single predictor state to a single criterion state, (b) a pattern of predictor states to a single criterion state, (c) a single predictor state to a pattern of criterion states, or (d) a pattern of predictor states to a pattern of criterion states. Patterns of states can arise from the same variable or from different variables in the case of multiple predictors or multiple criteria [15]. Within a set of predictions, it is not necessary to make a specific prediction for each predictor state or pattern of states. However, predictions, whether elementary or as a set, must be nontautological (must not describe hypotheses that cannot be disconfirmed). In addition, a set of predictions cannot contain contradictory elementary predictions.

Recall the Trautner et al. study. The set of elementary predictions for the hypothesized relation between performance on the ART and performance on the GCT, expressed using statement calculus, is as follows.

Let X be the predictor, performance on the ART, with two states (1 = failed, 2 = passed). Let Y be the criterion, performance on the GCT, with two states (1 = failed, 2 = passed). Then the set of prediction hypotheses, H , is $x_1 \rightarrow y_1, x_2 \rightarrow y_2$.

Recall the Casey study. The set of elementary predictions for the hypothesized agreement between the levels of the mirror image and shape-detail tasks, expressed using statement calculus, is as follows.

Let X be the predictor, level on the mirror-image task, with three states (1 = nonlearner, 2 = instructed learner, 3 = spontaneous responder). Let Y be the criterion, level on the shape-detail task, with three states (1 = nonlearner, 2 = instructed learner, 3 = spontaneous responder). Then the set of hypothesized predictions, H , is $x_1 \rightarrow y_1, x_2 \rightarrow y_2, x_3 \rightarrow y_3$.

A Proportionate Reduction in Error Measure for Evaluating Prediction Success

Hildebrand et al. use a PRE model to define *del*, the measure of prediction success. On the basis of the stated prediction, cells of a resultant cross-classification are designated as either *predicted cells* (cells that represent, for a given predictor state, the criterion state(s) that are predicted) or error cells (cells that represent, for a given predictor state, the criterion state(s) that are not predicted).

Defining w_{ij} as the error cell indicator, (1 if cell ij is an error cell; 0 if otherwise); then, the population measure of prediction success, ∇ (del), is calculated as

$$\nabla = 1 - \frac{(\sum_i \sum_j w_{ij} P_{ij})}{(\sum_i \sum_j w_{ij} P_{i.} P_{.j})} \quad (1)$$

where $P_{i.}$ and $P_{.j}$ are marginal probabilities indicating the probability that an observation lies in the i th row and j th column respectively, of the $Y \times X$ cross-classification and P_{ij} is the cell probability representing the probabilities that an observation belongs to states Y_i and X_j .

We define, above, an error weight value of 1 for all error cells. However, in situations where some errors are regarded as more serious or more important than others, differential weighting of errors is allowed.

By equation (1), ∇ represents a PRE definition of prediction success. As such, ‘ ∇_p measures the PRE in the number of cases observed in the set of error cells for P relative to the number expected under statistical independence, given knowledge of the actual probability structure’ [6]. The value of ∇ may vary in the range, $-\infty$ to 1. If ∇ is positive, then it measures the PRE afforded by the prediction. A negative value of ∇ indicates that the specified prediction is grossly incorrect about the nature of the relation. If the variables are statistically independent, then ∇ is zero (although, the converse does not hold: $\nabla = 0$ does not imply that two variables are statistically independent).

When comparing several competing theoretical predictions to determine the dominance of a particular prediction, one needs to consider *precision* in addition to the value of ∇ . Precision is determined, to a large extent, by the number of criterion states that are predicted for given predictor states. The overall precision of a prediction, U , is defined as

$$U = \sum_i \sum_j w_{ij} P_{i.} P_{.j} \quad (2)$$

Larger values of U represent greater precision ($0 \leq U \leq 1$). One prediction dominates another when the prediction has higher prediction success and at least as great prediction precision than the other, or a higher prediction precision and at least as great a prediction success [6].

A thorough analysis of prediction success would consider performance of each elementary prediction. The overall success of a prediction, ∇ , can be

expressed as the weighted average of the success achieved by the elementary predictions, $\nabla_{.j}$. That is,

$$\nabla = \sum_j \left[\frac{P_{.j} \sum_i w_{ij} P_{i.}}{\sum_j P_{.j} \sum_i w_{ij} P_{i.}} \right] \nabla_{.j} \quad (3)$$

By equation (3), ‘the greater the precision of the elementary prediction relative to the precision of the other elementary predictions, the greater the contribution of the elementary prediction’s success to the overall measure of prediction success’ [6].

Recall the Trautner et al. study. To measure the success of the prediction that performance on the ART predicts performance on the GCT, we identify the error cells, indicated as shaded cells in Table 1. Using (1), the calculated measure of overall prediction success, ∇ , is 0.76. This value indicates that a PRE of 76.0% is achieved in applying the prediction to children with known ART performance over that which is expected when the prediction is applied to a random selection of children whose ART performance is unknown. Furthermore, the calculated measure of prediction success for the first and second elementary predictions is 0.7804 and 0.7395 respectively. When weighted by their respected precision weights, one sees that each elementary prediction contributes about equally to the overall prediction success (0.37979 and 0.37961 respectively). We acknowledge that the value of del (0.76) is the same as the value of the Cohen’s kappa (k) measure reported in the Trautner et al. study. Hildebrand et al. note the equivalence of kappa to del for any square table given the proposition predicts the main diagonal in the $Y \times X$ cross-classification.

von Eye and Brandtstadter [11] present an alternative measure for evaluating predictions. Their measure of predictive efficiency (PE) focuses on predicted cells as opposed to error cells and, by definition, excludes events for which no prediction was made. ‘Del and PE are equivalent as long as a set of predictions contains only nontautological and admissible partial (elementary) predictions’ [12].

Statistical Inference

In the previous section, we presented ∇ as a population parameter. That is, we evaluated predictions with observations that represent an entire population. In practice, researchers almost always work with a sample taken from the population. As such, inference

is of concern. Earlier, we mentioned that the focus of this presentation is on predictions stated *a priori*. The distinction between *a priori* and *ex post* predictions is critical here in the statistical inference process because the statistical theory is different, dependent on whether the prediction is stated before or after the data are analyzed. We continue to focus on *a priori* predictions. A discussion on *ex post* analysis can be found in Hildebrand [6] and Szabat [8].

Hildebrand et al. [6] present bivariate statistical inference under three sample schemes that cover the majority of research applications. Given the specific sampling condition, the sample estimate of ∇ is defined by substituting observed sample proportions for any unknown population probabilities. The sampling distributions of ∇ estimators are based on standard methods of asymptotic theories; the estimators, ∇ -hat, are approximately normally distributed. Hildebrand et al. [6] provide formulas for the approximate variance of the ∇ estimators for each of the three sampling schemes. The variance calculations are quite tedious by hand, but can be programmed for computer calculation quite easily (see [14, 10]).

Given the sampling distribution of the ∇ estimators, it is possible to calculate P values for hypothesis testing. The form of the test statistic for testing the hypothesis that $\nabla > 0$ is

$$Z = \frac{(\nabla\text{-hat})}{[\text{est. var.}(\nabla\text{-hat})]^{1/2}} \quad (4)$$

The standard condition for applying the normal distribution to the binomial distribution, $5 \leq nU(1-\nabla) \leq n-5$, is the essential condition for the application of the normal approximation.

Recall the Trautner et al. study. Suppose we hypothesize the true value of ∇ to be greater than zero. Under sampling condition S2, the value for the estimate of ∇ , 0.76, yields a test statistic, $Z = 10.43$, which is compared with the normal probability table value, 1.65. We conclude at the 5% significance level that the true value of ∇ is statistically greater than zero.

Prediction Analysis: The von Eye et al. Approach

von Eye, Brandtstadter, and Rovine [12, 13] present nonstandard log-linear modeling as an alternative to the Hildebrand et al. approach. Rather than measuring prediction success of a set of predictions via

a percent reduction in error measure, the von Eye et al. approach models the prediction hypothesis and evaluates the extent to which the model adequately describes the data. This is achieved via nonhierarchical log-linear models that specifies the relationship between predictors and criteria (base model) as well as specifies the event predictions. This approach to prediction analysis allows for consideration of a wide range of models and prediction hypotheses. Model specification results from an understanding of the relation between logical formulation of variable relationship and base model for estimation of expected cell frequencies. For an extensive coverage of different log-linear base models for different prediction models, the reader is referred to *Statistical Analysis of Categorical Data in the Social and Behavioral Sciences* [15]. Here, we present prediction analysis for one type of base model, a base model that is saturated in both the predictors and criteria, and, like all base models for prediction analysis, assumes independence between predictors and criteria. This base model corresponds to the standard hierarchical log-linear model that underlies the estimation of expected cell frequencies in the Hildebrand et al. approach. In addition, the model involves one predictor and one criterion.

Specification of Prediction Analysis Hypotheses

As is the case for prediction analysis in the Hildebrand et al. tradition, the von Eye et al. approach starts with the formulation of prediction hypotheses. One uses prediction logic or statement calculus to specify a set of nontautological and noncontradictory elementary predictions. The process is described above.

The Nonstandard Log-linear Model for Evaluating Prediction Success

von Eye et al. use nonstandard **log-linear modeling** to evaluate prediction success. On the basis of the set of predictions, cells of the resultant cross-classification are identified as predicted or 'hit' cells (cells that confirm the prediction), error cells (cells that disconfirm the prediction), or irrelevant cells (cells that neither confirm nor disconfirm the prediction). A log-linear model is then specified. The form of the log-linear model to test prediction analysis hypotheses is derived by applying special design

matrices to the general form of the log-linear model,

$$\log(\mathbf{F}) = \mathbf{X}\lambda + \mathbf{e} \quad (5)$$

where $\mathbf{F}$ is the vector of frequencies, $\mathbf{X}$ is the indicator matrix, λ is the parameter vector, and $\mathbf{e}$ is the residual vector. The specially designed indicator matrix needed for prediction analysis contains both indicator variables required for the model to be saturated in both predictors and criteria, and indicator variables reflecting prediction hypotheses. Indicator variables that reflect the prediction hypotheses are defined as follows: 1 if the event confirms the elementary prediction, -1 if the event disconfirms the elementary prediction, and 0 if else. This nonstandard log-linear model approach explicitly contrasts predicted cells and error cells, which allows one to discriminate X to Y and Y to X predictions, a feature not possible with the Hildebrand et al. approach.

Once the model is specified, estimation of the nonstandard log-linear model parameters and expected cell frequencies occurs. There are a few major statistical software packages that can assist in this effort, namely, S+, SYSTAT, LEM, SPSS, and CDAS, (see [15] and **Software for Statistical Analyses**). Evaluation of the model is then assessed via statistical testing.

Recall the Casey study. The design matrix for nonstandard log-linear modeling is given in Table 3. The design matrix contains two groups of indicator variables. The first four vectors contain two main effect variables each for the predictor and criterion variable. The next three vectors are the three coding variables that result for the three elementary predictions in the set.

Table 3 Design matrix for nonstandard log-linear modeling: Casey Study

Variables	Design matrix						
XY							
11	1	0	1	0	1	0	0
12	0	1	1	0	-1	0	0
13	-1	-1	1	0	-1	0	0
21	1	0	0	1	0	-1	0
22	0	1	0	1	0	1	0
23	-1	-1	0	1	0	-1	0
31	1	0	-1	-1	0	0	-1
32	0	1	-1	-1	0	0	-1
33	-1	-1	-1	-1	0	0	1

Statistical Inference

As mentioned above, statistical testing is used to evaluate the model. Statistical evaluation occurs at two levels: (a) statistical evaluation of the overall hypothesis, via a goodness of fit test (likelihood ratio χ^2 or Pearson χ^2 (see **Goodness of Fit for Categorical Variables**), and (b) statistical evaluation of the each elementary hypothesis, via a test for significant parameter effect (z test). According to von Eye et al. [12], if statistical evaluation indicates overall model fit and at least one elementary prediction is successful (statistically significant), then the prediction set is considered successful. If statistical evaluation indicates some elementary predictions are successful, but the model does not fit, then parameter estimation is shaky. If statistical evaluation indicates overall model fit and not one of the elementary predictions are successful, then one assumes that the attempt at parameterization of deviations from independence failed.

With this approach, the number of elementary predictions is more severely limited than with the Hildebrand et al. approach. There are no more than $c - X - Y + 1$ degrees of freedom available for prediction analysis given c cells from X predictor and Y criterion states. Given a 2×2 table with 4 cells, one has $4 - 2 - 2 + 1 = 1$ degree of freedom left for prediction analysis, 'This excludes these tables from PA using the present approach because inserting an indicator variable implies a saturated model.' [12].

Recall the Casey study. Statistical evaluation shows that the log-linear *base model* does not adequately describe the data ($LR - \chi^2 = 13.742$, $p = 0.008$ and Pearson $\chi^2 = 12.071$, $p = 0.017$). When the parameters for the prediction hypotheses are added to the base model, statistical results indicate a good fit ($LR - \chi^2 = 0.273$, $p = 0.601$ and Pearson $\chi^2 = 0.160$, $p = 0.690$). However, only one of the parameters (the first one) for the predication analysis part of the model is statistically significant ($z = 2.665$, $p = 0.004$). Still, statistical evaluation indicates overall model fit and at least one elementary event is successful, therefore, we consider the prediction set successful.

Discussion

Above we present two approaches to prediction analysis of cross-classifications. In both approaches, formulation of prediction hypotheses is the same. Both

approaches evaluate sets of elementary predictions statistically, and both base evaluation of prediction success on deviations from an assumption of some form of independence between predictors and criteria [12]. The two approaches differ, however, with respect to hypotheses tested (that is, the meaning assigned to the link between predictors and criteria), model independence assumption, and method applied for statistical evaluation.

Two other prediction approaches warrant mention. Hubert [7] proposed use of matching models, instead of log-linear models, for evaluating prediction success. The approach provides an alternative method for constructing hypothesis tests that may require very cumbersome formulas; therefore, the approach is not presented here. Goodman and Kruskal [2, 3] proposed **quasi-independence** log-linear modeling for prediction analysis. In this approach, error cells are declared **structural zeros**, which result in loss of degrees of freedom. This severely limits the type of hypotheses one can test; therefore, the approach is not presented here.

Summary

Researchers in the social and behavioral sciences are often faced with investigations that concern variables measured in terms of discrete categories. The basic research design, whether cross-sectional or longitudinal in nature, for such studies involves the cross-classification of each subject with respect to relevant qualitative variables. Prediction analysis of cross-classifications approaches as developed by Hildebrand et al. and von Eye et al. provide a data analysis tool that allows researchers to statistically test custom-tailored sets of hypothesized predictions about the relation between such variables. Prediction analysis is a viable, valuable tool that researchers in the social and behavioral sciences can use in a wide range of research settings.

References

- [1] Casey, M.B. (1986). Individual differences in selective attention among prereaders: a key to mirror-image confusions, *Development Psychology* **22**, 58–66.
- [2] Goodman, L.A. & Kruskal, W.H. (1974a). Empirical evaluation of formal theory, *Journal of Mathematical Sociology* **3**, 187–196.
- [3] Goodman, L.A. & Kruskal, W.H. (1974b). More about empirical evaluation of formal theory, *Journal of Mathematical Sociology* **3**, 211–213.
- [4] Hildebrand, D.K., Laing, J.D. & Rosenthal, H. (1974a). Prediction logic: a method for empirical evaluation for formal theory, *Journal of Mathematical Sociology* **3**, 163–185.
- [5] Hildebrand, D.K., Laing, J.D. & Rosenthal, H. (1974b). Prediction logic and quasi-independence in empirical evaluation for formal theory, *Journal of Mathematical Sociology* **3**, 197–209.
- [6] Hildebrand, D.K., Laing, J.D. & Rosenthal, H. (1977). *Prediction Analysis of Cross-Classifications*, Wiley, New York.
- [7] Hubert, L.J. (1979). Matching models in the analysis of cross-classifications, *Psychometrika* **44**, 21–41.
- [8] Szabat, K.A. (1990). Prediction analysis, in *Statistical Methods in Longitudinal Research*, Vol. 2, A. von Eye, ed., Academic Press, San Diego, pp. 511–544.
- [9] Trautner, H.M., Geravi, J. & Nemeth, R. (2003). Appearance-reality distinction and development of gender constancy understanding in children, *International Journal of Behavioral Development* **27**(3), 275–283.
- [10] von Eye, A. (1997). Prediction analysis program for 32-bit operation systems, *Methods for Psychological Research-Online* **2**, 1–3.
- [11] von Eye, A. & Brandtstadter, J. (1988). Formulating and testing developmental hypotheses using statement calculus and nonparametric statistics, in *Life-Span Development and Behavior*, Vol. 8, P.B. Bates, D. Featherman & R.M. Lerner, eds, Lawrence Erlbaum, Hillsdale, pp. 61–97.
- [12] von Eye, A., Brandtstadter, J. & Rovine, M.J. (1993). Models for prediction analysis, *Journal of Mathematical Sociology* **18**, 65–80.
- [13] von Eye, A., Brandtstadter, J. & Rovine, M.J. (1998). Models for prediction analysis in longitudinal research, *Journal of Mathematical Sociology* **22**, 355–371.
- [14] von Eye, A. & Krampen, G. (1987). BASIC programs for prediction analysis of cross-classifications, *Educational and Psychological Measurement* **47**, 141–143.
- [15] von Eye, A. & Niedermeier, K.E. (1999). *Statistical Analysis of Longitudinal Categorical Data in the Social and Behavioral Sciences*, Lawrence Erlbaum, Hillsdale.

KATHRYN A. SZABAT

Prevalence

HANS GRIETENS

Volume 3, pp. 1579–1580

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Prevalence

Prevalence is defined as the proportion of individuals in a population who have a particular disease. It is computed by dividing the total number of cases by the total number of individuals in the population. For instance, the prevalence rate per 1000 is calculated as follows:

$$1000 \left(\frac{\text{Number of cases ill at one point in time}}{\text{Size of the population exposed to risk at that time}} \right)$$

Prevalence rates can be studied in general or in clinical populations and have to be interpreted as morbidity rates. In a general population study, the rates can be considered as base rates or baseline data. Studying the prevalence of diseases is a main aim of epidemiology, a subdiscipline of social medicine [6, 7], which has found its way into the field of psychiatry and behavioral science [5]. In these fields, prevalence refers to the proportion of individuals in a population who have a specified mental disorder (e.g., major depressive disorder) or who manifest specific behavioral problems (e.g., physical aggression). In a prevalence study, individuals are assessed at time T to determine whether they are manifesting a certain outcome (e.g., disease, mental disorder, behavioral problem) at that time [4]. Prevalences can easily be estimated by a single survey. Such surveys are useful to determine the needs for services in a certain area. Further, they permit study of variations in the prevalence of certain outcomes over time and place [2]. Prevalence rates are determined by the **incidence** and the duration of a disease. If one intends to determine whether an individual is at risk for a certain disease, incidence studies are to be preferred over prevalence studies. Prevalence rates may be distorted by mortality factors, treatment characteristics, and effects of preventive measures.

According to the period of time that is taken into account, three different forms of prevalence can be defined: (1) point prevalence is the prevalence of a disease at a certain point in time, (2) period prevalence is the prevalence at any time during a certain time period, and (3) lifetime prevalence refers to the presence of a disease at any time during a person's life. Most prevalence studies on behavioral problems

or mental disorders present period prevalence rates. This is, for instance, the case in all studies using the rating scales of Achenbach's System of Empirically Based Assessment [1] to measure behavior problems in 1.5-to-18-year old children. Depending on the child's age, these rating scales cover problem behaviors within a two-to-six month period. Examples of studies on adolescents and adults using lifetime prevalence rates of psychiatric disorders are the Epidemiological Catchment Area Study [8] and the National Comorbidity Survey [3], both conducted in the United States.

Prevalences rely on the identification of 'truly' disordered individuals. The identification of true cases is only possible by using highly predictive diagnostic criteria. To measure these criteria, one needs to have instruments that are both highly sensitive (with the number of false negatives being as low as possible) and highly specific (with the number of false positives being as low as possible).

References

- [1] Achenbach, T.M. & Rescorla, L.E. (2001). *Manual for the ASEBA School-Age Forms & Profiles*, University of Vermont, Research Center for Children, Youth, & Families, Burlington.
- [2] Cohen, P., Slomkowski, C. & Robins, L.N. (1999). *Historical and Geographical Influences on Psychopathology*, Lawrence Erlbaum, Mahwah.
- [3] Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., Hughes, M., Eshleman, S., Wittchen, H.U. & Kendler, K.S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the national comorbidity survey, *Archives of General Psychiatry* **51**, 8–19.
- [4] Kraemer, H.C., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S. & Kupfer, D.J. (1997). Coming to terms with the terms of risk, *Archives of General Psychiatry* **54**, 337–343.
- [5] Kramer, M. (1957). A discussion of the concept of incidence and prevalence as related to epidemiological studies of mental disorders, *Epidemiology of Mental Disease* **47**, 826–840.
- [6] Lilienfeld, M. & Stolley, P.D. (1994). *Foundations of Epidemiology*, Oxford University Press, New York.
- [7] MacMahon, B. & Trichopoulos, D. (1996). *Epidemiology Principles and Methods*, Little, Brown and Company, Boston.
- [8] Robins, L.N. & Regier, D.A., eds (1991). *Psychiatric Disorders in America*, Free Press, New York.

HANS GRIETENS

Principal Component Analysis

IAN JOLLIFFE

Volume 3, pp. 1580–1584

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Principal Component Analysis

Large datasets often include measurements on many variables. It may be possible to reduce the number of variables considerably while still retaining much of the information in the original dataset. A number of dimension-reducing techniques exist for doing this, and principal component analysis is probably the most widely used of these. Suppose we have n measurements on a vector $\mathbf{x}$ of p random variables, and we wish to reduce the dimension from p to q . Principal component analysis does this by finding linear combinations, $\mathbf{a}_1'\mathbf{x}$, $\mathbf{a}_2'\mathbf{x}$, $\dots$, $\mathbf{a}_q'\mathbf{x}$, called *principal components*, that successively have maximum variance for the data, subject to being uncorrelated with previous $\mathbf{a}_k'\mathbf{x}$ s. Solving this maximization problem, we find that the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\dots$, $\mathbf{a}_q$ are the **eigenvectors** of the **covariance matrix**, $\mathbf{S}$, of the data, corresponding to the q largest **eigenvalues**. These eigenvalues give the variances of their respective principal components, and the ratio of the sum of the first q eigenvalues to the sum of the variances of all p original variables represents the proportion of the total variance in the original dataset, accounted for by the first q principal components.

This apparently simple idea actually has a number of subtleties, and a surprisingly large number of uses. It was first presented in its algebraic form by **Hotelling** [7] in 1933, though **Pearson** [14] had given a geometric derivation of the same technique in 1901. Following the advent of electronic computers, it became feasible to use the technique on large datasets, and the number and varieties of applications expanded rapidly. Currently, more than 1000 articles are published each year with principal component analysis, or the slightly less popular terminology, principal components analysis, in keywords or title. Henceforth, in this article, we use the abbreviation PCA, which covers both forms. As well as numerous articles, there are two comprehensive general textbooks [9, 11] on PCA, and even whole books on subsets of the topic [3, 4].

An Example

As an illustration, we use an example that has been widely reported in the literature, and which is

Table 1 Vectors of coefficients for the first two principal components for data from [17]

Variable	$\mathbf{a}_1$	$\mathbf{a}_2$
$\mathbf{x}_1$	0.34	0.39
$\mathbf{x}_2$	0.34	0.37
$\mathbf{x}_3$	0.35	0.10
$\mathbf{x}_4$	0.30	0.24
$\mathbf{x}_5$	0.34	0.32
$\mathbf{x}_6$	0.27	-0.24
$\mathbf{x}_7$	0.32	-0.27
$\mathbf{x}_8$	0.30	-0.51
$\mathbf{x}_9$	0.23	-0.22
$\mathbf{x}_{10}$	0.36	-0.33

originally due to Yule et al. [17]. The data consist of scores between 0 and 20 for 150 children aged $4\frac{1}{2}$ to 6 years from the Isle of Wight, on 10 subtests of the Wechsler Pre-School and Primary Scale of Intelligence. Five of the tests were ‘verbal’ tests and five were ‘performance’ tests. Table 1 gives the vectors $\mathbf{a}_1$, $\mathbf{a}_2$ that define the first two principal components for these data.

The first component is a linear combination of the 10 scores with roughly equal weight (max. 0.36, min. 0.23) given to each score. It can be interpreted as a measure of the overall ability of a child to do well on the full battery of 10 tests, and represents the major (linear) source of variability in the data. On its own, it accounts for 48% of the original variability. The second component contrasts the first five scores on the verbal tests with the five scores on the performance tests. It accounts for a further 11% of the total variability. The form of this second component tells us that once we have accounted for overall ability, the next most important (linear) source of variability in the test scores is between those children who do well on the verbal tests *relative to* the performance tests, and those children whose test score profile has the opposite pattern.

Covariance or Correlation

In our introduction, we talked about maximizing variance and eigenvalues/eigenvectors of a covariance matrix. Often, a slightly different approach is adopted in order to avoid two problems. If our p variables are measured in a mixture of units, then it is difficult to interpret the principal components. What do we mean by a linear combination of weight, height, and

2 Principal Component Analysis

temperature, for example? Furthermore, if we measure temperature and weight in degrees Fahrenheit and pounds respectively, we get completely *different principal components* from those obtained from the *same data* but using degrees Celsius and kilograms. To avoid this arbitrariness, we standardize each variable to have zero mean and unit variance. Finding linear combinations of these standardized variables that successively maximize variance, subject to being uncorrelated with previous linear combinations, leads to principal components defined by the eigenvalues and eigenvectors of the **correlation matrix**, rather than the covariance matrix of the original variables. When all variables are measured in the same units, covariance-based PCA may be appropriate, but even here, there can be circumstances in which such analyses are uninformative. This occurs when a few variables have much larger variances than the remainder. In such cases, the first few components are dominated by the high-variance variables and tell us nothing that could not have been deduced by inspection of the original variances. There are certainly circumstances where covariance-based PCA is of interest, but they are not common. Most PCAs encountered in practice are correlation-based. Our example is a case where either approach would be appropriate. The results given above are based on the correlation matrix, but because the variances of all 10 tests are similar, results from a covariance-based analysis would be little different.

How Many Components?

We have talked about q principal components accounting for most of the variation in the p variables. What do we mean by ‘most’, and, more generally, how do we decide how many components to keep? There is a large literature on this topic – see, for example, [11, Chapter 6]. Perhaps, the simplest procedure is to set a threshold, say 80%, and stop when the first q components account for a percentage of total variation greater than this threshold. In our example, the first two components accounted for only 59% of the variation. We would usually want more than this – 70 to 90% are the usual sort of values, but it depends on the context of the dataset, and can be higher or lower. Other techniques are based on the values of the eigenvalues (for example *Kaiser’s rule* [13]) or on the differences between consecutive

eigenvalues (the *scree graph* [1]). Some of these simple ideas as well as more sophisticated ones [11, Chapter 6] have been borrowed from **factor analysis**. This is unfortunate because the different objectives of PCA and factor analysis (see below for more on this) mean that, typically, fewer dimensions should be retained in factor analysis than in PCA, so the factor analysis rules are often inappropriate. It should also be noted that, although it is usual to discard low-variance principal components, they can sometimes be useful in their own right, for example in finding outliers [11, Chapter 10] and in quality control [9].

Normalization Constraints

Given a principal component $\mathbf{a}'_k \mathbf{x}$, we can multiply it by any constant and not change its interpretation. To solve the maximization problem that leads to principal components, we need to impose a normalization constraint, $\mathbf{a}'_k \mathbf{a}_k = 1$. Having found the components, we are free to renormalize by multiplying $\mathbf{a}_k$ by some constant. At least two alternative normalizations can be useful. One that is sometimes encountered in PCA output from computer software is $\mathbf{a}'_k \mathbf{a}_k = l_k$, where l_k is the k th eigenvalue (variance of the k th component). With this normalization, the j th element of $\mathbf{a}_k$ is the correlation between the j th variable and the k th component for correlation-based PCA. The normalization $\mathbf{a}'_k \mathbf{a}_k = 1/l_k$ is less common, but can be useful in some circumstances, such as finding **outliers**.

Confusion with Factor Analysis

It was noted above that there is much confusion between principal component analysis and factor analysis. This is partially caused by a number of widely used software packages treating PCA as a special case of factor analysis, which it most certainly is not. There are several technical differences between PCA and factor analysis [10], but the most fundamental difference is that factor analysis explicitly specifies a model relating the observed variables to a smaller set of underlying unobservable factors. Although some authors [2, 16] express PCA in the framework of a model, its main application is as a descriptive, exploratory technique, with no thought of an underlying model. This descriptive nature means that distributional assumptions are unnecessary to apply PCA in its usual form. It can be used, although an

element of caution may be needed in interpretation, on discrete, and even binary, data, as well as continuous variables. One notable feature of factor analysis is that it is generally a two-stage procedure; having found an initial solution, it is rotated towards *simple structure* (see **Factor Analysis: Exploratory**). The purpose of *factor rotation* (see **Factor Analysis: Exploratory**) is to make the coefficients or loadings relating variables to factors as simple as possible, in the sense that they are either close to zero or far from zero, with few intermediate loadings. This idea can be borrowed and used in PCA; having decided to keep q principal components, we may rotate within the q -dimensional subspace defined by the components in a way that makes the axes as easy as possible to interpret. This is one of a number of techniques that attempt to simplify the results of PCA by postprocessing them in some way, or by replacing PCA with a modified technique [11, Chapter 11].

Uses of Principal Component Analysis

The basic use of PCA is as a dimension-reducing technique whose results are used in a descriptive/exploratory manner, but there are many variations on this central theme. Because the ‘best’ two- (or three-) dimensional representation of a dataset in a least squares sense (see **Least Squares Estimation**) is given by a plot of the first two- (or three-) principal components, the components provide a ‘best’ low-dimensional graphical display of the data. A plot of the components can be augmented by plotting variables as well as observations on the same diagram, giving a **biplot** [6].

PCA is often used as the first step, reducing dimensionality before undertaking another multivariate technique such as cluster analysis (see **Cluster Analysis: Overview**) or **discriminant analysis**. Principal components can also be used in **multiple linear regression** in place of the original variables in order to alleviate problems with *multicollinearity* [11, Chapter 8]. Several dimension-reducing techniques, such as **projection pursuit** [12] and **independent component analysis** [8], which may be viewed as alternatives to PCA, nevertheless, suggest preprocessing the data using PCA in order to reduce dimensionality, before proceeding to the technique of interest. As already noted, there are also occasions when low-variance components may be of interest.

Extensions to Principal Component Analysis

PCA has been extended in many ways. For example, one restriction of the technique is that it is linear. A number of nonlinear versions have, therefore, been suggested. These include the ‘Gifi’ [5] approach to **multivariate analysis**, and various nonlinear extensions that are implemented using **neural networks** [3]. Another area in which many variations have been proposed is when the data are **time series**, so that there is dependence between observations as well as between variables [11, Chapter 12]. A special case of this occurs when the data are functions, leading to *functional data analysis* [15]. This brief review of extensions is by no means exhaustive, and the list continues to grow – see [9, 11].

References

- [1] Cattell, R.B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research* **1**, 245–276.
- [2] Caussinus, H. (1986). Models and uses of principal component analysis: a comparison emphasizing graphical displays and metric choices, in *Multidimensional Data Analysis*, J. de Leeuw, W. Heiser, J. Meulman and F. Critchley eds, DSWO Press, Leiden, pp. 149–178.
- [3] Diamantaras, K.I. & Kung, S.Y. (1996). *Principal Component Neural Networks Theory and Applications*, Wiley, New York.
- [4] Flury, B. (1988). *Common Principal Components and Related Models*, Wiley, New York.
- [5] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester.
- [6] Gower, J.C. & Hand, D.J. (1996). *Biplots*, Chapman & Hall, London.
- [7] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441, 498–520.
- [8] Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*, Wiley, New York.
- [9] Jackson, J.E. (1991). *A User's Guide to Principal Components*, Wiley, New York.
- [10] Jolliffe, I.T. (1998). Factor analysis, overview, in *Encyclopedia of Biostatistics* Vol. 2, P. Armitage & T. Colton, eds, Wiley, New York 1474–1482.
- [11] Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd Edition, Springer, New York.
- [12] Jones, M.C. & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, A* **150**, 1–38. (including discussion).
- [13] Kaiser, H.F. (1960). The application of electronic computers to factor analysis, *Educational and Psychological Measurement* **20**, 141–151.

4 Principal Component Analysis

- [14] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* **2**, 559–572.
- [15] Ramsay, J.O. & Silverman, B.W. (2002). *Applied Functional Data Analysis, Methods and Case Studies*, Springer, New York.
- [16] Tipping, M.E. & Bishop, C.M. (1999). Probabilistic principal component analysis, *Journal of the Royal Statistical Society, B* **61**, 611–622.
- [17] Yule, W., Berger, M., Butler, S., Newham, V. & Tizard, J. (1969). The WPPSI: an empirical evaluation with a British sample, *British Journal of Educational Psychology* **39**, 1–13.

(See also **Multidimensional Scaling**)

IAN JOLLIFFE

Principal Components and Extensions

GEORGE MICHAILIDIS

Volume 3, pp. 1584–1594

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Principal Components and Extensions

Introduction

Principal components analysis (PCA) belongs to the class of projection methods, where one calculates linear combinations of the original variables with the weights chosen so that some measure of ‘interestingness’ is maximized. This measure for PCA corresponds to maximizing the variance of the linear combinations, which, loosely speaking, implies that the directions of maximum variability in the data are those of interest.

PCA is one of the oldest and most popular multivariate analysis techniques; its basic formulation dates back to **Pearson** [7] and in its most common derivation to **Hotelling** [18]. For time series data it corresponds to the Karhunen–Loeve decomposition of Watanabe [30].

In this paper, we primarily discuss extensions of PCA to handle categorical data. We also provide a brief summary of some other extensions that have appeared in the literature over the last two decades. The exposition of the extension of PCA for categorical data takes us into the realm of optimal scoring, and we discuss those connections to some length.

PCA Problem Formulation

The setting is as follows: suppose we have a data matrix X comprising p measurements on N objects. Without loss of generality, it is assumed that the columns have been centered to zero and scaled to have length one. The main idea is to construct $d < p$ linear combinations XW that capture the main characteristics of the data. Let us consider the singular value decomposition of the data matrix

$$X = U \Lambda V', \quad (1)$$

with Λ being a $p \times p$ diagonal matrix with decreasing nonnegative entries, U an $n \times p$ matrix of orthonormal columns and V a $p \times p$ orthogonal matrix. Then, the principal components scores for the objects correspond to the columns of $U \Lambda$.

Some of the properties of this solution are:

- The first $d < p$ columns of the principal components matrix give the linear projection of X into d dimensions with the largest variance.
- Consider the following reconstruction of the data matrix $\tilde{X} = U_d \Lambda_d V_d'$, where only the first d columns/rows of U , Λ , V have been retained. Then, as a consequence of the Eckart–Young theorem [12] $\tilde{X}$ is the best rank d approximation of X in the least squares sense (see **Least Squares Estimation**).

PCA for Nominal Categorical Data

We discuss next a formulation for *nominal* categorical data that leads to a joint representation in Euclidean space of objects and the categories they belong to.

The starting point is a ubiquitous coding of the data matrix. In this setting, data have been collected for N objects (e.g., the sleeping bags described in Appendix 1) on J categorical variables (e.g., price, quality, and fiber decomposition) with k_j categories per variable (e.g., three categories for price, three for quality, and two for fibers).

Let G_j be a matrix with entries $G_j(i, t) = 1$ if object i belongs to category t , and $G_j(i, t) = 0$ if it belongs to some other category, $i = 1, \dots, N$, $t = 1, \dots, k_j$. The collection of all the indicator matrices $G = [G_1 | G_2 | \dots | G_J]$ corresponds to the *super-indicator* matrix [22].

In order to be able to embed both the objects and the categories of the variables in Euclidean space, we need to assign to them scores, X and Y_j respectively. The object scores matrix is of order $N \times d$ and the category scores matrix for variable j is of order $k_j \times d$, where d is usually 2 or 3. The quality of the scoring of the objects and the variables is measured by the squared differences between the two, namely

$$\sigma(X; Y) = J^{-1} \sum_{j=1}^J \text{SSQ}(X - G_j Y_j). \quad (2)$$

The goal is to find object scores X and variable scores $Y = [Y_1, Y_2, \dots, Y_J]'$ that minimize these differences. In order to avoid the trivial solution $X = Y = 0$, we impose the normalization constraint $X'X = NI_d$, together with a centering constraint $u'X = 0$, where u is a $N \times 1$ vector comprised of ones.

2 Principal Components and Extensions

Some algebra (see [22]) shows that the optimal solution takes the form

$$Y_j = D_j^{-1} G_j' X, \quad (3)$$

where D_j contains the frequencies of the categories of variable j along its diagonal, and

$$X = J^{-1} \sum_{j=1}^J G_j Y_j. \quad (4)$$

Equation (3) is the driving force behind the interpretation of the results, since it says that a category score is located at the center of gravity of the objects that belong to it. This is known in the literature as the first centroid principle [22]. Although, (4) implies in theory an analogous principle for the object scores, such a relationship does not hold in practice due to the influence of the normalization constraint $X'X = NI_d$.

Remark This solution is known in the literature as the *homogeneity analysis* solution [11]. Its origins can be traced back to the work of Hirschfeld [17], Fisher [10] and especially Guttman [13], although

some ideas go further back to Pearson (see the discussion in de Leeuw [7]). It has been rediscovered many times by considering different starting points, such as analysis of variance consideration [24] or extensions to correspondence analysis [1]. A derivation of the above solution from a graph theoretic point of view is given in [23].

We summarize next some basic properties of the homogeneity analysis solution (for more details see Michailidis and de Leeuw [22]).

- Category quantifications and object scores are represented in a joint space.
- A category point is the centroid of objects belonging to that category.
- Objects with the same response pattern (identical profiles) receive identical object scores. In general, the distance between two object points is related to the ‘similarity’ between their profiles.
- A variable is more informative to the extent that its category points are further apart.
- If a category applies uniquely to only a single

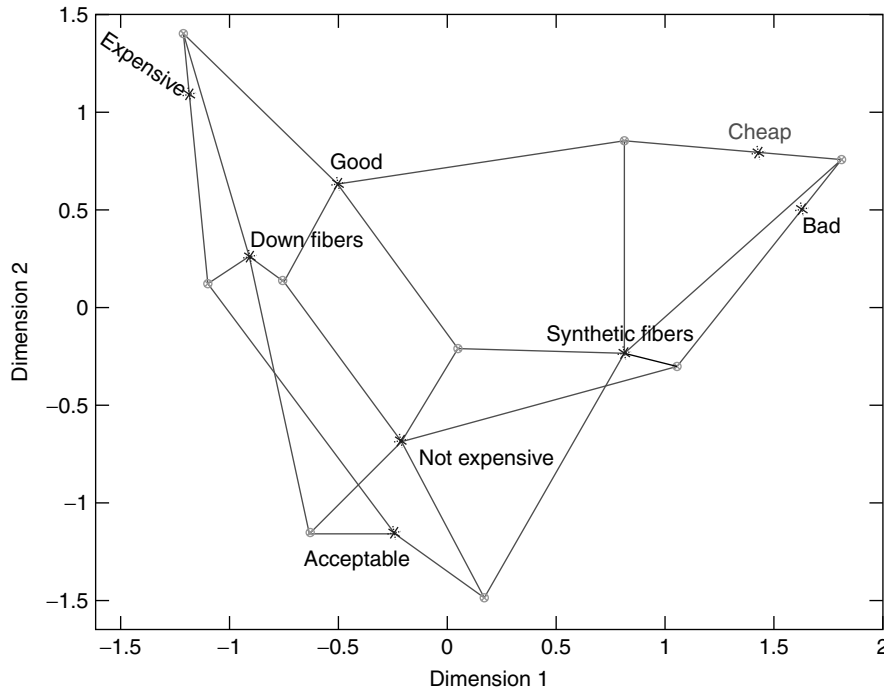


Figure 1 Graph plot of the homogeneity analysis solution of the sleeping bag data

object, then the object point and that category point will coincide.

- Objects with a ‘unique’ profile will be located further away from the origin of the joint space, whereas objects with a profile similar to the ‘average’ one will be located closer to the origin (direct consequence of the previous property).
- The solution is *invariant* under *rotations* of the object scores in p -dimensional space and of the category quantifications. To see this, suppose we select a different basis for the column space of the object scores X ; that is, let $X^\# = X \times R$, where R is a rotation matrix satisfying $R'R = RR' = I_p$. We then get from that $Y_j^\# = D_j^{-1}G_j'X^\# = \hat{Y}_jR$.

An application of the technique to the sleeping bags data set is given next.

We apply homogeneity analysis to this data set and we provide two views of the solution: (a) a joint map of objects and categories with lines connecting them (the so-called graph plot) (Figure 1) and (b) the maps of the three variables, that is, the categories of the variable and the objects connected to them

(the so-called **star plots**, see Figures 2, 3, and 4). Notice that objects with similar profiles are mapped to identical points on the graph plot, a property stemming from the centroid principle. That is the reason that fewer than 18 object points appear on the graph plot.

Several things become immediately clear. There are good, expensive sleeping bags filled with down fibers, and cheap, bad-quality sleeping bags filled with synthetic fibers. There are also some intermediate sleeping bags in terms of quality and price filled either with down or synthetic fibers. Finally, there are some expensive ones of acceptable quality and some cheap ones of good quality. However, there are no bad expensive sleeping bags.

Remark Some algebra (see [22]) shows that the homogeneity analysis solution presented above can also be obtained by a singular value decomposition of

$$J^{-1/2}LGD^{-1/2} = U\Lambda V', \quad (5)$$

where L is a centering operator that leaves the super-indicator matrix G containing the data in deviations from its column means, and D is a diagonal matrix

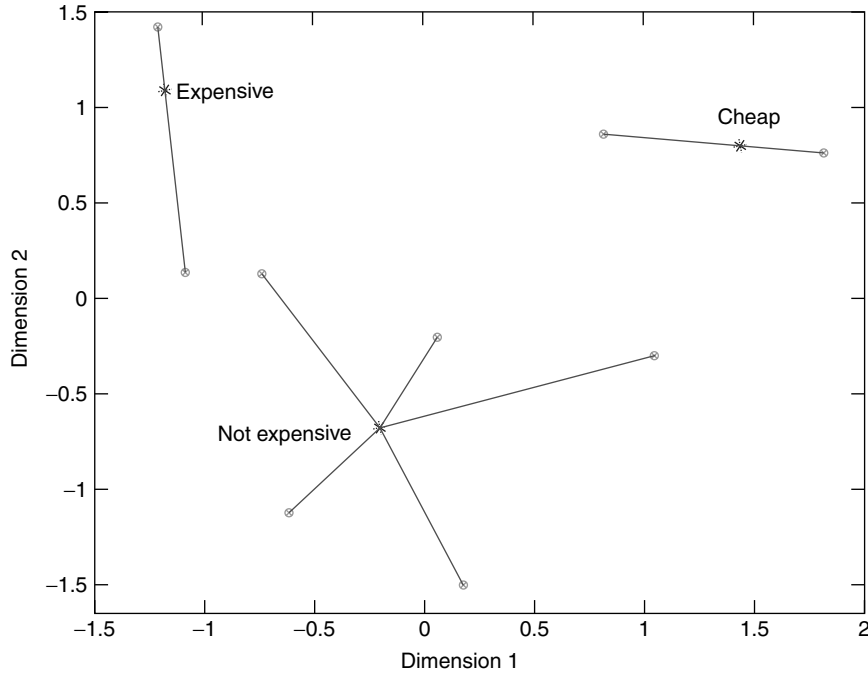


Figure 2 Star plot of variable price

4 Principal Components and Extensions

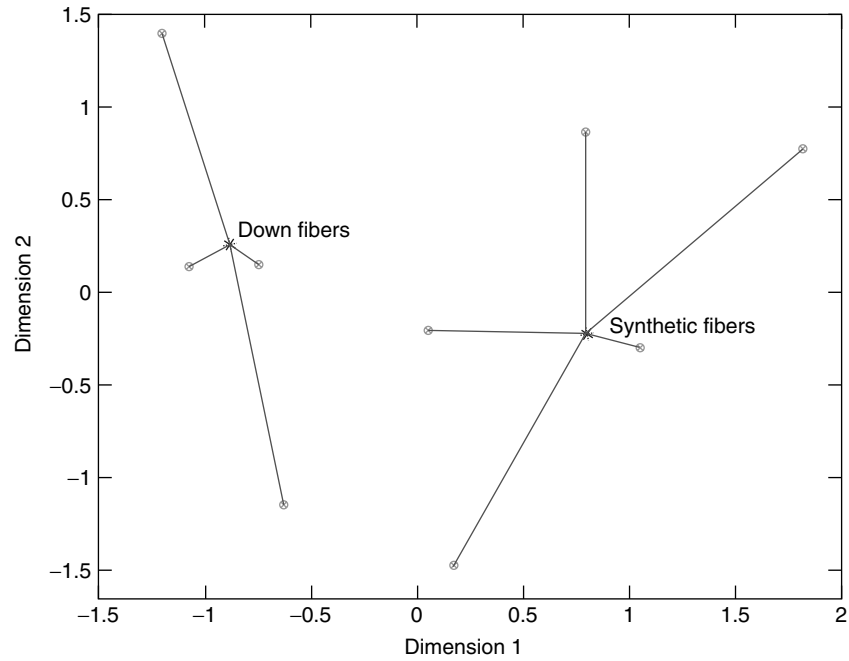


Figure 3 Star plot of variable fiber

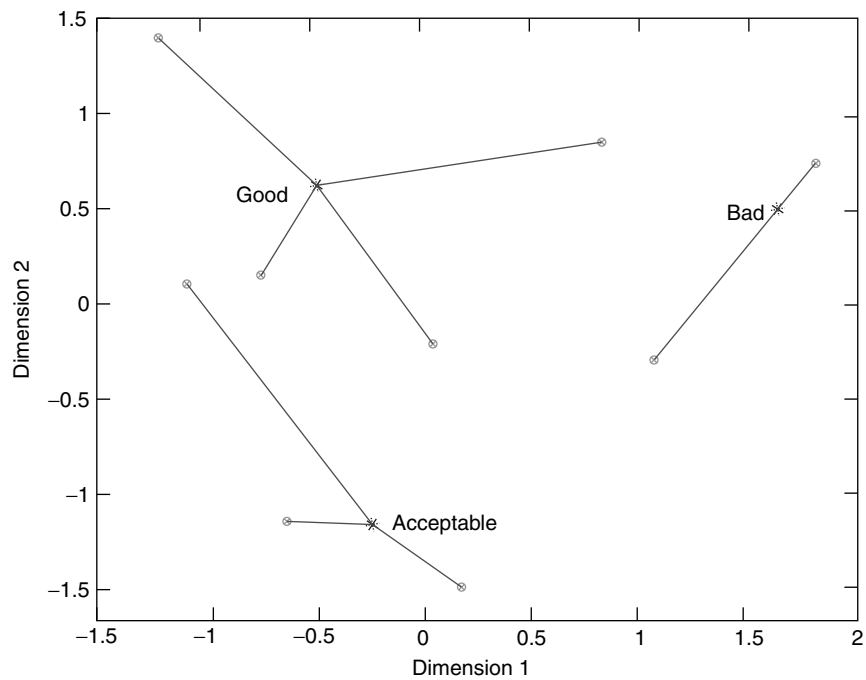


Figure 4 Star plot of variable quality

containing the univariate marginals of all the variables. The above expression shows why homogeneity analysis can be thought of as PCA for categorical data (see Section ‘PCA Problem Formulation’).

Incorporation of Ordinal Information

In many cases, the optimal scores for the categories of the variables exhibit nonmonotone patterns. However, this may not be a particularly desirable feature, especially when the categories encompass ordinal information. Furthermore, nonmonotone patterns may make the interpretation of the results a rather hard task. One way to overcome this difficulty is to further require

$$Y_j = q_j \beta_j', \quad (6)$$

where q_j is a k_j -column vector of *single* category scores for variable j , and β_j a d -column vector of weights (component loadings). Thus, each score matrix Y_j is restricted to be of rank-one, which implies that the scores in d dimensional space become proportional to each other.

As noted in [22], the introduction of the rank-one restrictions allows the existence of multidimensional solutions for object scores with a single quantification (optimal scaling) for the categories of the variables, and also makes it possible to incorporate the measurement level of the variables (ordinal, numerical) into the analysis (see **Scales of Measurement**).

The quality of the solution under such restrictions is still captured by the loss function given in (2), which is decomposed into two parts: the first component captures the loss (lack of fit) due to

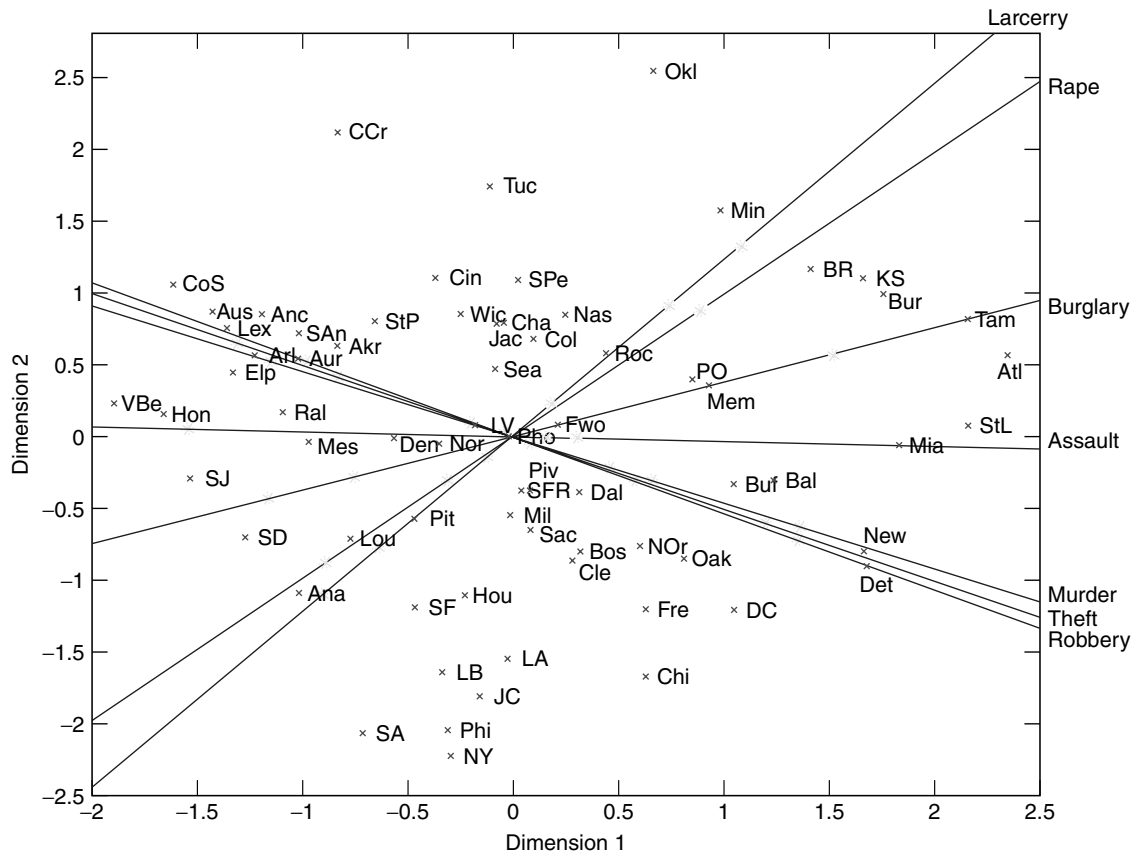


Figure 5 Variable quantifications (light grey points, not labeled) and US cities (The solid lines indicate the projections of the transformed variables. The placement of the variable labels indicates the high crime rate direction. (e.g., for the assault variable category 1 is on the left side of the graph, while category 6 on the right side))

6 Principal Components and Extensions

optimal scaling, while the second component captures the additional loss incurred due to the rank-one restrictions.

The optimal quantifications q_j are computed by (given that optimal variable scores $\hat{Y}_j$ have already been obtained)

$$\hat{q}_j = \frac{\hat{Y}_j \beta_j}{(\beta_j' \beta_j)}, \quad (7)$$

and the optimal loadings β_j by

$$\hat{\beta}_j = \frac{(\hat{Y}_j' D_j q_j)}{(q_j' D_j q_j)}. \quad (8)$$

An application of weighted monotone regression procedure [6], allows one to impose an ordinal constraint on the scores q_j . This procedure is known in the literature as the *Princals solution* (principal components analysis by means of alternating least squares), since an alternating least squares algorithm is used to calculate the optimal scores for the variables and their component loadings.

We demonstrate next the main features of the technique through an example. The data in this example give crime rates per 100 000 people in seven areas – murder, rape, robbery, assault, burglary, larceny, motor vehicle theft- for 1994 for each of the largest 72 cities in the United States. The data and their categorical coding is given in Appendix 2. In principle, we could have used homogeneity analysis to analyze and summarize the patterns in this data. However, we would like to incorporate into the analysis the underlying monotone structure in the data (higher crime rates are worse for a city) and thus we have treated all the variables as ordinal in a nonlinear principal components analysis. In Figure 5, the component loadings of the seven variables of a two-dimensional solution are shown. In case the loadings are of (almost) unit length, then the angle between any two of them reflects the value of the correlation coefficient between the two corresponding quantified variables. It can be seen that the first dimension (component) is a measure of overall crime rate, since all variables exhibit high loadings on it. On the other hand, the second component has high

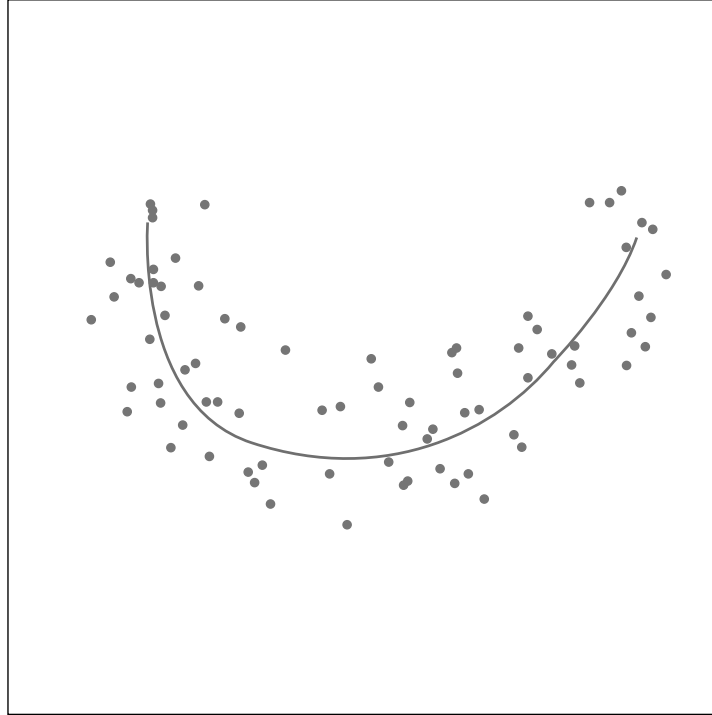


Figure 6 A schematic illustration of a principal curve in two dimensions

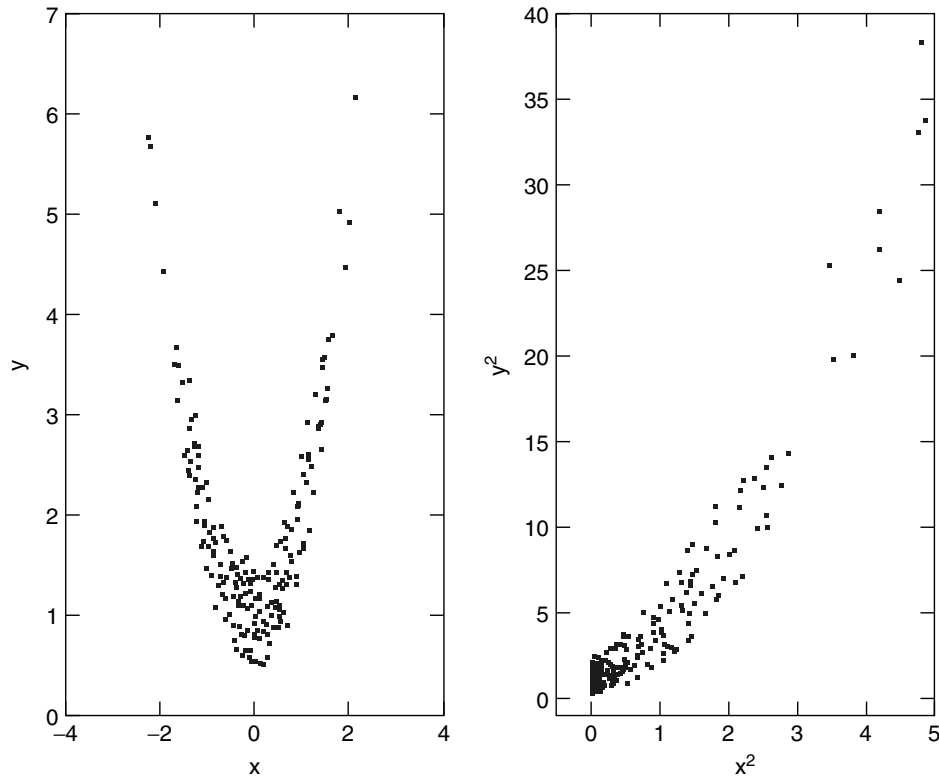


Figure 7 A two-dimensional nonlinear pattern (left panel) becomes more linear when additional dimensions are introduced. In the right panel the same data points are projected onto the x^2 , y^2 space

positive loadings on rape and larceny, and negative ones on murder, robbery, and auto theft. Thus, the second component will distinguish cities with large numbers of incidents involving larceny and rape from cities with high rates of auto thefts, murders, and robberies. Moreover, it can be seen that murder, robbery and auto theft are highly correlated, as are larceny and rape. The assault variable is also correlated, although to a lesser degree, with the first set of three variables and also with burglary. It is interesting to note that not all aspects of violent crimes are highly correlated (i.e., murder, rape, robbery, and assault) and the same holds for property crimes (burglary, larceny and auto thefts).

Software

The homogeneity analysis solution and its nonlinear principal components variety can be found in many popular software packages such as SPSS (under the

categories package), and SAS (under the corresponding procedure). A highly interactive version has been implemented in the xliisp-stat language [3], while a version with a wealth of plotting options for the results, as well as different ways of incorporating extraneous information through constraints can be found in R, as the homals package (*see Software for Statistical Analyses*).

Discussion

The main idea incorporated in homogeneity analysis and in nonlinear principal components analysis is that of optimal scoring (scaling) of the categorical variables. An extension of this idea to regression analysis with numerical variables can be found in the ACE methodology of Breiman and Friedman [5] and with categorical variables in the ALSOS system of Young et al. [31].

8 Principal Components and Extensions

The concept of **optimal scaling** has also been used in other multivariate analysis techniques (*see Multivariate Analysis: Overview*); for example, in canonical correlation analysis [20, 28], in flexible versions of discriminant analysis [16, 15] and in linear dynamical systems [2]. The book by Bekker and de Leeuw [29] deals with many theoretical aspects of optimal scaling (see also [8]).

Some Other Extensions of PCA

In this section, we briefly review some other extensions of PCA. All of the approaches covered try to deal with nonlinearities in the multivariate data.

Principal curves. If the data exhibit nonlinear patterns, then PCA that relies on linear associations would not be able to capture them. Hastie and Stuetzle [14] proposed the concept of a principal curve. The data are mapped to the closest point on the curve, or alternatively every point on the curve is the average of all the data points that are projected onto it (a different implementation of a centroid-like principle). An illustration of the principal curves idea is given in Figure 6. It can also be shown that the

regular principal components are the only straight lines possessing the above property. Some extensions of principal curves can be found in [9, 19].

Local PCA. If one would like to retain the conceptual simplicity of PCA, together with its algorithmic efficiency in the presence of nonlinearities in the data, he could resort to applying PCA locally. One possibility is to perform a **cluster analysis**, and then apply different principal components analyses to the various clusters [4]. Some recent advances on this topic are discussed in [21].

Kernel PCA. The idea behind kernel PCA [27] is that nonlinear patterns present in low-dimensional spaces can be linearized by projecting the data into high-dimensional spaces, at which point classical PCA becomes an effective. Although this concept is contrary to the idea of using PCA for data reduction purposes, it has proved successful in some application areas like handwritten digit recognition. An illustration of the linearization idea using synthetic data is given in Figure 7. In order to make this idea computationally tractable, kernels (to some extent they can be thought of as generalizations

Table 1 Sleeping bags data set

Sleeping bag	Expensive	Down	Good	Acceptable	Not expensive	Synthetic	Cheap	Bad
Foxfire	1	1	1	0	0	0	0	0
Mont Blanc	1	1	1	0	0	0	0	0
Cobra	1	1	1	0	0	0	0	0
Eiger	1	1	0	1	0	0	0	0
Viking	0	1	1	0	1	0	0	0
Climber Light	0	1	1	0	1	0	0	0
Traveler's Dream	0	1	1	0	1	0	0	0
Yeti Light	0	1	1	0	1	0	0	0
Climber	0	1	0	1	1	0	0	0
Cobra Comfort	0	1	0	1	1	0	0	0
Cat's Meow	0	0	1	0	1	1	0	0
Tyin	0	0	0	1	1	1	0	0
Donna	0	0	0	1	1	1	0	0
Touch the Cloud	0	0	0	1	1	1	0	0
Kompakt	0	0	0	1	1	1	0	0
One Kilo Bag	0	0	1	0	0	1	1	0
Kompakt Basic	0	0	1	0	0	1	1	0
Igloo Super	0	0	0	0	1	1	0	1
Sund	0	0	0	0	0	1	1	1
Finmark Tour	0	0	0	0	0	1	1	1
Interlight Lyx	0	0	0	0	0	1	1	1

of a covariance function) are used, that essentially calculate inner products of the original variables. It should be noted that kernels are at the heart of support vector machines, a very popular and successful classification technique [26].

Acknowledgments

The author would like to thank Jan de Leeuw for many useful suggestions and comments. This work was supported in part by NSF under grants IIS-9988095 and DMS-0214171 and NIH grant 1P41RR018627-01.

Appendix A

The sleeping bags data set was used in [25] and [22]. The data set is given in Table 1.

Appendix B

The data for this example are taken from table No. 313 of the 1996 Statistical Abstract of the United States. The coding of the variables is given next (see Table 2):

Murder 1: 0-10, 2: 11-20, 3: 21-40, 4: 40+

Rape 1: 0-40, 2: 41-60, 3: 61-80, 4: 81-100, 5: 100+

Robbery 1: 0-400, 2: 401-700, 3: 701-1000, 4: 1000+

Assault 1: 0-300, 2: 301-500, 3: 501-750, 4: 751-1000, 5: 1001-1250, 6: 1251+

Burglary 1: 0-1000, 2: 1001-1400, 3: 1401-1800, 4: 1801-2200, 5: 2200+

Larceny 1: 0-3000, 2: 3001-3500, 3: 3501-4000, 4: 4001-4500, 5: 4501-5000, 6: 5001-5500, 7: 5501-7000, 8: 7000+

Motor vehicle theft 1: 0-500, 2: 501-1000, 3: 1001-1500, 4: 1501-2000, 5: 2000+

Table 2 Crime rates data

City	City Code		City	City Code	
New York (NY)	NY	3134213	Los Angeles (CA)	LA	3235223
Chicago (IL)	Chi	3 ^a 46343	Houston (TX)	Hou	3223323
Philadelphia (PA)	Phi	3232114	San Diego (CA)	SD	1113223
Phoenix (AZ)	Pho	3213464	Dallas (TX)	Dal	3424354
Detroit (MI)	Det	4546445	San Antonio (TX)	SAn	2211362
Honolulu (HI)	Hon	1111252	San Jose (CA)	SJ	1213112
Las Vegas (NV)	LV	2323333	San Francisco (CA)	SF	2133253
Baltimore (MD)	Bal	4445474	Jacksonville (FL)	Jac	2424462
Columbus (OH)	Col	2522453	Milwaukee (WI)	Mil	3322244
Memphis (TN)	Mem	3533534	Washington, DC	DC	4246363
El Paso (TX)	ElP	1213152	Boston (MA)	Bos	2435245
Seattle (WA)	Sea	2223373	Charlotte (NC)	Cha	2325462
Nashville (TN)	Nas	2425373	Austin (TX)	Aus	1211262
Denver (CO)	Den	2312323	Cleveland (OH)	Cle	3533314
New Orleans (LA)	NOr	4433444	Fort Worth (TX)	FWo	3423363
Portland (OR)	Por	2426375	Oklahoma City (OK)	Okl	2514583
Long Beach (CA)	LB	2133324	Tucson (AZ)	Tuc	1314384
Kansas City (MO)	KS	3536574	Virginia Beach (VA)	VBe	1111131
Atlanta (GA)	Atl	4546585	Saint Louis (MO)	StL	4346585
Sacramento (CA)	Sac	2223455	Fresno (CA)	Fre	3234455
Tulsa (OK)	Tul	2314323	Miami (FL)	Mia	3246585
Oakland (CA)	Oak	3445454	Minneapolis (MN)	Min	2534573
Pittsburgh (PA)	Pit	2322223	Cincinnati (OH)	Cin	2523351
Toledo (OH)	Tol	2522453	Buffalo (NY)	Buf	3445533

Table 2 (continued)

City	City Code		City	City Code	
Wichita (KS)	Wic	2312463	Mesa (AZ)	Mes	1113353
Colorado Springs (CO)	Cos	1311151	Tampa (FL)	Tam	3546585
Santa Ana (CA)	SA	3122113	Arlington (VA)	Arl	1213242
Anaheim (CA)	Ana	1123223	Corpus Cristi (TX)	CCr	1313371
Louisville (KY)	Lou	2222312	St Paul (MN)	StP	2413332
Newark (NJ)	New	3346545	Birmingham (AL)	Bir	4536573
Norfolk (VA)	Nor	3322252	Anchorage (AK)	Anc	1313152
Aurora (CO)	Aur	1215252	St Petersburg (FL)	SPe	1426462
Riverside (CA)	Riv	2225434	Lexington (KY)	Lex	1213241
Rochester (NY)	Roc	3332562	Jersey City (NJ)	JC	2134414
Raleigh (NC)	Ral	2113341	Baton Rouge (LO)	BRo	3326584
Akron (OH)	Akr	2413232	Stockton (CA)	Sto	2224454

Note: The ^a for the Rape variable for Chicago denotes a missing observation.

References

- [1] Benzecri, J.P. (1973). *Analyse des Données*, Dunod, Paris.
- [2] Bijleveld, C.C.J.H. (1989). *Exploratory Linear Dynamic Systems Analysis*, DSWO Press, Leiden.
- [3] Bond, J. & Michailidis, G. (1996). Homogeneity analysis in Lisp-Stat, *Journal of Statistical Software* **1**(2), 1–30.
- [4] Bregler, C. & Omohundro, M. (1994). Surface learning with applications to lipreading, in *Advances in Neural Information Processing Systems*, S.J. Hanson, J.D. Cowan & C.L. Giles, eds, Morgan Kaufman, San Mateo.
- [5] Breiman, L. & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association* **80**, 580–619.
- [6] De Leeuw, J. (1977). Correctness of Kruskal's algorithms for monotone regression with ties, *Psychometrika* **42**, 141–144.
- [7] De Leeuw, J. (1983). On the prehistory of correspondence analysis, *Statistica Neerlandica* **37**, 161–164.
- [8] De Leeuw, J., Michailidis, G. & Wang, D. (1999). Correspondence analysis techniques, in *Multivariate Analysis, Design of Experiments and Survey Sampling*, S. Ghosh, ed., Marcel Dekker, New York, pp. 523–546.
- [9] Delicado, P. (2001). Another look at principal curves and surfaces, *Journal of Multivariate Analysis* **77**, 84–116.
- [10] Fisher, R.A. (1938). The precision of discriminant functions, *The Annals of Eugenics* **10**, 422–429.
- [11] Gifi, A. (1990). *Nonlinear Multivariate Analysis*, Wiley, Chichester.
- [12] Golub, G.H. & van Loan, C.F. (1989). *Matrix Computations*, Johns Hopkins University Press, Baltimore.
- [13] Guttman, L. (1941). The quantification of a class of attributes: a theory and a method of scale construction. *The Prediction of Personal Adjustment*. Horst et al., eds, Social Science Research Council, New York.
- [14] Hastie, T. & Stuetzle, W. (1989). Principal curves, *Journal of the American Statistical Association* **84**, 502–516.
- [15] Hastie, T., Buja, A. & Tibshirani, R. (1995). Penalized discriminant analysis, *Annals of Statistics* **23**, 73–102.
- [16] Hastie, T., Tibshirani, R. & Buja, A. (1994). Flexible discriminant analysis with optimal scoring, *Journal of the American Statistical Association* **89**, 1255–1270.
- [17] Hirschfeld, H.O. (1935). A connection between correlation and contingency, *Cambridge Philosophical Society Proceedings* **31**, 520–524.
- [18] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–441 and 498–520.
- [19] Kegl, B. & Krzyzak, A. (2002). Piecewise linear skeletonization using principal curves, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 59–74.
- [20] Koyak, R.A. (1987). On measuring internal dependence in a set of random variables, *The Annals of Statistics* **15**, 1215–1228.
- [21] Liu, Z.Y. & Xu, L. (2003). Topological local principal component analysis, *Neurocomputing* **55**, 739–745.
- [22] Michailidis, G. & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis, *Statistical Science* **13**, 307–336.
- [23] Michailidis, G. & de Leeuw, J. (2001). Data visualization through graph drawing, *Computational Statistics* **16**, 435–450.
- [24] Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Applications*, Toronto University Press, Toronto.
- [25] Prediger, S. (1997). Symbolic objects in formal concept analysis, in *Proceedings of the Second International Symposium on Knowledge, Retrieval, Use and Storage for Efficiency*, G. Mineau & A. Fall, eds.
- [26] Scholkopf, B. & Smola, A.J. (2002). *Learning with Kernels*, MIT Press, Cambridge.
- [27] Scholkopf, B., Smola, A. & Muller, K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* **10**, 1299–1319.

- [28] Van der Burg, E., de Leeuw, J. & Verdegaaal, R. (1988). Homogeneity analysis with K sets of variables: an alternating least squares method with optimal scaling features, *Psychometrika* **53**, 177–197.
- [29] Van Rijckevorsel, J. & de Leeuw, J. eds, (1988). *Component and Correspondence Analysis*, Wiley, Chichester.
- [30] Watanabe, S. (1965). Karhunen-Loeve expansion and factor analysis theoretical remarks and applications, *Transactions of the 4th Prague Conference on Information Theory*.
- [31] Young, F.W., de Leeuw, J. & Takane, Y. (1976). Regression with qualitative variables: an alternating least squares method with optimal scaling features, *Psychometrika* **41**, 505–529.

Further Reading

- De Leeuw, J. & Michailidis, G. (2000). Graph-layout techniques and multidimensional data analysis, *Papers in Honor of T.S. Ferguson*, Le Cam, L. & Bruss, F.T. (eds), 219–248 IMS Monograph Series, Hayward, CA.
- Michailidis, G. & de Leeuw, J. (2000). Multilevel homogeneity analysis with differential weighting *Computational Statistics and Data Analysis* **32**, 411–442.

GEORGE MICHAILIDIS

Probability: An Introduction

RANALD R. MACDONALD

Volume 3, pp. 1600–1605

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Probability: An Introduction

Probabilities are used by people to characterize the uncertainties they encounter in everyday life such as the winning of a sports event, the passing of an exam or even the vagaries of the weather, while in technical fields uncertainties are represented by probabilities in all the physical, biological, and social sciences. Probability theory has given rise to the disciplines of statistics, epidemiology, and actuarial studies; it even has a role in such disparate subjects as law, history, and ethics. Probability is useful for betting, insurance, risk assessment, planning, management, experimental design, interpreting evidence, predicting the future and determining what happened in the past (for more examples see [1, 3]).

A probability measures the uncertainty associated with the occurrence of an event on a scale from 0 to 1. When an event is impossible the probability is 0, when it is equally likely to occur or not to occur the probability is 0.5 and the probability is 1 when the event is certain. Probability is a key concept in all statistical inferences (*see Classical Statistical Inference: Practice versus Presentation*). What follows is intended as a short introduction to what probability is. Accounts of probability are included in most elementary statistics textbooks; Hacking [4] has written an introductory account of probability and Feller [2] is possibly the classic text on probability theory though it is not an introductory one.

Set theory can be used to characterize events and their probabilities where a set is taken to be a collection of events. Such representations are useful because new distinctions between events can be introduced which break down what was taken to be an event into more precisely defined events. For example the event of drawing an ace from a pack of cards can be broken down by suit, by player, by time, and by place.

Representing the Probability of a Single Event

In all probability applications there is one large set, consisting of the event in question and all other events that could happen should this event not occur. This

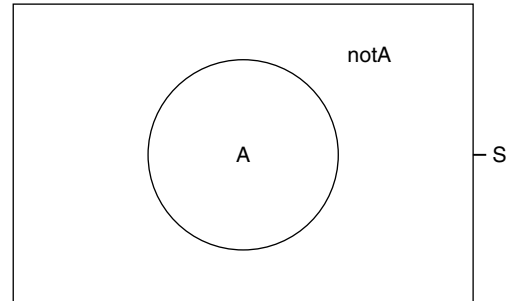


Figure 1 Venn diagram of event A and sample space S. A is the area in the circle and S the area in the rectangle

is called the *sample space* and is denoted by S. For instance, S might consist of all instances of drawing a card. In set theory, any event that might occur is represented by a set of some but not all of the possible events. Thus, it is a region of the sample space (a subset of S) and it can be represented as in Figure 1, where the event A might be drawing an ace and not A drawing any other card.

In Figure 1, called a *Venn diagram*, the sample space S is all the area enclosed by the rectangle, event A is the region within the circle and the event corresponding to the nonoccurrence of A (the complement of A) is represented by the region S-A (all the area within the rectangle except the area within the circle). The Venn diagram corresponds to a probability model where the areas within regions of the diagram correspond to the probabilities of the events they represent. The area within the circle is the probability of event A occurring and the area outside the circle within the diagram is the probability of A not occurring. A and not A are said to be *mutually exclusive* as A and not A cannot occur together and *exhaustive* as either A or not A must happen. The area within the rectangle is defined to be 1. Thus, $p(A) + p(\text{not}A) = p(S) = 1$.

Representing the Probability of Two Events

A Venn diagram gets more complicated when it is extended to include two events A and B which are not mutually exclusive (see Figure 2). For example, A might be drawing an ace and B drawing a club. (If A and B were mutually exclusive, the circles in Figure 2 would not overlap. This would be the case

2 Probability: An Introduction

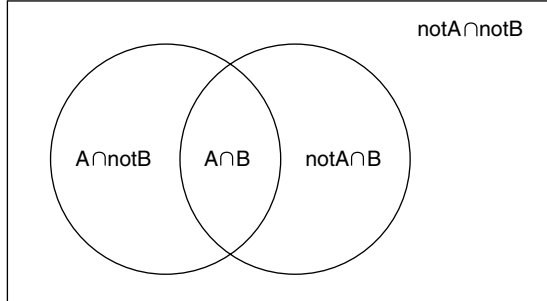


Figure 2 Venn diagram of events A and B. A is the area in the left circle, B the area in the right circle and $A \cap B$ the area the circles have in common

where A was drawing an ace and B was drawing a King from a pack of cards.) Figure 2 distinguishes between four different outcomes described below.

1. Both A and B occur. This can be written as $A \cap B$ and it is called the *intersection* of A & B or A ‘cap’ B.
2. A occurs and B does not, written as $A \cap \text{not}B$.
3. A does not occur and B does, written as $\text{not}A \cap B$.
4. Neither A or B occurs, written as $\text{not}A \cap \text{not}B$.

Again the Venn diagram corresponds to a probability model where the four regions correspond to the probabilities of the four possible outcomes.

It is also useful to introduce the concept of the *union* of two sets A and B written as $A \cup B$ which stands for the occurrence of either A or B or both. This is sometimes referred to as A union B or A ‘cup’ B.

$$A \cup B = (A \cap \text{not}B) + (\text{not}A \cap B) + (A \cap B). \quad (1)$$

From Figure 2 it can be seen that the probability of A union B

$$p(A \cup B) = p(A) + p(B) - p(A \cap B). \quad (2)$$

Example 1 Suppose there are a large number of boxes half of which contain a prize while the other half contain nothing. If one were presented with three boxes chosen at random (where each box has the same probability of being chosen) find

- (a) $p(A)$ the probability that at least one box contains a prize,

Table 1 Distinguishable outcomes when three boxes are randomly selected which either contain a prize (P) or are empty (E)

Box 1	P	P	P	P	E	E	E	E
Box 2	P	P	E	E	P	P	E	E
Box 3	P	E	P	E	P	E	P	E

- (b) $p(B)$ the probability that at least one box is empty,
- (c) $p(A \cap B)$ the probability that at least one contains a prize and that at least one is empty,
- (d) $p(A \cup B)$ the probability that at least one contains a prize or that at least one is empty.

The sample space of all distinguishable outcomes is given in Table 1.

Because the probability of a box containing a prize is .5 and the boxes have been chosen at random, all the patterns of boxes in Table 1 are equally likely and thus each pattern has a probability of 1/8.

- (a) The number of patterns of boxes with at least one prize is 7 and the sample space is 8 so the probability $p(A)$ that at least one box contains a prize is 7/8.
- (b) Similarly, the probability $p(B)$ that at least one box is empty is 7/8.
- (c) The number of patterns of boxes that have at least one box containing a prize and at least one empty is 6 and the sample space is 8 so the probability $p(A \cap B)$ of at least one box containing a prize and at least one empty box is $6/8 = 3/4$.
- (d) The probability $p(A \cup B)$ that at least one box contains a prize or that at least one is empty is 1 as all patterns have either a box containing a prize or an empty box. Alternatively it could be noted that

$$\begin{aligned} p(A \cup B) &= p(A) + p(B) - p(A \cap B) \\ &= \frac{7}{8} + \frac{7}{8} - \frac{6}{8} = 1. \end{aligned} \quad (3)$$

A key finding that allows us to treat the long run frequencies of events as probabilities is the law of large numbers (see **Laws of Large Numbers** and [3] for an account of its development). The law states that the average number of times a particular outcome occurs in a number of events (which are assumed to be equivalent with respect to the probability of the outcome in question) will asymptote to the probability of

the outcome as the number of events increases. Hacking [3] noted that subsequent generations ignored the need for equivalence and took the law to be an *a priori* truth about the stability of mass phenomena, but where equivalence can be assumed the law implies that good estimates of probabilities can be obtained from long run frequencies. This is how one can infer probabilities from incidence rates. Everitt [1] gives tables of the rate of death per 100,000 from various activities including, with the rates given in brackets, motorcycling (2000), hang gliding (80), and boating (5). From these it follows that the probability that someone taken at random from the sampled population dies from motorcycling, gliding, and boating are 0.02, 0.0008 and 0.00005 respectively.

Conditional Probabilities

To return to Figure 2, suppose it becomes known that one of the events, say B, has occurred. This causes the sample space to change to the events inside circle B. When the sample space can change in this way it is better to think of the areas in the Venn diagram as relative likelihoods rather than as probabilities. The probability of an event is now given by the ratio of the relative likelihood of the event in question over the relative likelihood of all the events that are considered possible. Thus, if it is known that B has occurred, the only possible regions of the diagram are $A \cap B$ and $\text{not}A \cap B$ and the probability of A becomes the likelihood of $A \cap B$ divided by the likelihood of $A \cap B$ and $\text{not}A \cap B$. This is known as the *conditional probability* of A given B and it is written as $p(A|B)$. Thus, the probability of A before anything is known is

$$p(A) = \frac{[p(A \cap B) + p(A \cap \text{not}B)]}{[p(A \cap B) + p(A \cap \text{not}B) + p(\text{not}A \cap B) + p(\text{not}A \cap \text{not}B)]}. \quad (4)$$

The probability of A after B has occurred is

$$\begin{aligned} p(A|B) &= \frac{[p(A \cap B)]}{[p(A \cap B) + p(\text{not}A \cap B)]} \\ &= \frac{[p(A \cap B)]}{[p(B)]}. \end{aligned} \quad (5)$$

One reason why conditional probabilities are useful is that all probabilities are conditional in that every probability depends on a number of assumptions.

Nevertheless, most of the time these assumptions are taken for granted and the conditioning is ignored. Conditional probabilities are useful when one is interested in the relationship between the occurrence of one event and the probability of another. For example, conditional probabilities are useful in signal detection theory, where an observer has to detect the presence of a signal on a number of trials. It was found that the probability of being correct on a trial (as measured by the relative frequency) was an unsatisfactory measure of detection ability as it was confounded by response bias. Signal detection theory uses both the conditional probabilities of a hit (detecting a signal given that a signal was present) and of a false alarm (the probability of detecting a signal given that no signal was present) to determine a measure of detection ability that is, at least to a first approximation, independent of response bias (see **Signal Detection Theory** and [6]).

Independence

Event A is said to be *independent* of event B when the probability of A remains the same regardless of whether or not B occurs. Thus,

$$p(A) = p(A|B) = p(A|\text{not}B)$$

but from the definition of conditional probability

$$p(A|B) = \frac{p(A \cap B)}{p(B)}. \quad (6)$$

Substituting $p(A)$ for $p(A|B)$ and rearranging the equation gives

$$p(A \cap B) = p(A) \times p(B). \quad (7)$$

Similar arguments show that

$$\begin{aligned} p(A \cap \text{not}B) &= p(A) \times p(\text{not}B) \\ p(\text{not}A \cap B) &= p(\text{not}A) \times p(B) \\ p(\text{not}A \cap \text{not}B) &= p(\text{not}A) \times p(\text{not}B). \end{aligned} \quad (8)$$

It also follows that whenever A is independent of B, B is independent of A because

$$\begin{aligned} p(B|A) &= \frac{p(A \cap B)}{p(A)} \\ &= \frac{p(A) \times p(B)}{p(A)} \text{ (from above)} = p(B). \end{aligned} \quad (9)$$

4 Probability: An Introduction

Table 2 Sample space when at least one box is empty

Box 1	P	P	P	E	E	E	E
Box 2	P	E	E	P	P	E	E
Box 3	E	P	E	P	E	P	E

Example 1 (continued) If three boxes are chosen at random, (e) what is the probability that at least one contains a prize given that at least one of the boxes is empty ($p(A|B)$) and (f) show that the probabilities of having at least one prize in the three boxes $P(A)$ and that at least one box is empty is $P(B)$ are not independent.

- (e) When at least one box is empty the sample space is reduced to that given in Table 2.
As all the possibilities are equally likely the probability $p(A|B)$ of at least one prize given at least one box is empty is $6/7$.
Alternatively $p(A|B) = [p(A \cap B)]/[p(B)] = [3/4]/[7/8] = 6/7$.
- (f) From (a) $p(A) = 7/8$ and from (e) $p(A|B) = 6/7$ so the probability of A (at least one prize) is related to B (the probability of at least one empty box). Alternatively it could be noted from (c) that $p(A \cap B) = 3/4$ and that this does not equal $p(A)*p(B)$ which substituting from (a) and (b) is $(7/8)^2$.

Independence greatly simplifies the problem of specifying the probability of an event as any variable that is independent of the event in question can be ignored. Furthermore, where a set of data can be regarded as composed of a number of independent events their combined probability is simply the product of the probabilities of each of the individual events. In most statistical applications, the joint probability of a set of data can be determined because the observations have been taken to be identically independently distributed and so the joint probability is the product of the probabilities of the individual observations. Little or nothing could be said about the effects of sampling error where the data was supposed to consist of events derived from separate unknown nonindependent probability distributions.

Applying Probability to a Particular Situation

In practice, applying probability to real life situations can be difficult because of the difficulty in identifying

the sample space and all the events in it that should be distinguished. The following apparently simple problem, called the *Monty Hall problem* after a particular quiz master, gave rise to considerable controversy [5] though the problem itself is very simple. In a quiz show, a contestant chooses one of three boxes knowing that there is a prize in one and only one box. The quiz master then reveals that a particular other box is empty and gives the contestant the option of choosing the remaining box instead of the originally chosen one. The problem is, should the contestant switch?

The nub of the problem involves noting that if contestants' first choices are correct they will win by sticking to their original choice; if they are wrong they will win by switching. To spell out the situation, take the sample space to be the nine cells specified in Table 3 where the boxes are labeled A, B & C. When a contestant's choice becomes known the sample space reduces to one particular column. Suppose a contestant's first choice is wrong; this implies that the chosen box is empty, so if another identified box is also empty the contestant will win by switching to the third box. On the other hand, if the contestant's first choice is correct the contestant will lose by switching. Since the probability of the first choice being correct is $1/3$, the probability of winning by sticking is $1/3$ whereas the probability of winning by switching is $2/3$ because the probability of the first choice being wrong is $2/3$. Table 3 gives the winning strategy for each cell. This answer was counter intuitive to a number of mathematicians [5].

Actually this is a case in which the world may be more complex than the probability model. The correct strategy should depend on what you think the quiz master is doing. Where the quiz master always makes the offer (which requires the quiz master to know where the prize is) the best strategy is to switch as described above. However, the quiz master might wish to keep the prize and knowing where the prize is, only offer the possibility of switching when the first choice is right. In this case, switching will always be

Table 3 Winning strategies on being showed an empty box in the Monty Hall problem

	A chosen	B chosen	C chosen
Prize in A	Stick	Switch	Switch
Prize in B	Switch	Stick	Switch
Prize in C	Switch	Switch	Stick

wrong. On the other hand, the quiz master might want to be helpful and only offer a switch when the first choice is wrong and under these circumstances the contestant will win every time by switching whenever possible. In other words, contestants should switch unless they think that the offer of a second choice is related to whether the first choice is correct in which case the best strategy depends on the size of the relationship between having chosen the correct box and being made the offer.

Example 2 In the Monty Hall problem, suppose the quiz master offers a choice of switching with a probability P if the contestant is wrong and kP if the contestant is correct (where k is a constant > 1). How large should k be in order for sticking to be the best policy?

Before any offer can be made the probability of the contestant's first choice being correct ($p(\text{FC})$) is $1/3$ and of being wrong ($p(\text{FW})$) $= 2/3$. However, when the quiz master adopts the above strategy being made an offer (O) gives information that $p(\text{FC})$ is greater than $1/3$. The position can be formalized in the following way:

From the definition of conditional probability:

$$p(\text{FC}|\text{O}) = \frac{p(\text{O} \cap \text{FC})}{[p(\text{O})]}. \quad (10)$$

Rewriting $p(\text{O} \cap \text{FC})$ as $p(\text{O}|\text{FC}) \times p(\text{FC})$, noting that $p(\text{O})$ is the sum of the probabilities of $\text{O} \cap \text{FC}$ and of $\text{O} \cap \text{FW}$ gives

$$p(\text{FC}|\text{O}) = \frac{p(\text{O}|\text{FC}) \times p(\text{FC})}{[p(\text{O} \cap \text{FC}) + p(\text{O} \cap \text{FW})]} \quad (11)$$

expanding the denominator according to the definition of conditional probability gives

$$\begin{aligned} p(\text{FC}|\text{O}) &= \frac{p(\text{O}|\text{FC}) \times p(\text{FC})}{[p(\text{O}|\text{FC}) \times p(\text{FC}) + p(\text{O}|\text{FW}) \times p(\text{FW})]} \\ &= \frac{p(\text{O}|\text{FC}) \times p(\text{FC})}{[kP \times (1/3) + P \times (2/3)]} \end{aligned} \quad (12)$$

and substituting yields

$$p(\text{FC}|\text{O}) = \frac{kP \times (1/3)}{[kP \times (1/3) + P \times (2/3)]}$$

$$= \frac{k}{(k+2)}. \quad (13)$$

Thus, when $k = 2$ and an offer is made the probability that the first choice is correct is a half and it makes no difference whether the contestant sticks or switches. When k is less than 2 the probability that the first choice is correct, given that an offer has been made, is less than a half and so the probability that the first choice is wrong is greater than a half. On being made an offer the contestant should switch because when the first choice is wrong the contestant will win by switching. However, if k is greater than 2 and the contestant is made an offer the probability that the first choice is correct is greater than a half and the contestant should stick as the probability of winning by switching is less than a half. This form of reasoning is called *Bayesian* (see **Bayesian Statistics**). It might be pointed out that this analysis could be improved on where, by examining the quiz master's reactions, it was possible to distinguish between offers of switches that were more likely to help as opposed to those offers that were more likely to hinder the contestant. (For a discussion of the subtleties of relating probabilities to naturally occurring events, see **Probability: Foundations of; Incompleteness of Probability Models**).

References

- [1] Everitt, B.S. (1999). *Chance Rules*, Copernicus, New York.
- [2] Feller, W. (1971). *An Introduction to Probability Theory*, Vol. 1, 3rd Edition, Wiley, New York.
- [3] Hacking, I. (1990). *The Taming of Chance*, Cambridge University Press, Cambridge.
- [4] Hacking, I. (2001). *Introduction to Probability and Inductive Logic*, Cambridge University Press, Cambridge.
- [5] Hoffman, P. (1998). *The Man who Loved Only Numbers*, Fourth Estate, London.
- [6] Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, Wiley, New York.

(See also **Bayesian Belief Networks**)

RANALD R. MACDONALD

Probability Plots

DAVID C. HOWELL

Volume 3, pp. 1605–1608

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Probability Plots

Probability plots are often used to compare two empirical distributions, but they are particularly useful to evaluate the shape of an empirical distribution against a theoretical distribution. For example, we can use a probability plot to examine the degree to which an empirical distribution differs from a normal distribution.

There are two major forms of probability plots, known as probability–probability (P-P) plots and quantile–quantile (Q-Q) plots (*see Empirical Quantile–Quantile Plots*). They both serve essentially the same purpose, but plot different functions of the data.

An Example

It is easiest to see the difference between P-P and Q-Q plots with a simple example. The data displayed in Table 1 are derived from data collected by Compas (personal communication) on the effects of stress on cancer patients. Each of the 41 patients completed the Externalizing Behavior scale of the Brief Symptom Inventory. We wish to judge the normality of the sample data. A histogram for these data is presented in Figure 1 with a normal distribution superimposed.

The sample mean is 12.10, and the sample standard deviation is 10.32.

P-P Plots

A P-P plot displays the cumulative probability of the obtained data on the x -axis and the corresponding expected cumulative probability for the reference distribution (in this case, the normal distribution) on the y -axis. For example, a raw score of 15 for our data had a cumulative probability of 0.756. It also had a z score, given the mean and standard deviation of the sample, of 0.28. For a normal distribution, we expect to find 61% of the distribution falling at or below $z = 0.28$. So, for this observation, we had considerably more scores (75.6%) falling at or below 15 than we would have expected if the distribution were normal (61%). From Table 1, you can see the results of similar calculations for the full data set. These are displayed in columns 4 and 6.

If we plot the obtained cumulative percentage on the x -axis and the expected cumulative percentage on the y -axis, we obtain the results shown in Figure 2. Here, a line drawn at 45° is superimposed. If the data had been exactly normally distributed, the points would have fallen on that line. It is clear from the figure that this was not the case. The points deviate noticeably from the line, and, thus, from normality.

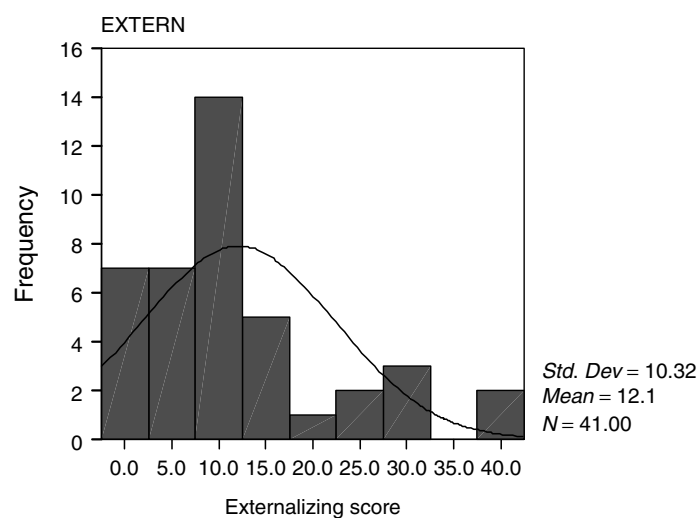


Figure 1 Histogram of externalizing scores with normal curve superimposed

2 Probability Plots

Table 1 Frequency distribution of externalizing scores with normal probabilities and quantiles

Score	Frequency	Percent	Cumulative percent	z	CDF normal	Normal quantile
0	3	7.3	7.3	-1.17	0.12	-3.01
1	2	4.9	12.2	-1.07	0.14	-0.03
2	2	4.9	17.1	-0.98	0.16	2.19
4	1	2.4	19.5	-0.78	0.22	3.13
5	3	7.3	26.8	-0.69	0.25	5.61
6	2	4.9	31.7	-0.59	0.28	7.08
7	1	2.4	34.1	-0.49	0.31	7.77
8	1	2.4	36.6	-0.40	0.35	8.46
9	6	14.6	51.2	-0.30	0.38	12.31
10	2	4.9	56.1	-0.20	0.42	13.58
11	3	7.3	63.4	-0.1	0.46	15.54
12	2	4.9	68.3	-0.01	0.50	16.92
14	2	4.9	73.2	0.18	0.57	18.39
15	1	2.4	75.6	0.28	0.61	19.16
16	2	4.9	80.5	0.38	0.65	20.87
19	1	2.4	82.9	0.67	0.75	21.81
24	1	2.4	85.4	1.15	0.88	22.88
25	1	2.4	87.8	1.25	0.89	24.03
28	1	2.4	90.2	1.54	0.94	25.35
29	1	2.4	92.7	1.64	0.95	27.01
31	1	2.4	95.1	1.83	0.97	29.08
40	1	2.4	97.6	2.70	0.99	32.41
42	1	2.4	100.0	2.90	1.00	36.02

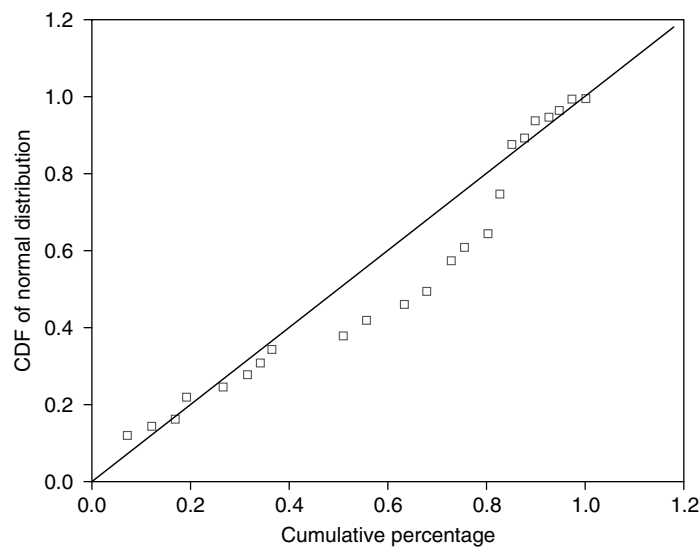


Figure 2 P-P plot of obtained distribution against the normal distribution

Q-Q Plots

A Q-Q plot resembles a P-P plot in many ways, but it displays the observed values of the data on the x -axis and the corresponding expected values from a normal

distribution on the y -axis. The expected values are based on the **quantiles** of the distribution. To take our example of a score of 15 again, we know that 75.6% of the observations fell at or below 15, placing 15 at the 75.6 percentile. If we had a normal distribution

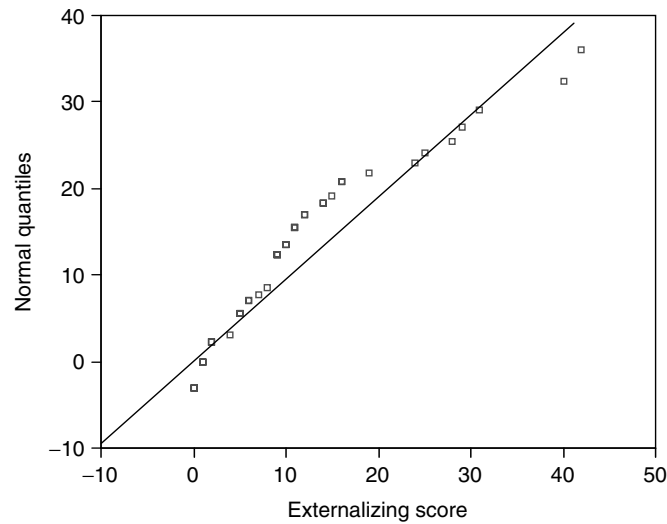


Figure 3 Q-Q plot of obtained distribution against the normal distribution

with a mean of 12.10 and a standard deviation of 10.32, we would expect the 75.6 percentile to fall at a score of 19.16. We can make similar calculations for the remaining data, and these are shown in column 7 of Table 1. (I treated the final cumulative percentage as 99.9 instead of 100 because the normal distribution runs to infinity and the quantile for 100% would be undefined.)

The plot of these values can be seen in Figure 3, where, again, a straight line is superimposed representing the results that we would have if the data were perfectly normally distributed. Again, we can see significant departure from normality.

DAVID C. HOWELL

Probits

CATALINA STEFANESCU, VANCE W. BERGER AND SCOTT L. HERSHBERGER

Volume 3, pp. 1608–1610

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Probits

Probit models have arisen in the context of analysis of dichotomous data. Let $Y_1, \dots, Y_n$ be n binary variables and let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ denote corresponding vectors of covariates. The flexible class of Probit models may be obtained by assuming that the response $Y_i (1 \leq i \leq n)$ is an indicator of the event that some unobserved continuous variable, Z_i say, exceeds a threshold, which can be taken to be zero, without loss of generality. Specifically, let $Z_1, \dots, Z_n$ be latent continuous variables and assume that

$$\begin{aligned} Y_i &= I_{\{Z_i > 0\}}, \quad \text{for } i = 1, \dots, n, \\ Z_i &= \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \end{aligned} \quad (1)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression parameters. In this formulation, $\mathbf{x}_i \boldsymbol{\beta}$ is sometimes called the *index function* [11]. The marginal probability of a positive response with covariate vector $\mathbf{x}$ is given by

$$\begin{aligned} p(\mathbf{x}) &= \Pr(Y = 1; \mathbf{x}) = \Pr(\mathbf{x} \boldsymbol{\beta} + \varepsilon > 0) \\ &= 1 - \Phi(-\mathbf{x} \boldsymbol{\beta}), \end{aligned} \quad (2)$$

where $\Phi(x)$ is the standard normal cumulative distribution function. Also,

$$\begin{aligned} \text{Var}(Y; \mathbf{x}) &= p(\mathbf{x})\{1 - p(\mathbf{x})\} \\ &= \{1 - \Phi(-\mathbf{x} \boldsymbol{\beta})\} \Phi(-\mathbf{x} \boldsymbol{\beta}). \end{aligned} \quad (3)$$

As a way of relating stimulus and response, the Probit model is a natural choice in situations in which an interpretation for a threshold approach is readily available. Examples include attitude measurement, assigning pass/fail gradings for an examination based on a mark cut-off, and categorization of illness severity based on an underlying continuous scale [10]. The Probit models first arose in connection with bioassay [4] – in toxicology experiments, for example, sets of test animals are subjected to different levels x of a toxin. The proportion $p(x)$ of animals surviving at dose x can then be modeled as a function of x , following (2). The surviving proportion is increasing in the dose when $\beta > 0$ and it is decreasing in the dose when $\beta < 0$. Surveys of the toxicology literature on Probit modeling are included in [7] and [9].

Probit models belong to the wider class of **generalized linear models** [13]. This class also includes the logit models, arising when the random errors ε_i in (1) have a logistic distribution. Since the logistic distribution is similar to the normal except in the tails, whenever the binary response probability p belongs to $(0.1, 0.9)$, it is difficult to discriminate between the logit and Probit functions solely on the grounds of goodness of fit. As Greene [11] remarks, ‘it is difficult to justify the choice of one distribution or another on theoretical grounds... in most applications, it seems not to make much difference’.

Estimation of the Probit model is usually based on maximum likelihood methods. The nonlinear likelihood equations require an iterative solution; the Hessian is always negative definite, so the log-likelihood is globally concave. The asymptotic covariance matrix of the **maximum likelihood estimator** can be estimated by using an estimate of the expected Hessian [2], or with the estimator developed by Berndt et al. [3]. Windmeijer [18] provides a survey of the many goodness-of-fit measures developed for binary choice models, and in particular, for Probits.

The maximum likelihood estimator in a Probit model is sometimes called a *quasi-maximum likelihood estimator (QMLE)* since the normal probability model may be misspecified. The QMLE is not consistent when the model exhibits any form of heteroscedasticity, nonlinear covariate effects, unmeasured heterogeneity, or omitted variables [11]. In this setting, White [17] proposed a robust ‘sandwich’ estimator for the asymptotic covariance matrix of the QMLE.

As an alternative to maximum likelihood estimation, Albert and Chib [1] developed a framework for estimation of latent threshold models for binary data, using data augmentation. The univariate Probit is a special case of this class of models, and data augmentation can be implemented by means of Gibbs sampling. Under this framework, the class of Probit regression models can be extended by using mixtures of normal distributions to model the latent data.

There is a large literature on the generalizations of the Probit model to the analysis of a variety of qualitative and limited dependent variables. For example, McKelvey and Zavoina [14] extend the Probit model to the analysis of ordinal dependent variables, while Tobin [16] discusses a class of models in which the dependent variable is limited in range. In particular,

the Probit model specified in (1) can be generalized by allowing the error terms ε_i to be correlated. This leads to a multivariate Probit model, useful for the analysis of clustered binary data (*see Clustered Data*). The multivariate Probit focuses on the conditional expectation given the cluster-level random effect, and thus it belongs to the class of cluster-specific approaches for modeling correlated data, as opposed to population-average approaches of which the most common example are the **generalized estimating equations** (GEE)-type methods [19].

The multivariate Probit model has several attractive features that make it particularly suitable for the analysis of correlated binary data. First, the connection to the Gaussian distribution allows for flexible modeling of the association structure and straightforward interpretation of the parameters. For example, the model is particularly attractive in marketing research of consumer choice because the latent correlations capture the cross-dependencies in latent utilities across different items. Also, within the class of cluster-specific approaches, the exchangeable multivariate Probit model is more flexible than other fully specified models (such as the beta-binomial), which use compound distributions to account for overdispersion in the data. This is due to the fact that both underdispersion and overdispersion can be accommodated in the multivariate Probit model through the flexible underlying covariance structure. Finally, due to the underlying threshold approach, the multivariate Probit model has the potential of extensions to the analysis of clustered mixed binary and continuous data or of multivariate binary data [12, 15].

Likelihood methods are one option for inference in the multivariate Probit model (see e.g., [5]), but they are computationally difficult due to the intractability of the expressions obtained by integrating out the latent variables. As an alternative, estimation can be done in a Bayesian framework [6, 8] where generic prior distributions may be employed to incorporate prior information. Implementation is usually done with **Markov chain Monte Carlo methods** – in particular, the Gibbs sampler is useful in models where some structure is imposed on the covariance matrix (e.g., exchangeability).

References

- [1] Albert, J.H. & Chib, S. (1997). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**, 669–679.
- [2] Amemiya, T. (1981). Qualitative response models: a survey, *Journal of Economic Literature* **19**, 481–536.
- [3] Berndt, E., Hall, B., Hall, R. & Hausman, J. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**, 653–665.
- [4] Bliss, C.I. (1935). The calculation of the dosage-mortality curve, *Annals of Applied Biology* **22**, 134–167.
- [5] Chan, J.S.K. & Kuk, A.Y.C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects, *Biometrics* **53**, 86–97.
- [6] Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models, *Biometrika* **85**, 347–361.
- [7] Cox, D. (1970). *Analysis of Binary Data*, Methuen, London.
- [8] Edwards, Y.D. & Allenby, G.M. (2003). Multivariate analysis of multiple response data, *Journal of Marketing Research* **40**, 321–334.
- [9] Finney, D. (1971). *Probit Analysis*, Cambridge University Press, Cambridge.
- [10] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold Publishing, London.
- [11] Greene, W.H. (2000). *Econometric Analysis*, 4th Edition, Prentice Hall, Englewood Cliffs.
- [12] Gueorguieva, R.V. & Agresti, A. (2001). A correlated probit model for joint modelling of clustered binary and continuous responses, *Journal of the American Statistical Association* **96**, 1102–1112.
- [13] McCullagh, P. & Nelder, J.A. (1989). *Generalised Linear Models*, 2nd Edition, Chapman & Hall, London.
- [14] McKelvey, R.D. & Zavoina, W. (1976). A statistical model for the analysis of ordinal level dependent variables, *Journal of Mathematical Sociology* **4**, 103–120.
- [15] Regan, M.M. & Catalano, P.J. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology, *Biometrics* **55**, 760–768.
- [16] Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica* **26**, 24–36.
- [17] White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **53**, 1–1.
- [18] Windmeijer, F. (1995). Goodness of fit measures in binary choice models, *Econometric Reviews* **14**, 101–116.
- [19] Zeger, S.L., Liang, K.Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equations approach, *Biometrics* **44**, 1049–1060.

CATALINA STEFANESCU, VANCE W. BERGER
AND SCOTT L. HERSHBERGER

Procrustes Analysis

INGWER BORG

Volume 3, pp. 1610–1614

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Procrustes Analysis

The Ordinary Procrustes Problem

The ordinary Procrustes¹ problem is concerned with fitting a configuration of points $\mathbf{X}$ to a fixed target configuration $\mathbf{Y}$ as closely as possible. The purpose of doing this is to eliminate nonessential differences between $\mathbf{X}$ and $\mathbf{Y}$. Consider an example. Assume that you have two **multidimensional scaling** (MDS) solutions with coordinate matrices

$$\mathbf{X} = \begin{pmatrix} 0.07 & 2.62 \\ 0.93 & 3.12 \\ 1.93 & 1.38 \\ 1.07 & 0.88 \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} 1.00 & 2.00 \\ -1.00 & 2.00 \\ -1.00 & -2.00 \\ 1.00 & -2.00 \end{pmatrix}.$$

These solutions appear to be quite different, but are the differences of the point coordinates due to the data that these configurations represent or are they a consequence of choosing particular coordinate systems for the MDS spaces? In MDS, it is essentially the ratios of the distances among the points that represent the meaningful properties of the data. Hence, all transformations that leave these ratios unchanged are admissible. We can therefore rotate, reflect, dilate, and translate MDS configurations in any way we like. So, if two or more MDS solutions are to be compared, it makes sense to first eliminate meaningless differences by such transformations. Figure 1 demonstrates that $\mathbf{X}$ can indeed be fitted to perfectly match $\mathbf{Y}$. In this simple two-dimensional case, we could possibly see the equivalence of $\mathbf{X}$ and $\mathbf{Y}$ by carefully studying the plots, but, in general, such ‘cosmetic’ fittings are needed to avoid seeing differences that are unfounded.

The Orthogonal Procrustes Problem

Procrustean procedures were first introduced in **factor analysis** because it frequently deals with relatively high-dimensional vector configurations that are hard to compare. Factor analysts almost always are interested in dimensions and, thus, comparing coordinate matrices is a natural research question to them. In addition, Procrustean methods can also be used in a confirmatory manner: one simply checks how well an empirical matrix of loadings $\mathbf{X}$ can be fitted to a hypothesized factor matrix $\mathbf{Y}$.

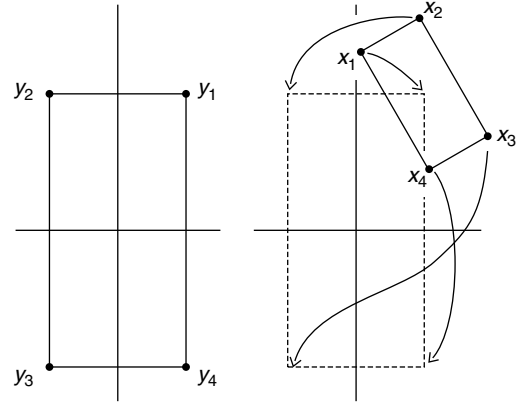


Figure 1 Illustration of fitting $\mathbf{X}$ to $\mathbf{Y}$ by a similarity transformation

In factor analysis, the group of admissible transformations is smaller than in MDS: Only rotations and reflections are admissible, because here the points must be interpreted as end-points of vectors emanating from a fixed origin. The lengths of these vectors and the angles between any two vectors represent the data. Procrustean fittings of $\mathbf{X}$ to $\mathbf{Y}$ are, therefore, restricted to rotations and reflections of the point configuration $\mathbf{X}$ or, expressed algebraically, to orthogonal transformations of the coordinate matrix $\mathbf{X}$.

Assume that $\mathbf{Y}$ and $\mathbf{X}$ are both of order $n \times m$. We want to fit $\mathbf{X}$ to $\mathbf{Y}$ by picking a best-possible matrix $\mathbf{T}$ out of the set of all orthogonal $\mathbf{T}$. By ‘best-possible’ one typically means a $\mathbf{T}$ that minimizes the sum of the squared distances between corresponding points of the fitted $\mathbf{X}$, $\hat{\mathbf{X}} = \mathbf{XT}$, and the target $\mathbf{Y}$. This leads to the loss function $L = \text{tr}[(\mathbf{Y} - \mathbf{XT})'(\mathbf{Y} - \mathbf{XT})]$ that must be minimized by picking a $\mathbf{T}$ that satisfies $\mathbf{T}'\mathbf{T} = \mathbf{TT}' = \mathbf{I}$. It can be shown that L is minimal if $\mathbf{T} = \mathbf{QP}'$, where $\mathbf{PAQ}'$ is the singular value decomposition (SVD) of $\mathbf{Y}'\mathbf{X}$ [8, 9, 12, 14].

Procrustean Similarity Transformations

We now extend the rotation/reflection task by also admitting all transformations that preserve the shape of $\mathbf{X}$. Figure 1 illustrates a case for such a similarity transformation. Algebraically, we here have $\hat{\mathbf{X}} = s\mathbf{XT} + \mathbf{1t}'$, where s is a dilation factor, $\mathbf{t}$ a translation vector, $\mathbf{T}$ a rotation/reflection matrix, and $\mathbf{1}$ a column vector of ones. The steps to find the optimal similarity transformation are [13]:

- (1) Compute $\mathbf{C} = \mathbf{Y}'\mathbf{J}\mathbf{X}$
- (2) Compute the SVD of $\mathbf{C} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}'$.
- (3) Compute the optimal rotation matrix as $\mathbf{T} = \mathbf{Q}\mathbf{P}'$.
- (4) Compute the optimal dilation factor as $s = (\text{tr } \mathbf{Y}'\mathbf{J}\mathbf{X}\mathbf{T})/(\text{tr } \mathbf{X}'\mathbf{J}\mathbf{X})$.
- (5) Compute the optimal translation vector as $\mathbf{t} = n^{-1}(\mathbf{Y} - s\mathbf{X}\mathbf{T})'\mathbf{1}$.

The matrix $\mathbf{J}$ is the ‘centering’ matrix, that is, $\mathbf{J} = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}'$ and $\mathbf{1}$ is a vector of n ones.

To measure the extent of similarity between $\hat{\mathbf{X}}$ and $\mathbf{Y}$, the **product-moment correlation coefficient** r computed over the corresponding coordinates of the matrices is a proper index. Langeheine [10] provides norms that allow one to assess whether this index is better than can be expected when fitting random configurations.

Special Issues

In practice, $\mathbf{Y}$ and $\mathbf{X}$ may not always have the same dimensionality. For example, if $\mathbf{Y}$ is a hypothesized loading pattern for complex IQ test items, one may only be able to theoretically predict a few dimensions, while $\mathbf{X}$ represents the empirical items in a higher-dimensional representation space. Technically, this poses no problem: all of the above matrix computations work if one appends columns of zeros to $\mathbf{Y}$ or $\mathbf{X}$ until both matrices have the same column order.

A further generalization allows for an $\mathbf{X}$ with missing cells [6, 15, 7]. A typical application is one where some points represent previously investigated variables and the remaining variables are ‘new’ ones. We might then use the configuration from a previous study as a partial target for the present data to check structural replicability.

A frequent practical issue is the case where the points of $\mathbf{X}$ and $\mathbf{Y}$ are not matched one-by-one but rather class-by-class. In this case, one may proceed by a more substantively motivated strategy, first averaging the coordinates of all points that belong to the same substantive class and then fit the matrices of centroids to each other. The transformations derived in this fitting can then be used to fit $\mathbf{X}$ to $\mathbf{Y}$ [1].

Robust Procrustean Analysis

In some cases, two configurations can be matched closely except for a few points. Using the

usual sum-of-squares loss functions (*see Least Squares Estimation*), these points can cause severe misfits. To deal with this problem, Verboon and Heiser [17] proposed more robust forms of Procrustes analysis. They first decompose the misfit into the contribution of each distance between corresponding points of $\mathbf{Y}$ and $\mathbf{X}$ to the total loss, $L = \text{tr}[(\mathbf{Y} - \mathbf{X}\mathbf{T})'(\mathbf{Y} - \mathbf{X}\mathbf{T})] = \sum_{i=1}^n (\mathbf{y}_i - \mathbf{T}'\mathbf{x}_i)'(\mathbf{y}_i - \mathbf{T}'\mathbf{x}_i) = \sum_{i=1}^n d_i^2$, where $\mathbf{y}_i$ and $\mathbf{x}_i$ denote rows i of $\mathbf{Y}$ and $\mathbf{X}$, respectively. The influence of outliers on $\mathbf{T}$ is reduced by minimizing a variant of L , $L_f = \sum_{i=1}^n f(d_i)$, where f is a function that satisfies $f(d_i) < d_i^2$. One obvious example for f is simply $|d_i|$, which minimizes the disagreement of two configurations in terms of point-by-point distances rather than in terms of squared distances. Algorithms for minimizing L_f with various robust functions can be found in [16].

Oblique Procrustean Analysis

Within the factor-analytic context, a particular form of Procrustean fitting has been investigated where $\mathbf{T}$ is only restricted to be a direction cosine matrix. This amounts to fitting $\mathbf{X}$ ’s column vectors to the corresponding column vectors of $\mathbf{Y}$ by individually rotating them about the origin. The main purpose for doing this is to fit a factor matrix to maximum similarity with a hypothesized factor pattern. The solutions proposed by Browne [3], Browne [5, 6], or [4] are formally interesting but not really needed anymore, because such testing can today be done more directly by **structural equation modeling** (SEM).

More than Two Configurations

Assume we have K different $\mathbf{X}_k$ and that we want to eliminate uninformative differences from them by similarity transformations. Expressed in terms of a loss function, this generalized Procrustes analysis amounts to $L = \sum_{k < l}^K \text{tr}(\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l)'(\tilde{\mathbf{X}}_k - \tilde{\mathbf{X}}_l) = \min$, where $\tilde{\mathbf{X}}_k = s_k \mathbf{X}_k \mathbf{T}_k + \mathbf{1}\mathbf{t}_k'$ and $\mathbf{T}_k' \mathbf{T}_k = \mathbf{I}$. To avoid a degenerate solution where $s_k = 0$, a restriction such as $\sum_k^K s_k^2 \text{tr } \tilde{\mathbf{X}}_k' \tilde{\mathbf{X}}_k = \sum_k^K \text{tr } \mathbf{X}_k' \mathbf{X}_k$ can be imposed [7].

This fitting can be done by several methods. For example, one can use the centroid configuration $\mathbf{Z}$ of

all $\tilde{\mathbf{X}}_k$'s, $\mathbf{Z} = (1/K) \sum_k \tilde{\mathbf{X}}_k$, and write the loss function as $L = \sum_{k=1}^K \text{tr}(\tilde{\mathbf{X}}_k - \mathbf{Z})(\tilde{\mathbf{X}}_k - \mathbf{Z})'$. This function is minimized by iteratively fitting each of the $\tilde{\mathbf{X}}_k$'s and successively updating the centroid configuration $\mathbf{Z}$. Geometrically, each of $\mathbf{Z}$'s points is the centroid of the corresponding points from the fitted individual configurations. Thus, L is small if these centroids lie somewhere in the middle of a tight cluster of K points, where each single point belongs to a different $\tilde{\mathbf{X}}_k$.

Procrustean Individual Differences Scaling

A rather obvious idea in Procrustean analysis of MDS spaces is to attempt to explain each $\mathbf{X}_k$ by a simple transform of $\mathbf{Z}$ (or of a fixed $\mathbf{Y}$). The most common model is the 'dimensional salience' or INDSCAL model (see **Three-mode Component and Scaling Methods**). If you think of $\mathbf{Z}$, in particular, as the space that characterizes a whole group of K individuals ('group space'), then each individual $\mathbf{X}_k$ may result from $\mathbf{Z}$ by differentially stretching or shrinking it along its dimensions. Figure 2 illustrates this notion. Here, the group space perfectly explains the two individual configurations if it is stretched by the factor 2 along dimension 2 and dimension 1, respectively.

The weighted centroid configuration for individual k can be expressed as $\mathbf{Z}\mathbf{W}_k$, where $\mathbf{W}_k$ is an $m \times m$ diagonal matrix of nonzero dimension weights. The Procrustean fitting problem requires one to generate both a dimensionally weighted target $\mathbf{Z}$, $\mathbf{Z}\mathbf{W}_k$, and the individual configurations $\mathbf{X}_k$ ($k = 1, K$) that are optimally fitted to this 'elastic target' by similarity transformations.

The optimal $\mathbf{W}_k$ is easily found by regression methods. Yet, there may be no particular reason for stretching $\mathbf{Z}$ along the given dimensions, and

stretching it in other directions obviously can result in different shears of the configuration. Hence, in general, what is needed is also a rotation matrix $\mathbf{S}$ so that $\mathbf{Z}\mathbf{S}\mathbf{W}_k$ optimally explains all K configurations $\mathbf{X}_k$ fitted to it.

To find $\mathbf{S}$, one may consider the individual case first, where $\mathbf{S}_k$ is an idiosyncratic rotation, that is, a different rotation $\mathbf{S}_k$ for each individual k . Hence, we want to find a rotation $\mathbf{S}_k$ and a set of dimensional weights $\mathbf{W}_k$ that transform the group space so that it optimally explains each (admissibly transformed) individual space. A direct solution for this problem is known only for the two-dimensional case (Lingoes & 1; Commandeur, 1990). This solution can be used iteratively for each plane of the space. The average of all $\mathbf{Z}\mathbf{S}_k$'s is then used as a target to solve for $\mathbf{Z}\mathbf{S}$.

In general, we thus obtain a group space that is uniquely rotated by $\mathbf{S}$. However, this uniqueness is often not very strong in the sense that other rotations are not much worse in terms of fit. This means that one must be careful when interpreting the uniqueness of dimensions delivered by programs that optimize the dimension-weighting model only, not providing fit indices for the unweighted case as benchmarks.

Another possibility to fit $\mathbf{Z}$ to each (admissibly transformed) $\mathbf{X}_k$ is to weight the rows of $\mathbf{Z}$, that is, using the model $\mathbf{V}_k\mathbf{Z}$. Geometrically, $\mathbf{V}_k$ acts on the points of $\mathbf{Z}$, shifting them along their position vectors. Thus, these weights are called 'vector weights'. Except for special cases ('perspective model', see [2]), however, vector weighting is not such a compelling psychological model as dimension weighting. However, it may provide valuable index information. For example, if we find that an optimal fitting of $\mathbf{Z}$ to each individual configuration can be done only with weights varying considerably around +1, then it makes little sense to consider the centroid

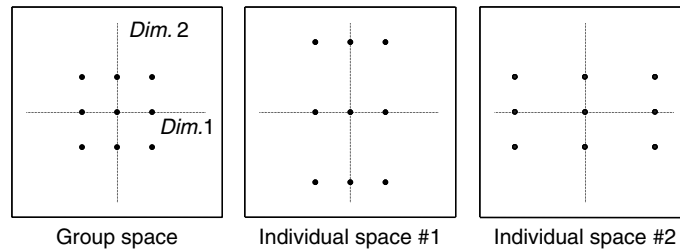


Figure 2 A group space and two individual spaces related to it by dimensional weightings

configuration $\mathbf{Z}$ as a structure that is common to all individuals.

Finding optimal vector weights is simple in the 2D case, but simultaneously to find all transformations ($\mathbf{V}_k$ and all those on $\mathbf{X}_k$) in higher-dimensional spaces appears intractable. Hence, to minimize the loss, we have to iterate over all planes of the space [11].

Finally, as in the idiosyncratic rotations in the dimension-weighting model, we can generalize the vector-weighting transformations to a model with an idiosyncratic origin. In other words, rather than fixing the perspective origin externally either at the centroid or at some other more meaningful point, it is also possible to leave it to the model to find an origin that maximizes the correspondence of an individual configuration and a transformed $\mathbf{Z}$. Formally, using both dimension and vector weighting is also possible, but as a model this fitting becomes too complex to be of any use.

The various individual differences models can be fitted by the program PINDIS [11] which is also integrated into the NewMDSX package. Langeheine [10] provides norms for fitting random configurations by unweighted Procrustean transformations and by all of the dimension and vector-weighting models.

Note

1. Procrustes is an innkeeper from Greek mythology who would fit his guests to his iron bed by stretching them or cutting their legs off.

References

- [1] Borg, I. (1978). Procrustean analysis of matrices with different row order, *Psychometrika* **43**, 277–278.
- [2] Borg, I. & Groenen, P.J.F. (1997). *Modern Multidimensional Scaling*, Springer, New York.
- [3] Browne, M.W. (1969a). On oblique Procrustes rotations, *Psychometrika* **34**, 375–394.
- [4] Browne, M.W. & Kristof, W. (1969b). On the oblique rotation of a factor matrix to a specified target, *British Journal of Mathematical and Statistical Psychology* **25**, 115–120.
- [5] Browne, M.W. (1972a). Oblique rotation to a partially specified target, *British Journal of Mathematical and Statistical Psychology* **25**, 207–212.
- [6] Browne, M.W. (1972b). Orthogonal rotation to a partially specified target, *British Journal of Mathematical and Statistical Psychology* **25**, 115–120.
- [7] Commandeur, J.J.F. (1991). *Matching Configurations*, Leiden, The Netherlands, DSWO Press.
- [8] Cliff, N. (1966). Orthogonal rotation to congruence, *Psychometrika* **31**, 33–42.
- [9] Kristof, W. (1970). A theorem on the trace of certain matrix products and some applications, *Journal of Mathematical Psychology* **7**, 515–530.
- [10] Langeheine, R. (1982). Statistical evaluation of measures of fit in the Lingo-Borg procrustean individual differences scaling, *Psychometrika* **47**, 427–442.
- [11] Lingo, J.C. & Borg, I. (1978). A direct approach to individual differences scaling using increasingly complex transformations, *Psychometrika* **43**, 491–519.
- [12] Schönemann, P.H. (1966). A generalized solution of the orthogonal Procrustes problem, *Psychometrika* **31**, 1–10.
- [13] Schönemann, P.H. & Carroll, R.M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion, *Psychometrika* **35**, 245–256.
- [14] Ten Berge, J.M. (1977). Orthogonal procrustes rotation for two or more matrices, *Psychometrika* **42**, 267–276.
- [15] Ten Berge, J.M.F., Kiers, H.A.L. & Commandeur, J.J.F. (1993). Orthogonal Procrustes rotation for matrices with missing values, *British Journal of Mathematical and Statistical Psychology* **46**, 119–134.
- [16] Verboon, P. & Heiser, W.J. (1992). Resistant orthogonal Procrustes analysis, *Journal of Classification* **9**, 237–256.
- [17] Verboon, P. (1994). *A Robust Approach to Nonlinear Multivariate Analysis*, DSWO Press, Leiden.

Further Reading

- Carroll, J.D. & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition, *Psychometrika* **35**, 283–320.
- Gower, J.C. (1975). Generalized Procrustes analysis, *Psychometrika* **40**, 33–51.
- Horan, C.B. (1969). Multidimensional scaling: combining observations when individuals have different perceptual structures, *Psychometrika* **34**, 139–165.
- Kristof, W. & Wingersky, B. (1971). Generalization of the orthogonal Procrustes rotation procedure for more than two matrices, in *Proceedings of the 79th Annual Convention of the American Psychological Association*, Washington, DC, pp. 89–90.
- Mulaik, S.A. (1972). *The Foundations of Factor Analysis*, McGraw-Hill, New York.

INGWER BORG

Projection Pursuit

WERNER STUETZLE

Volume 3, pp. 1614–1617

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Projection Pursuit

A time-honored method for detecting unanticipated ‘structure’ – clusters (*see* **Cluster Analysis: Overview**), outliers, skewness, concentration near a line or a curve – in bivariate data is to look at a **scatterplot**, using the ability of the human perceptual system for instantaneous pattern discovery. The question is how to bring this human ability to bear if the data are high-dimensional.

Scanning all 45 pairwise scatterplots of a 10-dimensional data set already tests the limits of most observers’ patience and attention span, and it is easy to construct examples where there is obvious structure in the data that will not be revealed in any of those plots. This fact is illustrated in Figures 1 and 2.

Figure 1 shows a two-dimensional data set consisting of two clearly separated clusters. We added eight independent standard Gaussian ‘noise’ variables and then rotated the resulting 10-dimensional data set into a random orientation. Visual inspection of all 45 pairwise scatterplots of the resulting 10-dimensional data fails to reveal the clusters; the scatterplot which, subjectively, appears to be most structured is shown in Figure 2.

However, we know that there do exist planes for which the projection is clustered; the question is how to find one.

Looking for Interesting Projections

The basic idea of Projection Pursuit, suggested by Kruskal [15] and first implemented by Friedman and Tukey [10], is to define a *projection index* $I(\mathbf{u}, \mathbf{v})$ measuring the degree of ‘interestingness’ of the projection onto the plane spanned by the (orthogonal) vectors $\mathbf{u}$ and $\mathbf{v}$ and then use numerical optimization to find a plane maximizing the index.

A key issue is the choice of the projection index. Probably the most familiar projection index is the variance of the projected data. A plane maximizing this index can be found by linear algebra – it is spanned by the two largest principal components (*see* **Principal Component Analysis**). In our example, however, projection onto the largest principal components (Figure 3) does not show any clustering –

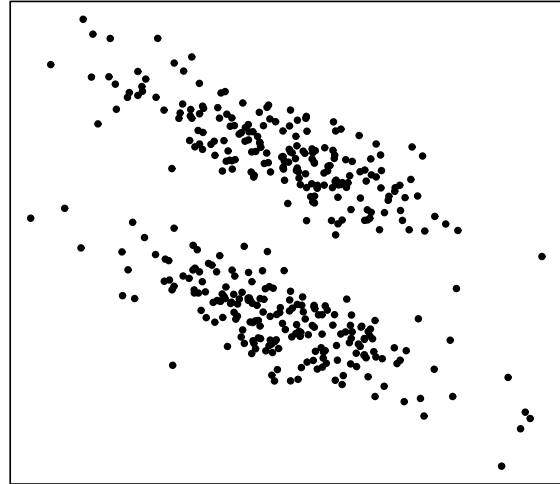


Figure 1 Sample from a bivariate mixture of two Gaussians

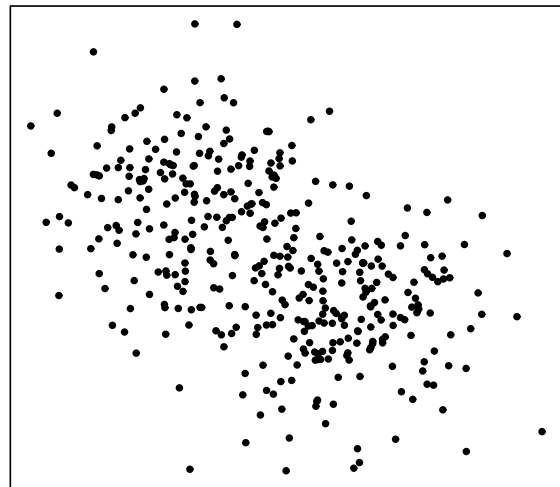


Figure 2 Most ‘structured’ pairwise scatterplot

variance is not necessarily a good measure of ‘interestingness’.

Instead, a better approach is to first sphere the data (transform it to have zero mean and unit covariance) and then use an index measuring the deviation of the projected data from a standard Gaussian distribution. This choice is motivated by two observations. First, if the data are multivariate Gaussian (*see* **Catalogue of Probability Density Functions**), then all projections will be Gaussian and Projection Pursuit will not find

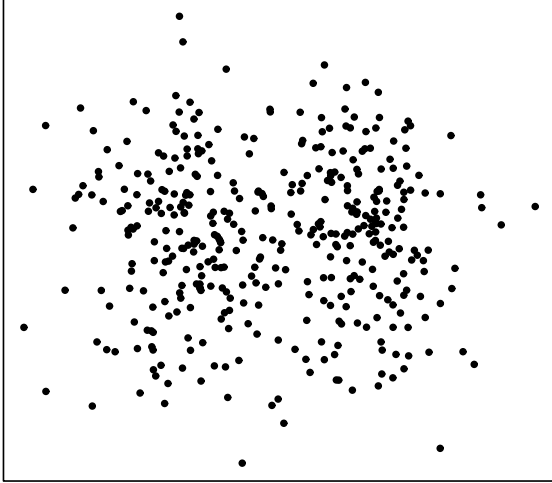


Figure 3 Projection onto largest principal components

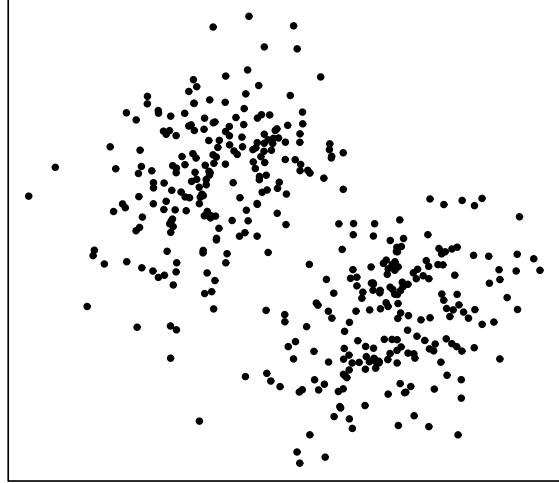


Figure 4 Projection onto plane maximizing the 'holes' index

any interesting projections. This is good, because a multivariate Gaussian distribution is completely specified by its mean and covariance matrix, and there is nothing more to be found. Second, Diaconis and Freedman [3] have shown that under appropriate conditions most projections of multivariate data are (approximately) Gaussian, which suggests regarding non-Gaussian projections as interesting.

Many projection indices measuring deviation from Gaussianity have been devised; see, for example [2, 11–14]. Figure 4 shows the projection of our simulated data onto a plane maximizing the 'holes' index [1]; the clusters are readily apparent.

Example: The Swiss Banknote Data

The Swiss Banknote data set [4] consists of measurements of six variables (width of bank note; height on left side; height on right side; lower margin; upper margin; diagonal of inner box) on 100 genuine and 100 forged Swiss bank notes. Figure 5 shows a projection of the data onto the first two principal components. The genuine bank notes, labeled '+', are clearly separated from the false ones.

Applying projection pursuit (with a Hermite index of order 7) results in the projection shown in Figure 6 (adapted from [14]).

This picture (computed without use of the class labels) suggests that there are two distinct groups

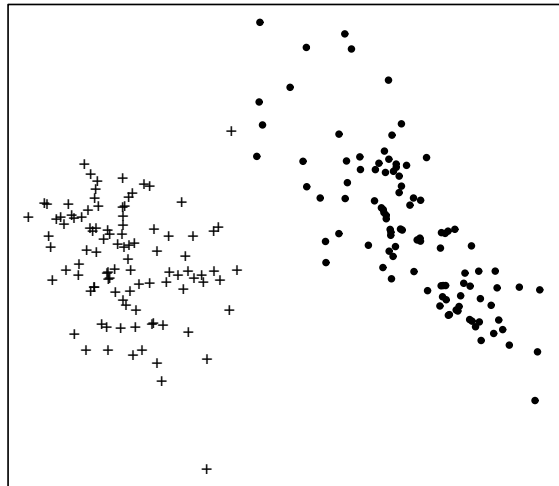


Figure 5 Projection of Swiss Banknote data onto largest principal components

of forged notes, a fact that was not apparent from Figure 5.

Projection Pursuit Modeling

In general there may be multiple interesting views of the data, possibly corresponding to multiple local maxima of the projection index. This suggests using

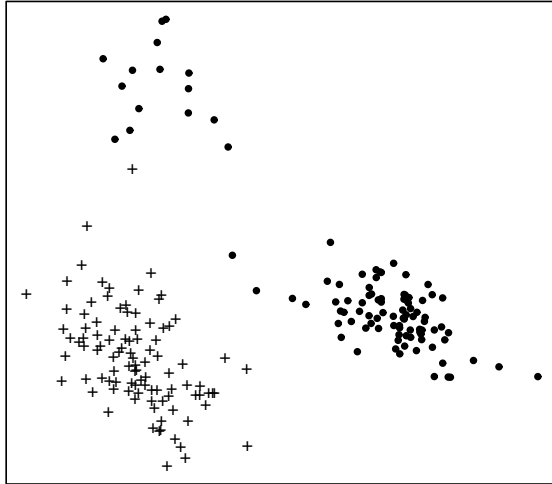


Figure 6 Projection of Swiss Banknote data onto plane maximizing the ‘Hermite 7’ index

multiple starting values for the nonlinear optimization, such as planes in random orientation (see **Optimization Methods**). A more principled approach is to remove the structure revealed in consecutive solution projections, thereby deflating the corresponding local maxima of the index. In the case where a solution projection shows multiple clusters, structure can be removed by partitioning the data set and recursively applying Projection Pursuit to the individual clusters. The idea of alternating between Projection Pursuit and structure removal was developed into a general *projection pursuit paradigm for multivariate analysis* by Friedman and Stuetzle [9]. The Projection Pursuit paradigm has been applied to density estimation [6, 8, 12, 13], regression [7], and classification [5].

Software

Projection Pursuit is one of the many tools for visualizing and analyzing multivariate data that together make up the *Ggobi Data Visualization System*. Ggobi is distributed under an AT&T open source license. A self-installing Windows binary or Linux/Unix versions as well as accompanying documentation can be downloaded from www.ggobi.org.

References

- [1] Cook, D., Buja, A. & Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions,

Journal of Computational and Graphical Statistics **2**(3), 225–250.

- [2] Cook, D., Buja, A., Cabrera, J. & Hurley, C. (1995). Grand tour and projection pursuit, *Journal of Computational and Graphical Statistics* **4**(3), 155–172.
- [3] Diaconis, P. & Freedman, D. (1984). Asymptotics of graphical projection pursuit, *Annals of Statistics* **12**, 793–815.
- [4] Flury, B. & Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces, *Journal of the American Statistical Association* **76**, 757–765.
- [5] Friedman, J.H. (1985). Classification and multiple regression through projection pursuit, Technical Report LCS-12, Department of Statistics, Stanford University.
- [6] Friedman, J.H. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**, 249–266.
- [7] Friedman, J.H. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.
- [8] Friedman, J.H., Stuetzle, W. & Schroeder, A. (1984). Projection pursuit density estimation, *Journal of the American Statistical Association* **79**, 599–608.
- [9] Friedman, J.H. & Stuetzle, W. (1982). Projection pursuit methods for data analysis, in *Modern Data Analysis*, R.L., Launer & A.F., Siegel, eds, Academic Press, New York, pp. 123–147.
- [10] Friedman, J.H. & Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE Transactions on Computers* **C-23**, 881–890.
- [11] Hall, P. (1989). Polynomial projection pursuit, *Annals of Statistics* **17**, 589–605.
- [12] Huber, P.J. (1985). Projection pursuit, *Annals of Statistics* **13**, 435–525.
- [13] Jones, M.C. & Sibson, R. (1987). What is projection pursuit, (with discussion), *Journal of the Royal Statistical Society, Series A* **150**, 1–36.
- [14] Klinke, S. (1995). Exploratory projection pursuit—the multivariate and discrete case, in W. Kloesgen, P. Nanopoulos & A. Unwin, eds, *Proceedings of NTTS '95*, Bonn, 247–262.
- [15] Kruskal, J.B. (1969). Towards a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new “index of condensation”, in *Statistical Computation*, R.C., Milton & J.A., Nelder, eds, Academic Press, New York, pp. 427–440.

(See also **Hierarchical Clustering; k-means Analysis; Minimum Spanning Tree; Multidimensional Scaling**)

WERNER STUETZLE

Propensity Score

RALPH B. D'AGOSTINO JR

Volume 3, pp. 1617–1619

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Propensity Score

Observational studies occur frequently in behavioral research. In these studies, investigators have no control over the treatment assignment. Therefore, large differences on observed covariates in the two groups may exist, and these differences could lead to biased estimates of treatment effects. The *propensity score* for an individual, defined as the conditional probability of being treated given the individual's **covariates**, can be used to balance the covariates in the two groups, and thus reduce this bias.

In a randomized experiment, the randomization of units (i.e., subjects) to different treatments guarantees that on average there should be no systematic differences in observed or unobserved covariates (i.e., bias) between units assigned to the different treatments (*see Clinical Trials and Intervention Studies*). However, in a nonrandomized observational study, investigators have no control over the treatment assignment, and therefore direct comparisons of outcomes from the treatment groups may be misleading. This difficulty may be partially avoided if information on measured covariates is incorporated into the study design (e.g., through matched sampling (*see Matching*)) or into estimation of the treatment effect (e.g., through **stratification** or covariance adjustment). Traditional methods of adjustment (matching, stratification, and covariance adjustment) are often limited since they can only use a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this limitation.

Formally, the propensity score for an individual is the probability of being treated conditional on (or based only on) the individual's covariate values. Intuitively, the propensity score is a measure of the likelihood that a person would have been treated using only their covariate scores. The propensity score is a balancing score and can be used in observational studies to reduce bias through the adjustment methods mentioned above.

Definition

With complete data, the propensity score for subject i ($i = 1, \dots, N$) is the conditional probability of

assignment to a particular treatment ($Z_i = 1$) versus control ($Z_i = 0$) given a vector of observed covariates, x_i :

$$e(x_i) = pr(Z_i = 1 | X_i = x_i). \quad (1)$$

where it is assumed that, given the X 's, the Z_i are independent:

$$\begin{aligned} pr(Z_1 = z_1, \dots, Z_N = z_N | X_1 = x_1, \dots, X_N = x_N) \\ = \prod_{i=1}^N e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i}. \end{aligned} \quad (2)$$

The propensity score is the 'coarsest function' of the covariates that is a balancing score, where a balancing score, $b(X)$, is defined as 'a function of the observed covariates X such that the conditional distribution of X given $b(X)$ is the same for treated ($Z = 1$) and control ($Z = 0$) units' [2]. For a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score if the treatment assignment is strongly ignorable given the covariates. Thus, matching, subclassification (stratification), or regression (covariance) adjustment on the propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable. Treatment assignment is considered strongly ignorable if the treatment assignment, Z , and the response, Y , are known to be conditionally independent given the covariates, X (i.e., when $Y \perp Z | X$).

When covariates contain no missing data, the propensity score can be estimated using **discriminant analysis** or **logistic regression**. Both of these techniques lead to estimates of probabilities of treatment assignment conditional on observed covariates. Formally, the observed covariates are assumed to have a multivariate normal distribution (conditional on Z) when discriminant analysis is used, whereas this assumption is not needed for logistic regression.

A frequent question is, 'Why must one estimate the probability that a subject receives a certain treatment since it is known for certain which treatment was given?' An answer to this question is that if one uses the probability that a subject would have been treated (i.e., the propensity score) to adjust the

estimate of the treatment effect, one can create a ‘quasirandomized’ experiment; that is, if two subjects are found, one in the treated group and one in the control, with the same propensity score, then one could imagine that these two subjects were ‘randomly’ assigned to each group in the sense of being equally likely to be treated or controlled. In a controlled experiment, the randomization, which assigns pairs of individuals to the treated and control groups, is better than this because it does not depend on the investigator conditioning on a particular set of covariates; rather, it applies to any set of observed or unobserved covariates. Although the results of using the propensity scores are conditional only on the observed covariates, if one has the ability to measure many of the covariates that are believed to be related to the treatment assignment, then one can be fairly confident that approximately unbiased estimates for the treatment effect can be obtained.

Uses of Propensity Scores

Currently in observational studies, propensity scores are used primarily to reduce bias and increase precision. The three most common techniques that use the propensity score are matching, stratification (also called *subclassification*), and regression adjustment. Each of these techniques is a way to make an adjustment for covariates prior to (matching and stratification) or while (stratification and regression adjustment) calculating the treatment effect. With all three techniques, the propensity score is calculated the same way, but once it is estimated, it is applied differently. Propensity scores are useful for these techniques because by definition the propensity score is the conditional probability of treatment given the observed covariates $e(U) = pr(Z = 1|X)$, which implies that Z and X are conditionally independent given $e(X)$. Thus, subjects in treatment and control groups with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on their background covariates. Exact

adjustments made using the propensity score will, on average, remove all of the bias in the background covariates. Therefore, bias-removing adjustments can be made using the propensity scores rather than all of the background covariates individually.

Summary

Propensity scores are being widely used in statistical analyses in many applied fields. The propensity score methodology appears to produce the greatest benefits when it can be incorporated into the design stages of studies (through matching or stratification). These benefits include providing more precise estimates of the true treatment effects as well as saving time and money. This savings results from being able to avoid recruitment of subjects who may not be appropriate for particular studies. In addition, it is important to note that propensity scores can (and often should) be used in addition to traditional methods of analysis rather than in place of these other methods. The propensity score should be thought of as an additional tool available to the investigators as they try to estimate the effects of treatments in studies.

Further references on the propensity score can be found in D’Agostino Jr [1] where applied illustrations are presented and in Rosenbaum and Rubin [2] where the theoretical properties of the propensity score are developed in depth.

References

- [1] D’Agostino Jr, R.B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, *Statistics in Medicine* **17**, 225–2281.
- [2] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

RALPH B. D’AGOSTINO JR

Prospective and Retrospective Studies

JANICE MARIE DYKACZ

Volume 3, pp. 1619–1621

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Prospective and Retrospective Studies

In mathematical or statistical applications to ‘real-world’ problems, the researcher often classifies variables as either x -variables (explanatory variables, factors, or covariates) or y -variables (response or outcome variables). Often the important research question is if and how an x -variable affects a y -variable. As an example, in a medical study, the x -variable may be smoking status (smoker or nonsmoker), and the y -variable may be lung cancer status (present or absent). Does smoking status affect the probability of acquiring lung cancer? There is a built-in direction between the x - and y -variables; x affects y . Mathematically, y is a function of x .

In a prospective study, the values of x are set or specified. Individuals (or perhaps units such as families or experimental animals) are then followed over time, and the value of y is determined. For example, individuals who do not have lung cancer are classified as smokers or nonsmokers (the x -variable) and followed over a period of years. The occurrence of lung cancer (the y -variable) is determined for the smokers and nonsmokers. By contrast, in a **retrospective study**, the values of y are set or specified. We examine the histories of individuals in the study, and the values of the x -variable are then determined. For example, lung cancer cases (those who have lung cancer) and controls (those who do not have lung cancer) are selected, the histories of these individuals are obtained, and then their smoking status is determined. In short, a prospective study starts with values of the x -variable; a retrospective study starts with values of the y -variable. For both types of studies, we would like to know if x influences y . Prospective studies are sometimes called *longitudinal* or **cohort studies**. Retrospective studies are sometimes called **case-control studies**.

Table 1 represents a hypothetical example of a prospective study. The x -variable has two values, Exposed ($x = 1$) and Not exposed ($x = 0$). The study starts with 1000 individuals with the x -variable = 1 (Exposed) and 1000 individuals with the x -variable = 0 (Not exposed). Individuals are followed over time, and at some point the number of individuals who acquire the disease are counted. We determine the

Table 1 Prospective study

	Disease present $y = 1$	Disease absent $y = 0$	
Exposed ($x = 1$)	100	900	1000
Not exposed ($x = 0$)	10	990	1000

number of those with the y -variable = 1 (Disease Present) and $y = 0$ (Disease Absent).

In a prospective study, the **relative risk** can be directly calculated. In the example above, the probability that individuals acquire the disease given that they are exposed is $100/1000 = 0.1$. The probability that individuals acquire the disease given that they are not exposed is $10/1000 = 0.01$. The relative risk is the ratio of these two probabilities. Comparing those exposed with those not exposed, the relative risk is $0.1/0.01 = 10$. Thus, those exposed are ten times more likely to acquire the disease than those not exposed.

Table 2 represents a hypothetical example of a retrospective study. The y -variable has two values, Disease Present ($y = 1$) and Disease Absent ($y = 0$). Those in the group with Disease Present are called *cases*, and those in the group with Disease Absent are called *controls*. The study starts with 110 individuals having y -variable = 1 (the cases), and with 1890 individuals having y -variable = 0 (the controls). We examine the histories of these individuals and determine whether the individual was exposed or not. That is, we count those with the x -variable = 1 (Exposed) and $y = 0$ (Not exposed).

In a retrospective study, the relative risk cannot be directly calculated, but the odds ratio can be calculated. In the example above, for those exposed, the odds of those acquiring the disease relative to those not acquiring the disease are $100/900$. For those not exposed, the odds of those acquiring the disease relative to those not acquiring the disease are $10/990$.

Table 2 Retrospective study

	Disease present $y = 1$	Disease absent $y = 0$
Exposed ($x = 1$)	100	900
Not exposed ($x = 0$)	10	990
	110	1890

2 Prospective and Retrospective Studies

The odds ratio is the ratio of these two fractions and is $(100/900)/(10/990)$ or 11. Recall that the same numbers in the table were used for the example of a prospective study, and the relative risk was 10. When the disease is rare, the odds ratio is used as an approximation to the relative risk.

In a prospective study, the relative risk is often used to measure the association between x and y . However, if a logistic regression model is used to determine the effect of x on y , the association between x and y can be described by the odds ratio, which occurs naturally in this model. In a retrospective study, the odds ratio is often used to measure the association between x and y . However, the odds ratio is sometimes used as an approximation to the relative risk.

Prospective studies can be expensive and time consuming because individuals or units are followed over time to determine the value of the y -variable (for example, whether an individual acquires the disease or not). However, prospective studies provide a direct estimate of relative risk and a direct estimate of the probability of acquiring the disease, given the value of the x -variable (that is, for those exposed and for those not exposed). In a prospective study, many y -variables (diseases) can be examined. Examples of prospective studies in the medical literature are found in [1, 2, 5].

By contrast, a retrospective study is generally less expensive. Retrospective studies are often used to study a rare disease because very few cases of the disease might be found in a prospective study. However, the odds ratio, and not the relative risk, is directly estimated. Only one y -variable can be studied because individuals are selected at the beginning of the study, for say, $y = 1$ (Disease Present) and $y = 0$ (Disease Absent). Because the x -variable is determined by looking into the past, there are concerns about the recollections of individuals. Examples of retrospective studies in the medical literature are found in [3, 4].

Related types of studies are **cross-sectional** studies and randomized **clinical trials** (RCT). In a cross-sectional study, the values of a x -variable and a

y -variable are measured for individuals or units at a point in time. Individuals are not followed over time, so, for example, the probability of acquiring a disease cannot be measured. Cross-sectional studies are used to determine the magnitude of a problem or to assess the nature of associations between the x - and y -variables. Randomized clinical trials are similar to prospective studies, but individuals are randomly assigned values of the x -variable (for example, an individual is randomly assigned to receive a drug, $x = 1$, or not, $x = 0$). Individuals are followed over time to determine whether they acquire a disease ($y = 1$) or not ($y = 0$). **Randomization** generally guarantees that individuals are alike except for the value of their x -variable (drug or no drug).

References

- [1] Calle, E.E., Rodriguez, C., Walker-Thurmond, K. & Thun, M.J. (2003). Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults, *The New England Journal of Medicine* **348**(17), 1625–1638.
- [2] Iribarren, C., Tekawa, I.S., Sidney, S. & Friedman, G.D. (1999). Effect of cigar smoking on the risk of cardiovascular disease, chronic obstructive pulmonary disease, and cancer in men, *The New England Journal of Medicine* **340**(23), 1773–80.
- [3] Karagas, M.R., Stannard, V.A., Mott, L.A., Slattery, M.J., Spencer, S.K. & Weinstock, M.A. (2002). Use of tanning devices and risk of basal cell and squamous cell skin cancers, *Journal of the National Cancer Institute*, **94**(3), 224–6.
- [4] Koepsell, T., McCloskey, L., Wolf, M., Moudon, A.V., Buchner, D., Kraus, J. & Patterson, M. (2002). Crosswalk markings and the risk of pedestrian-motor vehicle collisions in older pedestrians, *JAMA* **288**(17), 2136–2143.
- [5] Mukamal, K.J., Conigrave, K.M., Mittleman, M.A., Camargo Jr, C.A., Stampfer, M.J., Willett, W.C. & Rimm, E.B. (2003). Roles of drinking pattern and type of alcohol consumed in coronary heart disease in men, *The New England Journal of Medicine* **348**(2), 109–118.

JANICE MARIE DYKACZ

Proximity Measures

HANS-HERMANN BOCK

Volume 3, pp. 1621–1628

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Proximity Measures

Introduction: Definitions and Examples

Proximity measures characterize the *similarity* or *dissimilarity* that exists between the objects, items, stimuli, or persons that underlie an empirical study. In contrast to cases where we distinguish only between ‘similar’ and ‘dissimilar’ objects i and j (in a binary, qualitative way), proximities measure the *degree* of similarity by a real number s_{ij} , typically between 0 and 1: the larger the value s_{ij} , the larger the similarity between i and j , and $s_{ij} = 1$ means maximum similarity. A dual approach measures the *dissimilarity* between i and j by a numerical value $d_{ij} \geq 0$ (with 0 the minimum dissimilarity). In practice, there are many ways to find appropriate values s_{ij} or d_{ij} , for example,

- s_{ij} the degree of friendship between two persons i and j
- s_{ij} the relative frequency of common descriptors shared by two documents i and j
- s_{ij} the (relative) number of symptoms that are shared by two patients i and j
- d_{ij} the road distance (or transportation costs) between two cities i and j
- d_{ij} the Euclidean distance between two (data) points x_i and x_j in $\mathbb{R}^p$
- d_{ij} the number of nonmatching scores in the results of two test persons i and j .

Given a set $\mathcal{X} = \{1, \dots, n\}$ of n objects, the corresponding *similarity matrix* $S = (s_{ij})$ of size $n \times n$, or its dual (Table 1), a *dissimilarity matrix* $D = (d_{ij})$ (Table 2) provides a quantitative information on the overall similarity structure of the set of objects, items, and so on. Such information is the basis for various statistical and data analytical techniques, in particular:

Table 1 A 6×6 similarity matrix

$S =$	$\begin{pmatrix} 1.0 & 0.8 & 0.4 & 0.7 & 0.1 & 0.3 \\ 0.8 & 1.0 & 0.5 & 0.9 & 0.2 & 0.1 \\ 0.4 & 0.5 & 1.0 & 0.3 & 0.7 & 0.8 \\ 0.7 & 0.9 & 0.3 & 1.0 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.7 & 0.1 & 1.0 & 0.5 \\ 0.3 & 0.1 & 0.8 & 0.2 & 0.5 & 1.0 \end{pmatrix}$
-------	--

Table 2 A 6×6 dissimilarity matrix

$D =$	$\begin{pmatrix} 0.0 & 3.0 & 8.0 & 1.5 & 10.5 & 6.0 \\ 3.0 & 0.0 & 5.0 & 1.5 & 7.5 & 3.0 \\ 8.0 & 5.0 & 0.0 & 6.5 & 2.5 & 2.0 \\ 1.5 & 1.5 & 6.5 & 0.0 & 9.0 & 4.5 \\ 10.5 & 7.5 & 2.5 & 9.0 & 0.0 & 4.5 \\ 6.0 & 3.0 & 2.0 & 4.5 & 4.5 & 0.0 \end{pmatrix}$
-------	--

- for a *detailed analysis* of the network of similarities among selected (classes of) individuals,
- for an *illustrative visualization* of the similarity structure by graphical displays, and
- for *clustering* the objects of $\mathcal{X}$ into homogeneous classes of mutually similar elements.

Formal Specifications

Formally, given a set of objects (individuals, items, points, elements) $\mathcal{X}$, a *dissimilarity measure* d on $\mathcal{X}$ is a function that assigns to each pair (x, y) of elements x, y of $\mathcal{X}$ a real number $d(x, y)$ such that for all x, y the conditions

- $d(x, y) \geq 0$
- $d(x, x) = 0$ (symmetry) (1)
- $d(x, y) = d(y, x)$

are fulfilled. d is called a *pseudometric* if, additionally, the *triangle inequality*

- $d(x, y) \leq d(x, z) + d(z, y)$ (2)
- for all $x, y, z \in \mathcal{X}$

is fulfilled. A pseudometric with $[(v)d(x, y) = 0 \text{ holds only for } x = y]$ (definiteness) is called a *metric* or a *distance* measure. Note that in practice, the term ‘distance’ is often used for a dissimilarity of whatever type. Analogous conditions may be formulated for characterizing a *similarity measure* $S = (s_{ij})$:

- $0 \leq s_{ij} \leq 1$ (i')
- $s_{ii} = 1$ (ii')
- $s_{ij} = s_{ji}$ (iii') (3)

for all i, j .

Commonly Used Proximity Measures

Depending on the practical situation and the data, various different proximity measures can be selected.

2 Proximity Measures

This selection must match the underlying or intended similarity concept from, for example, psychology, medicine, or documentation. In the following we list some formal definitions which are commonly used in statistics and data analysis (see [2, 11, 10, 9]).

Dissimilarities

If $x = (x_1, \dots, x_p)'$ and $y = (y_1, \dots, y_p)'$ are two elements of $\mathbb{R}^p$, for example, two data points with p real-valued components, the *Euclidean distance*

$$d(x, y) := \|x - y\|_2 := \left[\sum_{\ell=1}^p (x_\ell - y_\ell)^2 \right]^{1/2} \quad (4)$$

and the *Manhattan distance*

$$d(x, y) := \|x - y\|_1 := \sum_{\ell=1}^p |x_\ell - y_\ell| \quad (5)$$

provide dissimilarity measures that are symmetric in all p components. If the components provide different contributions for the intended similarity concept, we may use weighted versions such as

$$d(x, y) := \left[\sum_{\ell=1}^p w_\ell (x_\ell - y_\ell)^2 \right]^{1/2} \quad \text{and} \quad d(x, y) := \sum_{\ell=1}^p w_\ell |x_\ell - y_\ell| \quad (6)$$

with suitable weights $w_1, \dots, w_p > 0$. The *Mahalanobis distance*

$$d_M(x, y) := \|x - y\|_{\Sigma^{-1}} := \sqrt{(x - y)' \Sigma^{-1} (x - y)} = \left[\sum_{\ell=1}^p \sum_{s=1}^p (x_\ell - y_\ell) \sigma^{\ell s} (x_s - y_s) \right]^{1/2} \quad (7)$$

takes into account an eventual dependence between the p components that is summarized by a corresponding covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$ (theoretical, empirical, within/between classes, etc.) and its inverse $\Sigma^{-1} = (\sigma^{\ell s})$. In fact, (7) is the Euclidean distance $d_M(x, y) = \|\tilde{x} - \tilde{y}\|_2$ of the linearly transformed data points $\tilde{x} = \Sigma^{-1/2}x$ and $\tilde{y} = \Sigma^{-1/2}y \in \mathbb{R}^p$ in a space of **principal components**.

In the case of *qualitative data*, each component x_ℓ of the data vector $x = (x_1, \dots, x_p)'$ refers to a property such as nationality (French, German, Italian, ...), color (red, blue, yellow, ...), or material (iron, copper, zinc, ...) and takes its 'values' in a finite set $\mathcal{Z}_\ell$ of categories or alternatives. In this case we measure the dissimilarity between two data configurations $x = (x_1, \dots, x_p)'$ and $y = (y_1, \dots, y_p)'$ typically by the number of nonmatching components, that is, the *Hamming distance*

$$d_H(x, y) := \#\{ \ell \mid x_\ell \neq y_\ell \}. \quad (8)$$

From a wealth of other dissimilarity measures we mention just the *Levenshtein distance* d_L that measures, for example, the dissimilarity of two words (DNA strains, messages, ...) $x = (x_1, \dots, x_p)$ and $y = (y_1, \dots, y_q)$: Here $d_L(x, y)$ is the minimum number of operations (= deletion of a letter, insertion of a letter, insertion of a space) that is needed in order to transform the word x into the other one y (see [14]).

Similarity Measures

Quite generally, a similarity measure s can be obtained from a dissimilarity measure d by a decreasing function h such as, for example, $s = h(d) = 1 - e^{-d}$ or $s = (d_0 - d)/d_0$ (where d_0 is the maximum observed dissimilarity value). However, a range of special definitions has been proposed for *binary data* where each component takes only two alternatives 1 and 0 (e.g., female/male, yes/no, present/absent) such that $\mathcal{Z}_\ell = \{0, 1\}$ for all ℓ . In this case, the matching properties of two data vectors x and y are typically summarized in a 2×2 table (see Table 3) where the integers a, b, c, d (with $a + b + c + d = p$) denote the number of components of x and y with $x_\ell = y_\ell = 0$ (negative matching), $(x_\ell, y_\ell) = (0, 1)$, and $(x_\ell, y_\ell) = (1, 0)$ (mismatches), and $x_\ell = y_\ell = 1$ (positive matching), respectively. An example is given

Table 3 2×2 matching table

x/y	0	1	Σ
0	a	b	$a + b$
1	c	d	$c + d$
Σ	$a + c$	$b + d$	p

Table 4 Binary data vectors x, y with $n = 10$, $a = 4$, $b = 1$, $c = 2$, $d = 3$

$x =$	(0 0 1 1 0 1 1 0 0 1)
$y =$	(0 0 1 1 0 1 0 0 1 0)

in Table 4. In this notation, the following similarity measures have been proposed (see, e.g., [2]):

- the simple *matching coefficient*

$$s_M(x, y) := \frac{a + d}{a + b + c + d}$$

- the *Tanimoto coefficient*

$$s_T(x, y) = \frac{a + d}{a + 2(b + c) + d}$$

that are both symmetric with respect to the categories 1 and 0. On the other hand, there are applications where the simultaneous absence of an attribute in both x and y bears no similarity information. Then we may use asymmetric measures such as

- the *Jaccard* or *S coefficient*

$$s_S(x, y) := \frac{d}{d + b + c}.$$

All these measures have the form $s(x, y) = (d + \tau a) / (d + \tau a + \sigma(b + c))$ with suitable factors σ, τ .

Mixed Data

Problems are faced when considering *mixed data* where some components of the data vectors x and y are qualitative (categorical or ordinal) and some are quantitative (numerical) (see **Scales of Measurement**). In this case, weighted averages of the type

$$d(x, y) := \alpha_1 d_1(x', y') + \alpha_2 d_2(x'', y'') + \alpha_3 d_3(x''', y''') \quad (9)$$

are typically used where $d_1(x', y')$, $d_2(x'', y'')$, $d_3(x''', y''')$ are partial dissimilarities defined for the quantitative, categorical, and ordinal parts (x', y') , (x'', y'') , (x''', y''') of (x, y) whereas α_1, α_2 , and α_3 are suitable weights (see [12]). A main problem here

is to find a weighting scheme that guarantees the comparability of, and balance between, the three different types of components.

Inequalities for Dissimilarity Measures

All practically relevant definitions of a dissimilarity measure d between elements x, y of a set $\mathcal{X}$ share various common-sense properties in analogy to, for example, a geographical distance (in the plane or on the sphere). Some of these properties are expressed by inequalities that relate the dissimilarities of several pairs of objects (such as the triangle inequality (2)) and are motivated by graphical visualizations that are possible just for a special dissimilarity type (see the section on “Visualization of Dissimilarity Structures”).

Ultrametrics and the Ultrametric Inequality

A dissimilarity measure d on a set $\mathcal{X}$ is called a *ultrametric* or a *ultrametric distance* if and only if it fulfills the ultrametric inequality

$$d(x, y) \leq \max \{d(x, z), d(z, y)\} \quad (10)$$

for all x, y, z in $\mathcal{X}$

Ultrametric distances are related to hierarchical classifications (see section on ‘Hierarchical Clustering and Ultrametric Distances’), but are also met in the context of number theory and physics. Any ultrametric is a metric since (10) implies (2).

Additive Distances and the Four-point Inequality

A dissimilarity measure d is called an *additive* or *tree distance* if and only if it fulfills the *four-points inequality*

$$\begin{aligned} & d(x, y) + d(u, v) \\ & \leq \max \{d(x, u) + d(y, v), d(x, v) + d(y, u)\} \\ & \text{for all } x, y, u, v \text{ in } \mathcal{X} \end{aligned} \quad (11)$$

Such a distance corresponds to a tree-like visualization of the objects (see Figure 4).

Visualization of Dissimilarity Structures

A major approach for interpreting and understanding a similarity structure of n objects $i = 1, \dots, n$ proceeds by visualizing the observed dissimilarity values d_{ij} in a graphical display where the objects are represented by n points (vertices) $y_1, \dots, y_n$ and the dissimilarity d_{ij} of two objects i, j is reproduced or approximated by the ‘closeness’ (in some sense) of the vertices y_i, y_j in this display.

Euclidean Embedding and Multidimensional Scaling (MDS)

Multidimensional Scaling represents the n objects by n points $y_1, \dots, y_n$ in a low-dimensional Euclidean space $\mathbb{R}^s$ (usually with $s = 2$ or $s = 3$) such that their *Euclidean distances* $\delta(y_i, y_j) := \|y_i - y_j\|_2$ approximate as much as possible the given dissimilarities d_{ij} . An optimum configuration $Y := \{y_1, \dots, y_n\}$ is typically found to be minimizing an optimality criterion such as

$$\text{STRESS}(Y) := \sum_{i=1}^n \sum_{j=1}^n (d_{ij} - \delta(y_i, y_j))^2 \rightarrow \min_Y \quad (12)$$

$$\text{SSTRESS}(Y) := \sum_{i=1}^n \sum_{j=1}^n (d_{ij}^2 - \delta(y_i, y_j)^2)^2 \rightarrow \min_Y. \quad (13)$$

Nonmetric Scaling assumes a linear relationship $d = a + b\delta$ or a monotone function $d = g(\delta)$ between the d_{ij} and the Euclidean distances $\delta(y_i, y_j)$ and minimizes:

$$\sum_{i=1}^n \sum_{j=1}^n (d_{ij} - a - b\delta(y_i, y_j))^2 \quad \text{or} \quad \sum_{i=1}^n \sum_{j=1}^n (d_{ij}^2 - g(\delta(y_i, y_j)))^2 \quad (14)$$

with respect to Y and the coefficients $a, b \in \mathbb{R}$ or, respectively, the function g . There are many variants of this approach with suitable software support (see [7]).

If the minimum STRESS value in (12) is 0 for some dimension s , this means that there exists

a configuration $Y = (y_1, \dots, y_n)$ in $\mathbb{R}^s$ such that $\delta(y_i, y_j) = d_{ij}$ holds for all pairs (i, j) : Such a matrix D is called a *Euclidean dissimilarity matrix* since it allows an *exact* embedding of the objects in a Euclidean space $\mathbb{R}^s$. A famous theorem states that this case occurs exactly if the row-and-column centered matrix $\tilde{D} = (d_{ij}^2)$ is negative semidefinite with a rank less or equal to s (see, e.g., [7]).

Hierarchical Clustering and Ultrametric Distances

Ultrametric distances (see **Ultrametric Trees**) are intimately related to hierarchical classifications and their visualizations. A **hierarchical clustering** or dendrogram $(\mathcal{H}, h)$ for a set $\mathcal{X} = \{1, \dots, n\}$ of objects is a system $\mathcal{H} = \{A, B, C, \dots\}$ of classes of objects that are hierarchically nested as illustrated in Figure 1 and where each class A is displayed at a (heterogeneity) level $h(A) \geq 0$ such that larger classes have a higher level: $h(A) \leq h(B)$ whenever $A \subset B$. There exist plenty of clustering methods in order to construct a hierarchical clustering of the n objects on the basis of a given *dissimilarity matrix* $D = (d_{ij})$. The resulting dendrogram may then reveal and explain the underlying similarity structure.

Inversely, given a dendrogram $(\mathcal{H}, h)$, we can measure the dissimilarity of two objects i, j within the hierarchy by the level $h(B)$ of the smallest class

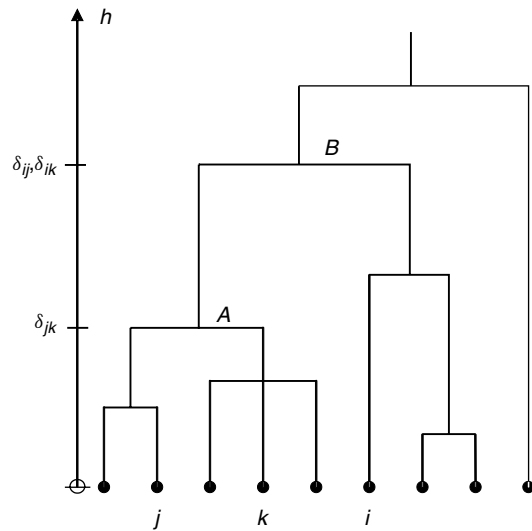


Figure 1 Dendrogram for $n = 9$ objects

B that contains both objects i and j (see Figure 1). If this level is denoted by $\delta_{ij} := h(B)$, we obtain a new dissimilarity matrix $\Delta = (\delta_{ij})$. As a matter of fact, it can be shown that Δ is an ultrametric distance.

Even more: It appears that dendrograms and ultrametrics are equivalent insofar as a dissimilarity matrix $D = (d_{ij})_{n \times n}$ is an ultrametric if and only if there exists a hierarchical clustering $(\mathcal{H}, h)$ of the n objects such that the resulting ultrametric Δ is identical to D : $d_{ij} = \delta_{ij}$ for all i, j . This suggests that in cases when D does not share the ultrametric property, the criterion

$$\Phi(\Delta) := \sum_{i=1}^n \sum_{j=1}^n (\delta_{ij} - d_{ij})^2 \quad (15)$$

may be used for checking if $(\mathcal{H}, h)$ is an adequate classification for the given dissimilarity D , and looking for a good hierarchical clustering for D can also be interpreted as looking for an optimum ultrametric approximation Δ of D in the sense of minimizing $\Phi(\Delta)$ (see [8]).

Pyramids and Robinsonian Dissimilarities

If the elements d_{ij} of a dissimilarity matrix $D = (d_{ij})$ are monotonely increasing when moving away from the diagonal (with zero values) along a row or a column, D is called a *Robinsonian dissimilarity* (see Table 2 after rearranging the rows and columns in the order 1-4-2-6-3-5). There exists a close relationship between Robinsonian matrices and a special type of nested clusterings (pyramids) that is illustrated in Figure 2: A *pyramidal clustering* $(\mathcal{P}, <, h)$ is a system $\mathcal{P} = \{A, B, C, \dots\}$ of classes of objects, together with an order $<$ on the set $\mathcal{O}$ of objects such that each class $A \in \mathcal{P}$ is an interval with respect to $<$: $A = \{k \in \mathcal{O} \mid i < k < j\}$ for some objects $i < j$. As before, $h(A)$ is the heterogeneity level of a class A . Figure 2 illustrates also the general theorem that (a) the intersection $A \cap B$ of two classes of $A, B \in \mathcal{H}$ is either empty or belongs to $\mathcal{H}$ as well and (b) that any class $C \in \mathcal{H}$ can have at most two ‘daughter’ classes A, B from $\mathcal{H}$ (see, e.g., [13, 5]). In practice, the corresponding order $<$ provides a *seriation* of the n objects in a series that may describe, e.g., temporal changes, chronological periods, or geographical gradients.

Similarly as in section on ‘Hierarchical Clustering and Ultrametric Distances’, we can define, for a

given pyramid, a dissimilarity r_{ij} between two objects i, j *within the pyramid* by the smallest heterogeneity level $h(B)$ where i and j meet in the same class B of $\mathcal{H}$. A major result states essentially (a) that this dissimilarity matrix $(r_{ij})_{n \times n}$ is Robinsonian and (b) that, inversely, any Robinsonian dissimilarity matrix can be obtained in this way from an appropriate pyramidal classification $(\mathcal{P}, <, h)$. Insofar both concepts are equivalent. For example, the pyramid in Figure 2 corresponds to the dissimilarity matrix from Table 1. For details see, e.g., [15].

As an application, we may calculate, for a given arbitrary dissimilarity matrix $D = (d_{ij})$, an optimum Robinsonian approximation (r_{ij}) (in analogy to (15)). Then the corresponding pyramidal clustering provides not only insight into the similarity structure of D , but also suggests what kind of seriation may be appropriate and interpretable for the underlying objects.

Tree-like Configurations, Additive and Quadripolar Distances

Trivially, any dissimilarity matrix $D = (d_{ij})$ may be visualized by a weighted graph G where each of the n objects i is represented by a vertex y_i and each of the $n(n-1)/2$ pairs $\{i, j\}$ of vertices are linked by an edge $\bar{ij}$ that has the weight (length) $d_{ij} > 0$ (Figure 3). For a large number of objects, this graph is too complex for interpretation, therefore we should concentrate on simpler graphs.

A major approach considers the interpretation of $D = (d_{ij})$ in terms of a weighted tree T with n vertices and $n-1$ edges (Figure 4) such that d_{ij} is the length of the shortest path from i to j in this tree.

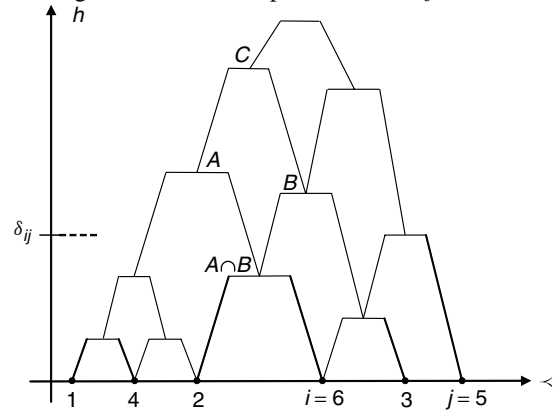


Figure 2 Pyramid for $n = 6$ objects with order $<$

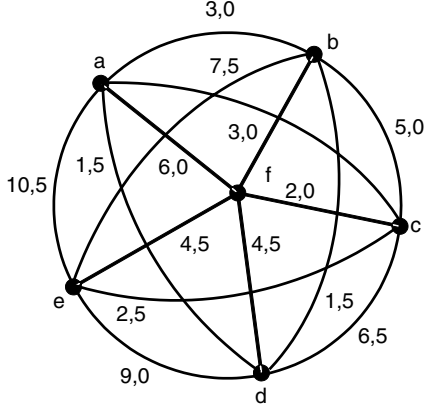
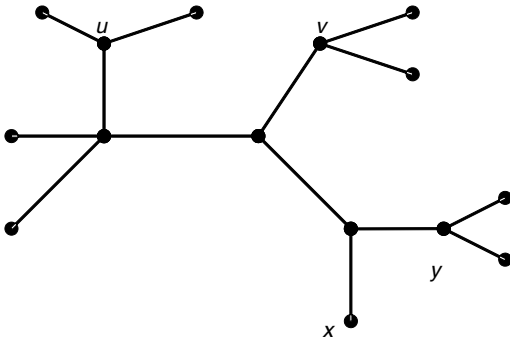


Figure 3 Graphical display of Table by a graph 2

Figure 4 A tree-like visualization T

A theorem of Buneman and Dobson states that, for a given matrix D , there exists such a weighted tree if and only if the four-points inequality (11) is fulfilled (see Figure 4). Then D is called an *additive tree* or *quadripolar distance*. If D is not a tree distance, we may find an optimum tree approximation to D by looking for a tree distance (δ_{ij}) with minimum deviation from D in analogy to (15) and interpret the corresponding tree T in terms of our application. For this and more general approaches see, e.g., [1], [8], [15].

Probabilistic Aspects

Dissimilarities from Random Data

When data were obtained from a random experiment, all dissimilarities calculated from these data must be

considered as random variables with some probability distribution. In some cases, this distribution is known and can be used for finding dissimilarity thresholds, e.g., for distinguishing between ‘similar’ and ‘dissimilar’ objects or clusters. We cite three examples:

1. If $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ are independent samples of p -dimensional random vectors X, Y with normal distributions $\mathcal{N}_p(\mu_1, \Sigma)$ and $\mathcal{N}_p(\mu_2, \Sigma)$, respectively, then the squared Mahalanobis distance $d_M^2(\bar{X}, \bar{Y}) := \|\bar{X} - \bar{Y}\|_{\Sigma^{-1}}^2$ of the class centroids $\bar{X}, \bar{Y}$ is distributed as $c^2 \chi_{p, \delta^2}^2$ where χ_{p, δ^2}^2 is a noncentral χ^2 with p degrees of freedom and noncentrality parameter $\delta^2 := \|\mu_1 - \mu_2\|_{\Sigma^{-1}}^2 / c^2$ with the factor $c^2 := n_1^{-1} + n_2^{-1}$.
2. On the other hand, if $X_1, \dots, X_n$ are independent $\mathcal{N}_p(\mu, \Sigma)$ variables with estimated covariance matrix $\hat{\Sigma} := (1/n) \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$, then the rescaled Mahalanobis distance $d_M^2(X_i, X_j) := (X_i - X_j)' \hat{\Sigma}^{-1} (X_i - X_j) / (2n)$ has a *Beta_I* distribution with parameters $p/2$ and $(n - p - 1)/2$ with expectation $2pn/(n - 1)$ ([2]).
3. Similarly, in the case of two binary data vectors $U = (U_1, \dots, U_p)'$ and $V = (V_1, \dots, V_p)'$ with $2p$ independent components, all with values 0 and 1 and the same hitting probability $\pi = P(U_i = 1) = P(V_i = 1)$ for $i = 1, \dots, p$, the Hamming distance $d_H(U, V)$ has the binomial distribution $\text{Bin}(p, 2\pi(1 - \pi))$ with expectation $2p\pi(1 - \pi)$.

Dissimilarity Between Probability Distributions

Empirical as well as theoretical investigations lead often to the problem to evaluate the (dis-)similarity of two populations that are characterized by two distributions P and Q , respectively, for a random data vector X . Typically, this comparison is conducted by using a variant of the ϕ -divergence measure of Csiszár [6]:

$$D(P; Q) := \int_{\mathcal{X}} q(x) \phi\left(\frac{p(x)}{q(x)}\right) dx \quad \text{resp.}$$

$$D(P; Q) := \sum_{x \in \mathcal{X}} q(x) \phi\left(\frac{p(x)}{q(x)}\right) \quad (16)$$

where $p(x), q(x)$ are the distribution densities (probability functions) of P and Q in the continuous (discrete) distribution case. Here $\phi(\lambda)$ is a convex function of a likelihood ratio $\lambda > 0$ (typically with $\phi(1) = 0$) and can be chosen suitably. Special cases include:

- (a) the *Kullback–Leibler discriminating information distance* for $\phi(\lambda) := -\log \lambda$:

$$D_{KL}(P; Q) := \int_{\mathcal{X}} q(x) \log \left(\frac{q(x)}{p(x)} \right) dx \geq 0 \quad (17)$$

with its symmetrized variant for $\tilde{\phi}(\lambda) := (\lambda - 1) \log \lambda$:

$$\begin{aligned} \tilde{D}(P; Q)_{KL} &:= D_{KL}(Q; P) + D_{KL}(P; Q) \\ &= \int_{\mathcal{X}} p(x) \tilde{\phi} \left(\frac{p(x)}{q(x)} \right) dx \end{aligned} \quad (18)$$

- (b) the *variation distance* resulting for $\phi(\lambda) := |\lambda - 1|$:

$$D_{var}(P; Q) := \int_{\mathcal{X}} |q(x) - p(x)| dx \quad (19)$$

- (c) the χ^2 -distance resulting from $\phi(\lambda) := (\lambda - 1)^2$:

$$D_{chi}(P; Q) := \int_{\mathcal{X}} \frac{(p(x) - q(x))^2}{q(x)} dx. \quad (20)$$

Other special cases include Hellinger and Bhattacharyya distance, Matusita distance, and so on (see [3, 4]). Note that only (18) and (19) yield a symmetric dissimilarity measure.

References

- [1] Barthélemy, J.-P. & Guénoche, A. (1988). *Les Arbres et Les Représentations Des Proximités*, Masson, Paris. English translation: *Trees and proximity representations*. Wiley, New York, 1991.
- [2] Bock, H.-H. (1974). *Automatische Klassifikation*, Göttingen, Vandenhoeck & Ruprecht.
- [3] Bock, H.-H. (1992). A clustering technique for maximizing ϕ -divergence, noncentrality and discriminating power, in *Analyzing and Modeling Data and Knowledge*, M. Schader, ed., Springer, Heidelberg, pp. 19–36.
- [4] Bock, H.-H. (2000). Dissimilarity measures for probability distributions, in *Analysis of Symbolic Data*, H.-H. Bock & E. Diday, eds, Springer, Heidelberg, pp. 153–160.
- [5] Brito, P. (2000). Hierarchical and pyramidal clustering with complete symbolic objects, in *Analysis of Symbolic Data*, H.-H. Bock & E. Diday, eds, Springer, Heidelberg, pp. 312–341.
- [6] Csiszár, L. (1967). Information-type measures of difference of probability distributions and indirect observations, *Studia Scientiarum Mathematicarum Hungarica* **2**, 299–318.
- [7] de Leeuw, J. & Heiser, W. (1982). Theory of multidimensional scaling, in *Classification, Pattern Recognition, and Reduction of Dimensionality*, P.R. Krishnaiah & L.N. Kanal, eds, North Holland, Amsterdam, pp. 285–316.
- [8] De Soete, G. & Carroll, J.D. (1996). Tree and other network models for representing proximity data, in *Clustering and Classification*, P. Arabie, L.J. Hubert & G. De Soete, eds, World Scientific, Singapore, pp. 157–197.
- [9] Esposito, F., Malerba, D. & Tamma, V. (2000). Dissimilarity measures for symbolic data, in *Analysis of Symbolic Data*, H.-H. Bock & E. Diday, eds, Springer, Heidelberg, pp. 165–185.
- [10] Everitt, B.S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, Oxford University Press.
- [11] Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data. Kendall's Library of Statistics, Vol. 4*, Arnold Publishers, London.
- [12] Gower, J.C. (1971). A general coefficient of similarity and some of its properties, *Biometrics* **27**, 857–874.
- [13] Lasch, R. (1993). *Pyramidale Darstellung Multivariater Daten*, Verlag Josef Eul, Bergisch Gladbach – Köln.
- [14] Sankoff, D., Kruskal, J. & Nerbonne, J. (1999). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, CSLI Publications, Stanford.
- [15] Van Cutsem, B., ed. (1994). *Classification and Dissimilarity Analysis, Lecture Notes in Statistics no. 93*, Springer, New York.

HANS-HERMANN BOCK

Psychophysical Scaling

HANS IRTEL

Volume 3, pp. 1628–1632

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Psychophysical Scaling

Introduction

A large part of human cognition is devoted to the development of a mental representation of the physical environment. This is necessary for planned interaction and for successful anticipation of dangerous situations. Survival in a continuously changing physical world will only be possible if the organism's mental representation of its environment is sufficiently valid. This actually is the case for most of our perceptual abilities. Our senses convey a rather valid view of the physical world, at least as long as we restrict the world to that part of the environment that allows for unaided interaction. And this is the basic idea of psychophysical scaling: Formulate a theory that allows the computation of perceived stimulus properties from purely physical attributes.

A problem complex as this requires simplification and reduction in order to create proper experiments and models that can be used to describe, or even explain, the results of these experiments. Thus, most of the experimental examples and most of the theories we will discuss here will be small and simplified cases. Most cases will assume that we have only a single relevant physical independent variable, and a single and unidimensional dependent variable. Simplification like this is appropriate as long as we keep in mind the bigger aim of relating the properties of the physical world to behavior, or, better, to the mental representation of the world that guides behavior.

Subject Tasks

Before going into the theoretical foundations of psychophysical scaling it might be useful to look at the experimental conditions which give rise to problems of psychophysical scaling. So we first will look at subject tasks which may be found in experiments involving psychophysical scaling. The most basic tasks involved in psychophysical experiments are detection and discrimination tasks. A *detection* task involves one or more time intervals during which a stimulus may be presented. The subject's task is to tell which if any time interval contained the target stimulus. Single time interval tasks usually are called *yes/no*-tasks while multiple time interval tasks

are called *forced choice*-tasks. *Discrimination* tasks are similar to multiple interval detection tasks. The major difference being that there are no empty intervals but the distractor intervals contain a reference or standard stimulus and the subject's task is to tell, whether the target is different from the reference or which of multiple intervals contains the target. The data of detection and discrimination experiments usually are captured by the probability that the respective target is detected or discriminated from its reference.

When scaling is involved, then discrimination frequently involves ordering. In this case, the subject is asked whether the target stimulus has more of some attribute as a second target that may be presented simultaneously or subsequently. These tasks are called *paired comparison* (see **Bradley-Terry Model**) tasks, and frequently involve the comparison of stimulus pairs. An example may be a task where the difference between two stimuli x , y with respect to some attribute has to be compared with the difference between two stimuli u , v . The data of these comparisons may also be handled as probabilities for finding a given ordering between pairs.

Another type of psychophysical task involves the assignment of numeric labels to stimuli. A simple case is the assignment of single stimuli to a small number of numeric categories. Or subjects may be required to directly assign real number labels to stimuli or pairs of stimuli such that the numbers describe some attribute intensity associated with physical stimulus properties. The most well-known task of this type is *magnitude estimation*: here the subject may be asked to assign a real number label to a stimulus such that the real number describes the appearance of some stimulus attribute such as loudness, brightness, or heaviness. These tasks usually involve stimuli that show large differences with respect to the independent physical attribute such that discrimination would be certain if two of them were presented simultaneously. In many cases, the numerical labels created by the subjects are treated as numbers such that the data will be mean values of subject responses.

The previously described task type may be modified such that the subject does not produce a number label but creates a stimulus attribute such that its appearance satisfies a certain numeric relation to a given reference. An example is *midpoint production*: here the subject adjusts a stimulus attribute such that it has an intensity that appears to be the midpoint between two given reference stimuli. Methods of this

2 Psychophysical Scaling

type are called *production methods* since the subject produces the respective stimulus attribute. This is in contrast to the *estimation methods*, where the subjects produce a numerical estimation of the respective attribute intensity.

Discrimination Scaling

Discrimination scaling is based on an assumption that dates back to **Fechner** [3]. His idea was that psychological measurement should be based on discriminability of stimuli. Luce & Galanter [5] used the phrase ‘equally often noticed differences are equal, unless always or never noticed’ to describe what they called *Fechner’s Problem*:

Suppose $P(x, y)$ is the discriminability of stimuli x and y . Does there exist a transformation g of the physical stimulus intensities x and y , such that

$$P(x, y) = F[g(x) - g(y)], \quad (1)$$

where F is a strictly increasing function of its argument? Usually, $P(x, y)$ will be the probability that stimulus x is judged to be of higher intensity as stimulus y with respect to some attributes. A solution g to (1) can be considered a psychophysical scale of the respective attribute in the sense that equal differences along the scale g indicate equal discriminability of the respective stimuli. An alternative but empirically equivalent formulation of (1) is

$$P(x, y) = G\left[\frac{h(x)}{h(y)}\right], \quad (2)$$

where $h(x) = e^{g(x)}$ and $G(x) = F(\log x)$.

Response probabilities $P(x, y)$ that have a representation like (1) have to satisfy the *quadruple condition*: $P(x, y) \geq P(u, v)$ if and only if $P(x, u) \geq P(y, v)$. This condition, however, is not easy to test since no statistical methods exist that allow for appropriate decisions based on estimates of P .

Response probabilities that have a Fechnerian representation in the sense of (1) allow the definition of a *sensitivity function* ξ : Let $P(x, y) = \pi$ and define $\xi_\pi(y) = x$. Thus, $\xi_\pi(y)$ is that stimulus intensity which, when compared to y , results in response probability π . From (1) with $h = F^{-1}$, we get

$$\xi_\pi(y) = g^{-1}[g(y) + h(\pi)], \quad (3)$$

where $h(\pi)$ is independent of x .

Fechner’s idea was that a fixed response probability value π corresponds to a single unit change on the sensation scale g : We take $h(\pi) = 1$ and look at the so called *Weber function* δ , which is such that $\xi_\pi(x) = x + \delta_\pi(x)$. The value of the Weber function $\delta_\pi(x)$ is that stimulus increment that has to be added to stimulus x such that the response probability $P(x + \delta_\pi(x), x)$ is π . *Weber’s law* states that $\delta_\pi(x)$ is proportional to x [16]. In terms of response probabilities, this means that $P(cx, cy) = P(x, y)$ for any multiplicative factor c . Weber’s law in terms of the sensitivity function ξ means that $\xi_\pi(cx) = c\xi_\pi(x)$. A generalization is $\xi_\pi(cx) = c^\beta \xi_\pi(x)$, which has been termed the *near-miss-to-Weber’s-law* [2]. Its empirical equivalent is $P(c^\beta x, cy) = P(x, y)$ with the representation $P(x, y) = G(y/x^{1/\beta})$. The corresponding Fechnerian representation then has the form

$$P(x, y) = F\left[\frac{1}{\beta} \log x - \log y\right]. \quad (4)$$

In case of $\beta = 1$, we get *Fechner’s law* according to which the sensation scale grows with the logarithmic transformation of stimulus intensity.

Operations on the Stimulus Set

Discrimination scaling is based on stimulus confusion. Metric information is derived from data that somehow describe a subject’s uncertainty when discriminating two physically distinct stimuli. There is no guarantee that the concatenation of just noticeable differences leads to a scale that also describes judgments about stimulus similarity for stimuli that are never confused. This has been criticized by [14].

An alternative method for scale construction is to create an operation on the set of stimuli such that this operation provides metric information about the appearance of stimulus differences. A simple case is the midpoint operation: the subject’s task is to find that stimulus $m(x, y)$, whose intensity appears to be the midpoint between the two stimuli x and y . For any sensation scale g , this should mean that

$$g[m(x, y)] = \frac{g(x) + g(y)}{2}. \quad (5)$$

The major empirical condition that guarantees that the midpoint operation m may be represented in this way is *bisymmetry* [11]:

$$m[m(x, y), m(u, v)] = m[m(x, u), m(y, v)], \quad (6)$$

which can be tested empirically. If bisymmetry holds, then a representation of the form

$$g[m(x, y)] = pg(x) + qg(y) + r \quad (7)$$

is possible. If, furthermore, $m(x, x) = x$ holds, then $p + q = 1$ and $r = 0$, and if $m(x, y) = m(y, x)$, then $p = q = 1/2$. Note that bisymmetry alone does not impose any restriction on the form of the psychophysical function g . However, Krantz [4] has shown that an additional empirically testable condition restricts the possible forms of the psychophysical function strongly. If the sensation scale satisfies (5), and the midpoint operation satisfies the homogeneity condition $m(cx, cy) = cm(x, y)$, then there remain only two possible forms of the psychophysical function g : $g(x) = \alpha \log x + \beta$, or $g(x) = \alpha x^\beta + \gamma$, for two constants $\alpha > 0$ and β . Falmagne [2] makes clear that the representation of m by an arithmetic mean is arbitrary. A geometric mean would be equally plausible, and the empirical conditions given above do not allow any distinction between these two options. Choosing the geometric mean as a representation for the midpoint operation m , however, changes the possible forms of the psychophysical function g [2].

The midpoint operation aims at the single numerical weight of $1/2$ for multiplying sensation scale values. More general cases have been studied by [9] in the context of magnitude estimation, which will be treated later. Methods similar to the midpoint operation have also been suggested by [14] under the label *ratio* or *magnitude production* with *fractionation* and *multiplication* as subcases. Pfanzagl [11], however, notes that these methods impose much less empirical constraints on the data such that almost any monotone psychophysical function will be admissible.

Magnitude Estimation, Magnitude Production, and Cross-modal Matching

Magnitude estimation is one of the classical methods proposed by [14] in order to create psychophysical scales that satisfy proper measurement conditions. Magnitude estimation requires the subject to assign numeric labels to stimuli such that the respective numbers are proportional to the magnitude of perceived stimulus intensity. Often, there will be a reference stimulus that is assigned a number label, ‘10’, say, by the experimenter, and the subject is instructed to map the relative sensation of the target with respect

to the reference. It is common practice to take the subjects’ number labels as proper numbers and compute average values from different subjects or from multiple replications with the same subject. This procedure remains questionable as long as no structural conditions are tested that validate the subjects’ proper handling of numeric labels. In magnitude production experiments, the subject is not required to produce number labels but to produce a stimulus intensity that satisfies a given relation on the sensation continuum to a reference, such as being 10 times as loud.

A major result of **Stevens’** research tradition is that average data from magnitude estimation and production frequently are well described by power functions: $g(x) = \alpha x^\beta$. Validation of the power law is done by fitting the respective power function to a set of data, and by *cross-modal matching*. This requires the subject to match a pair of stimuli from one continuum to a second pair of stimuli from another continuum. An example is the matching of a loudness interval defined by a pair of acoustic stimuli to the brightness interval of a pair of light stimuli [15]. If magnitude estimates of each single sensation scale are available and follow the power law, then the exponent of the matching function from one to the other continuum can be predicted by the ratio of the exponents. Empirical evidence for this condition is not unambiguous [2, 7]. In addition, a power law relation between matching sensation scales for different continua is also predicted by logarithmic sensation scales for the single continua [5].

The power law for sensation scales satisfies Stevens’ credo ‘that equal stimulus ratios produce equal subjective ratios’ ([14], p 153). It may, in fact, be shown that this rule implies a power law for the sensation scale g . But, as shown by several authors [2, 5], the notion ‘equal subjective ratios’ has the same theoretical status as Fechner’s assumption that just noticeable stimulus differences correspond to a single unit difference on the sensation scale. Both assumptions are arbitrary as long as there is no independent foundation of the sensation scale.

A theoretical foundation of magnitude estimation and cross-modal matching has been developed by [4] based on ideas of [12]. He combined magnitude estimates, ratio estimates, and cross-modal matches, and formulated a set of empirically testable conditions that have to hold if the psychophysical functions are power functions of the corresponding physical attribute. A key assumption of the Shepard–Krantz

theory is to map all sensory attributes to the single sensory continuum of perceived length, which is assumed to behave like physical length. The main assumption, then, is that for the reference continuum of perceived length, the condition $L(cx, cy) = L(x, y)$ holds. Since, furthermore, $L(x, y)$ is assumed to behave like numerical ratios, this form of invariance generates the power law [2].

A more recent theory of magnitude estimation for ratios has been developed by [9]. The gist of this theory is a strict separation of number labels, as they are used by subjects, and mathematical numbers used in the theory. Narens derived two major predictions: The first is a commutativity property. Let 'p' and 'q' be number labels, and suppose the subject produces stimulus y when instructed to produce p-times x , and then produces z when instructed to produce q-times y . The requirement is that the result is the same when the sequence of 'p' and 'q' is reversed. The second prediction is a multiplicative one. It requires that the subject has to produce the same stimulus as in the previous sequence when required to produce pq-times x . Empirical predictions like this are rarely tested. Exceptions are [1] or for a slightly different but similar approach [18]. In both cases, the model was only partially supported by the data.

While [6] describes the foundational aspects of psychophysical measurement, [8] presents a thorough overview over the current practices of psychophysical scaling in Stevens' tradition. A modern characterization of Stevens' measurement theory [13] is given by [10].

For many applications, the power law and the classical psychophysical scaling methods provide a good starting point for the question how stimulus intensity transforms into sensation magnitude. Even models that try to predict sensation magnitudes in rather complex conditions may incorporate the power law as a basic component for constant context conditions. An example is the CIE 1976 ($L^*a^*b^*$) color space [17]. It models color similarity judgments and implements both an adaptation- and an illumination-dependent component. Its basic transform from stimulus to sensation space, however, is a power law.

References

- [1] Ellermeier, W. & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens' ratio-scaling approach: I. loudness production, *Perception & Psychophysics* **62**, 1505–1511.
- [2] Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*, Oxford University Press, New York.
- [3] Fechner, G.Th. (1860). *Elemente der Psychophysik*, Breitkopf und Härtel, Leipzig.
- [4] Krantz, D.H. (1972). A theory of magnitude estimation and cross-modality matching, *Journal of Mathematical Psychology* **9**, 168–199.
- [5] Luce, R.D. & Galanter, E. (1963). Discrimination, in *Handbook of Mathematical Psychology*, Vol. I, R.D. Luce, R.R. Bush & E. Galanter, eds, Wiley, New York, pp. 191–243.
- [6] Luce, R.D. & Suppes, P. (2002). Representational measurement theory, in *Stevens' Handbook of Experimental Psychology*, Vol. 4, 3rd Edition, H. Pashler, ed., Wiley, New York, pp. 1–41.
- [7] Marks, L.E. & Algorn, D. (1998). Psychophysical scaling, in *Measurement, Judgement, and Decision Making*, M.H. Birnbaum, ed., Academic Press, San Diego, pp. 81–178.
- [8] Marks, L.E. & Gescheider, G.A. (2002). Psychophysical scaling, in *Stevens' Handbook of Experimental Psychology*, H. Pashler, ed., Vol. 4, 3rd Edition, Wiley, New York, pp. 91–138.
- [9] Narens, L. (1996). A theory of ratio magnitude estimation, *Journal of Mathematical Psychology* **40**, 109–129.
- [10] Narens, L. (2002). The irony of measurement by subjective estimations, *Journal of Mathematical Psychology* **46**, 769–788.
- [11] Pfanzagl, J. (1971). *Theory of Measurement*, 2nd Edition, Physica-Verlag, Würzburg.
- [12] Shepard, R.N. (1981). Psychophysical relations and psychophysical scales: on the status of 'direct' psychophysical measurement, *Journal of Mathematical Psychology* **24**, 21–57.
- [13] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.
- [14] Stevens, S.S. (1957). On the psychophysical law, *The Psychological Review* **64**, 153–181.
- [15] Stevens, J.C. & Marks L.E. (1965). Cross-modality matching of brightness and loudness, *Proceedings of the National Academy of Sciences*, **54**, 407–411.
- [16] Weber, E.H. (1834). *De Pulsu, Resorptione, Auditu et Tactu. Annotationes Anatomicae et Physiologicae*, Köhler, Leipzig.
- [17] Wyszecki, G. & Stiles, W.S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd Edition, Wiley, New York.
- [18] Zimmer, K., Luce, R.D. & Ellermeier, W. (2001). Testing a new theory of psychophysical scaling: temporal loudness integration, in *Fechner Day 2001. Proceedings of the Seventeenth Annual Meeting of the International Society for Psychophysics*, E. Sommerfeld, R. Kompuss & T. Lachmann, eds, Pabst, Lengerich, pp. 683–688.

(See also **Harmonic Mean**)

HANS IRTEL

Qualitative Research

MICHAEL QUINN PATTON

Volume 3, pp. 1633–1636

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Qualitative Research

An anthropologist studies initiation rites among the Gourma people of Burkina Faso in West Africa. A sociologist observes interactions among bowlers in their weekly league games. An evaluator participates fully in a leadership training program she is documenting. A naturalist studies Bighorn sheep beneath Powell Plateau in the Grand Canyon. A policy analyst interviews people living in public housing in their homes. An agronomist observes farmers' Spring planting practices in rural Minnesota. What do these researchers have in common? They are in the field, studying the real world as it unfolds. This is called *naturalistic inquiry*, and it is the cornerstone of qualitative research. Such qualitative investigations typically begin with detailed narrative descriptions, then constructing in-depth case studies of the phenomenon under study, and, finally, moving to comparisons and the interpretive search for patterns that cut across cases.

Qualitative research with human beings involves three kinds of data collection: (a) in-depth, open-ended interviews; (b) direct observations; and (c) written documents. *Interviews* yield direct quotations from people about their experiences, opinions, feelings, and knowledge. The data from *observations* consist of detailed descriptions of people's activities, behaviors, actions, and the full range of interpersonal interactions and organizational processes that are part of observable human experience. *Document analysis* includes studying excerpts, quotations or entire passages from organizational, clinical or program records; memoranda and correspondence; official publications and reports; personal diaries; and open-ended written responses to questionnaires and surveys.

The data for qualitative research typically come from fieldwork. During fieldwork, the researcher spends time in the setting under study – a program, organization, or community, where change efforts can be observed, people interviewed, and documents analyzed. The researcher makes firsthand observations of activities and interactions, sometimes engaging personally in those activities as a 'participant observer'. For example, a researcher might participate in all or part of an educational program under study, participating as a student. The qualitative researcher

talks with people about their experiences and perceptions. More formal individual or group interviews may be conducted. Relevant records and documents are examined. Extensive field notes are collected through these observations, interviews, and document reviews. The voluminous raw data in these field notes are organized into readable narrative descriptions with major themes, categories, and illustrative case examples extracted *inductively* through content analysis. The themes, patterns, understandings, and insights that emerge from research fieldwork and subsequent analysis are the fruit of qualitative inquiry.

Qualitative findings may be presented alone or in combination with quantitative data. At the simplest level, a questionnaire or interview that asks both fixed-choice (closed) questions and open-ended questions is an example of how quantitative measurement and qualitative inquiry are often combined.

The quality of qualitative data depends to a great extent on the methodological skill, sensitivity, and integrity of the researcher. Systematic and rigorous observation involves far more than just being present and looking around. Skillful interviewing involves much more than just asking questions. Content analysis requires considerably more than just reading to see what is there. Generating useful and credible qualitative findings through observation, interviewing, and content analysis requires discipline, knowledge, training, practice, creativity, and hard work.

Examples of Qualitative Applications

Fieldwork is the fundamental method of cultural anthropology. Ethnographic inquiry takes as its central and guiding assumption that any human group of people interacting together for a period of time will evolve a culture. Ethnographers study and describe specific cultures, then compare and contrast cultures to understand how cultures evolve and change. Anthropologists have traditionally studied nonliterate cultures in remote settings, what were often thought of as 'primitive' or 'exotic' cultures. Sociologists subsequently adapted anthropological fieldwork approaches to study urban communities and neighborhoods.

Qualitative research has also been important to management studies, which often rely on case studies of companies. One of the most influential books in organizational development has been *In Search*

of Excellence: Lessons from Americas' Best-Run Companies. Peters and Waterman [4] based the book on case studies of highly regarded companies. They visited companies, conducted extensive interviews, and studied corporate documents. From that massive amount of data, they extracted eight qualitative attributes of excellence: (a) a bias for action; (b) close to the customer; (c) autonomy and entrepreneurship; (d) productivity through people; (e) hands-on, value-driven; (f) stick to the knitting; (g) simple form, lean staff; and (h) simultaneous loose-tight properties. Their book devotes a chapter to each theme with case examples and implications. Their research helped launch the quality movement that has now moved from the business world to not-for-profit organizations and government.

A different kind of qualitative finding is illustrated by Angela Browne's book *When Battered Women Kill* [1]. Browne conducted in-depth interviews with 42 women from 15 states who were charged with a crime in the death or serious injury of their mates. She was often the first to hear these women's stories. She used one couple's history and vignettes from nine others, representative of the entire sample, to illuminate the progression of an abusive relationship from romantic courtship to the onset of abuse through its escalation until it was ongoing and eventually provoked a homicide. Her work helped lead to legal recognition of *battered women's syndrome* as a legitimate defense, especially in offering insight into the common outsider's question: Why does not the woman just leave? Getting an insider perspective on the debilitating, destructive, and all-encompassing brutality of battering reveals that question for what it is: the facile judgment of one who has not been there. The effectiveness of Browne's careful, detailed, and straightforward descriptions and quotations lies in their capacity to take us inside the abusive relationship. Offering that inside perspective powers qualitative reporting.

Qualitative methods are often used in program evaluations because they tell the program's story by capturing and communicating the participants' stories. Research case studies have all the elements of a good story. They tell what happened when, to whom, and with what consequences. The purpose of such studies is to gather information and generate findings that are *useful*. Understanding the program's and participant's stories is useful to the extent that those stories illuminate the processes and outcomes

of the program for those who must make decisions about the program. The methodological implication of this criterion is that the intended users must value the findings and find them credible. They must be interested in the stories, experiences, and perceptions of program participants beyond simply knowing how many came into the program, how many completed it, and how many did what afterwards. Qualitative findings in evaluation can illuminate the people behind the numbers and put faces on the statistics to deepen understanding [3].

Purposeful Sampling

Perhaps, nothing better captures the difference between quantitative and qualitative methods than the different logics that undergird sampling approaches (see **Survey Sampling Procedures**). Qualitative inquiry typically focuses in-depth on relatively small samples, even single cases ($n = 1$), selected *purposefully*. Quantitative methods typically depend on larger samples selected randomly. Not only are the techniques for sampling different, but the very logic of each approach is unique because the purpose of each strategy is different. The logic and power of random sampling derives from statistical probability theory. In contrast, the logic and power of purposeful sampling lies in selecting *information-rich cases* for study in depth. Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the inquiry, thus the term *purposeful* sampling. What would be 'bias' in statistical sampling, and, therefore, a weakness, becomes the intended focus in qualitative sampling, and, therefore, a strength. Studying information-rich cases yields insights and in-depth understanding rather than empirical generalizations. For example, if the purpose of a program evaluation is to increase the effectiveness of a program in reaching lower-socioeconomic groups, one may learn a great deal more by studying in-depth a small number of carefully selected poor families than by gathering standardized information from a large, statistically representative sample of the whole program. Purposeful sampling focuses on selecting information-rich cases whose study will illuminate the questions under study. There are several different strategies for purposefully selecting information-rich cases. The logic of each strategy serves a particular purpose.

Extreme or deviant case sampling involves selecting cases that are information-rich because they are unusual or special in some way, such as outstanding successes or notable failures. The influential study of high performing American companies published as *In Search of Excellence* [4] exemplifies the logic of purposeful, extreme group sampling. The sample of 62 companies was never intended to be representative of the US industry as a whole, but, rather, was purposefully selected to focus on innovation and excellence. In the early days of AIDS research, when HIV infections almost always resulted in death, a small number of cases of people infected with HIV who did not develop AIDS became crucial outlier cases that provided important insights into directions researchers should take in combating AIDS.

In program evaluation, the logic of extreme case sampling is that lessons may be learned about unusual conditions or extreme outcomes that are relevant to improving more typical programs. Suppose that we are interested in studying a national program with hundreds of local sites. We know that many programs are operating reasonably well, that other programs verge on being disasters, and that most programs are doing 'okay'. We know this from knowledgeable sources who have made site visits to enough programs to have a basic idea about what the variation is. If one wanted to precisely document the natural variation among programs, a random sample would be appropriate, one of sufficient size to be representative, and permit generalizations to the total population of programs. However, with limited resources and time, and with the priority being how to improve programs, an evaluator might learn more by intensively studying one or more examples of really poor programs and one or more examples of really excellent programs. The evaluation focus, then, becomes a question of understanding under what conditions programs get into trouble, and under what conditions programs exemplify excellence. It is not

even necessary to randomly sample poor programs or excellent programs. The researchers and intended users involved in the study think through *what cases they could learn the most from*, and those are the cases that are selected for study.

Examples of other purposeful sampling strategies are briefly described below:

- *Maximum variation sampling* involves purposefully picking a wide range of cases to get variation on dimensions of interest. Such a sample can document variations that have emerged in adapting to different conditions as well as identify important common patterns that cut across variations (cut through the noise of variation).
- *Homogenous sampling* is used to bring focus to a sample, reduce variation, simplify analysis, and facilitate group interviewing (focus groups).
- *Typical case sampling* is used to illustrate or highlight what is typical, normal, average, and gives greater depth of understanding to the qualitative meaning of a statistical mean.

For a full discussion of these and other purposeful sampling strategies, and the full range of qualitative methods and analytical approaches, see [2, 3].

References

- [1] Browne, A. (1987). *When Battered Women Kill*, Free Press, New York.
- [2] Denzin, N. & Lincoln, Y., eds (2000). *Handbook of Qualitative Research*, 2nd Edition, Sage Publications, Thousand Oaks.
- [3] Patton, M.Q. (2002). *Qualitative Research and Evaluation Methods*, 3rd Edition, Sage Publications, Thousand Oaks.
- [4] Peters, T. & Waterman, R. (1982). *In Search of Excellence*, Harper & Row, New York.

MICHAEL QUINN PATTON

Quantiles

PAT LOVIE

Volume 3, pp. 1636–1637

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quantiles

A quantile of a distribution is the value such that a proportion p of the population values are less than or equal to it. In other words, the p th quantile is the value that cuts off a certain proportion p of the area under the probability distribution function. It is sometimes known as a *theoretical quantile* or *population quantile* and is denoted by $Q_i(p)$. Quantiles for common distributions are easily obtained using routines for the inverse cumulative distribution function (inverse c.d.f.) found in many popular statistical packages.

For example, the 0.5 quantile (which we also recognize as the median or 50th **percentile**) of a standard normal distribution is $P(x \leq x_{0.5}) = \Phi^{-1}(0.5) = 0$, where $\Phi^{-1}(\cdot)$ denotes the inverse c.d.f. of that distribution. Equally easily, we locate the 0.5 quantile for an exponential distribution (see **Catalogue of Probability Density Functions**) with mean of 1 as 0.6931. Other useful quantiles such as the lower and upper quartiles (0.25 and 0.75 quantiles) can be obtained in a similar way.

However, what we are more likely to be concerned with are quantiles associated with data. Here, the *empirical quantile* $Q(p)$ is that point on the data scale that splits the data (the empirical distribution) into two parts so that a proportion p of the observations fall below it and the rest lie above. Although each data point is itself some quantile, obtaining a *specific* quantile requires a more pragmatic approach.

Consider the following 10 observations, which are the times (in milliseconds) between reported reversals of orientation of a Necker cube obtained from a study on visual illusions. The data have been ordered from the smallest to the largest.

274, 302, 334, 430, 489, 703, 978, 1656, 1697, 2745.

Imagine that we are interested in the 0.5 and 0.85 quantiles for these data. With 10 data points, the point $Q(0.5)$ on the data scale that has half of the observations below it and half above does not actually correspond to a member of the sample but lies *somewhere* between the middle two values 489 and 703. Also, we have no way of splitting off exactly 0.85 of the data. So, where do we go from here?

Several different remedies have been proposed, each of which essentially entails a modest redefinition of quantile. For most practical purposes, any

differences between the estimates obtained from the various versions will be negligible. The only method that we shall consider here is the one that is also used by Minitab and SPSS (but see [1] for a description of an alternative favored by the **Exploratory Data Analysis** (EDA) community).

Suppose that we have n ordered observations x_i ($i = 1$ to n). These split the data scale into $n + 1$ segments: one below the smallest observation, $n - 1$ between each adjacent pair of values and one above the largest. The proportion p of the distribution that lies below the i th observation is then estimated by $i/(n + 1)$. Setting this equal to p gives $i = p(n + 1)$. If i is an integer, the i th observation is taken to be $Q(p)$. Otherwise, we take the integer part of i , say j , and look for $Q(p)$ between the j th and $j + 1$ th observations. Assuming simple linear interpolation, $Q(p)$ lies a fraction $(i - j)$ of the way from x_j to x_{j+1} and is estimated by

$$Q(p) = x_j + (x_{j+1} - x_j) \times (i - j). \quad (1)$$

A point to note is that we cannot determine quantiles in the tails of the distribution for $p < 1/(n + 1)$ or $p > n/(n + 1)$ since these take us outside the range of the data. If extrapolation is unavoidable, it is safest to define $Q(p)$ for all P values in these two regions as x_1 and x_n , respectively.

We return now to our reversal data and our quest for the 0.5 and 0.85 quantiles:

For the 0.5 quantile, $i = 0.5 \times 11 = 5.5$, which is not an integer so $Q(0.5) = 489 + (703 - 489) \times (5.5 - 5) = 596$, which is exactly half way between the fifth and sixth observations. For the 0.85 quantile, $i = 0.85 \times 11 = 9.35$, so the required value will lie 0.35 of the way between the ninth and tenth observations, and is estimated by $Q(0.85) = 1697 + (2745 - 1697) \times (9.35 - 9) = 2063.8$.

From a plot of the empirical quantiles $Q(p)$, that is, the data, against proportion p as in Figure 1, we can get some feeling for the distributional properties of our sample. For example, we can read off rough values of important quantiles such as the median and quartiles. Moreover, the increasing steepness of the slope on the right-hand side indicates a lower density of data points in that region and thus alerts us to skewness and possible outliers. Note that we have also defined $Q(p) = x_1$ when $p < 1/11$ and $Q(p) = x_{10}$ when $p > 10/11$. Clearly, extrapolation would not be advisable, especially at the upper end.

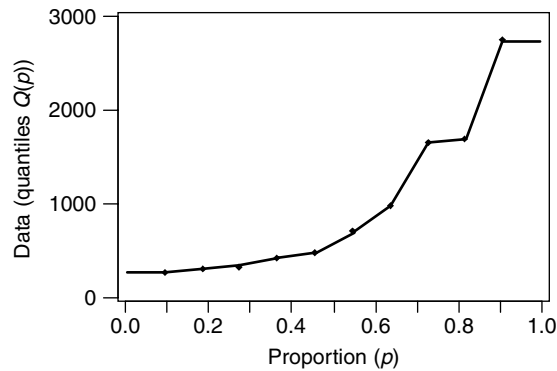


Figure 1 Quantile plot for Necker cube reversal times

It is possible to obtain **confidence intervals** for quantiles, and information on how these are constructed can be found, for example, in [2]. Routines for these procedures are available within certain statistical packages such as SAS and S-PLUS.

Quantiles play a central role within exploratory data analysis. The median ($Q(0.5)$), upper and lower quartiles ($Q(0.75)$ and $Q(0.25)$), and the maximum and minimum values, which constitute the so-called five number summary, are the basic elements of a box plot. An **empirical quantile–quantile** plot (EQQ) is a **scatterplot** of the paired empirical quantiles of

two samples of data, while the **symmetry plot** is a scatterplot of the paired empirical quantiles from a single sample, which has been split and folded at the median. Of course, for these latter plots there is no need for the P values to be explicitly identified.

Finally, it is worth noting that the term quantile is a general one referring to the class of statistics, which split a distribution according to a proportion p that can be *any* value between 0 and 1. As we have seen, well-known measures that have specific P values, such as the median, quartile, and percentile (on Galton’s original idea of splitting by 100ths [3]), belong to this group. Nowadays, however, quantile and percentile are used almost interchangeably: the only difference is whether p is expressed as a decimal fraction or as a percentage.

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.
- [2] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [3] Galton, F. (1885). Anthropometric percentiles, *Nature* **31**, 223–225.

PAT LOVIE

Quantitative Methods in Personality Research

R. CHRIS FRALEY AND MICHAEL J. MARKS

Volume 3, pp. 1637–1641

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quantitative Methods in Personality Research

The aims of personality psychology are to identify the ways in which people differ from one another and to elucidate the psychological mechanisms that generate and sustain those differences. The study of personality is truly multidisciplinary in that it draws upon data and insights from disciplines as diverse as sociology, social psychology, cognitive science, development, physiology, genetics, clinical psychology, and evolutionary biology. As with other subdisciplines within psychology, quantitative methods play a crucial role in personality theory and research. In this entry, we identify quantitative methods that have facilitated discovery, assessment, and model testing in personality science.

Quantitative Methods as a Tool for Discovery

There are hundreds, if not thousands, of ways that people differ from one another in their thoughts, motives, behaviors, and emotional experiences. Some people are highly creative and prolific, such as Robert Frost, who penned 11 books of poetic works and J. S. Bach who produced over 1000 masterpieces during his lifetime. Other people have never produced a great piece of musical or literary art and seem incapable of understanding the difference between a sonnet and sonata. One of the fundamental goals of personality psychology is to map the diverse ways in which people differ from one another. A guiding theme in this taxonomic work is that the vast number of ways in which people differ can be understood using a smaller number of organizing variables. This theme is rooted in the empirical observation that people who tend to be happier in their personal relationships, for example, also take more risks and tend to experience higher levels of positive affect. The fact that such characteristics – qualities that differ in their surface features – tend to covary positively in empirical samples suggests that there is a common factor or trait underlying their covariation.

Quantitative methods such as **factor analysis** have been used to decompose the covariation among

behavioral tendencies or person descriptors. Cattell [3], Tupes and Christal [14], and other investigators demonstrated that the covariation (*see Covariance*) among a variety of behavioral tendencies can be understood as deriving from a smaller number of latent factors (*see Latent Variable*). In contemporary personality research, most investigators focus on five factors derived from factor analytic considerations (extraversion, agreeableness, conscientiousness, neuroticism, openness to experience). These ‘Big Five’ are considered by many researchers to represent the fundamental trait dimensions of personality [8, 16].

Although factor analysis has primarily been used to decompose the structure of variable $\times$ variable correlation matrices (i.e., matrices of correlations between variables, computed for N people), factor analytic methods have also been used to understand the structure of person $\times$ person correlation matrices (i.e., matrices of correlations between people, computed for k variables). In *Q-factor analysis*, a person $\times$ person correlation matrix is decomposed so that the factors represent latent profiles or ‘prototypes,’ and the loadings represent the influence of the latent prototypes on each person’s profile of responses across variables [17]. Q-factor analysis is used to develop taxonomies of people, as opposed to taxonomies of variables [11]. Two individuals are similar to one another if they share a similar profile of trait scores or, more importantly, if they have high loadings on the same latent prototype. Although regular factor analysis and Q-factor analysis are often viewed as providing different representations of personality structure [11], debates about the extent to which these different approaches yield divergent and convergent sources of information on personality structure have never been fully resolved [2, 13].

Quantitative Methods in Personality Assessment and Measurement

One of the most important uses of mathematics and statistics in personality research is for measurement. Personality researchers rely primarily upon classical test theory (CTT) (*see Classical Test Models*) as a general psychometric framework for psychological measurement. Although CTT has served the field well, researchers are slowly moving toward modern test theory approaches such as **item response theory (IRT)**. IRT is a model-based approach that relates

variation in a latent trait to the probability of an item or behavioral response. IRT holds promise for personality research because it assesses measurement precision differentially along different levels of a trait continuum, separates item characteristics from person characteristics, and assesses the extent that a person's item response pattern deviates from that assumed by the measurement model [5]. This latter feature of IRT is important for debates regarding the problem of 'traitedness' – the degree to which a given trait domain is (or is not) relevant for characterizing a person's behavior. By adopting IRT procedures, it is possible to scale individuals along a latent trait continuum while simultaneously assessing the extent to which the assumed measurement model is appropriate for the individual in question [10].

Another development in quantitative methodology that is poised to influence personality research is the development of taxometric procedures – methods used to determine whether a latent variable is categorical or continuous. Academic personality researchers tend to conceptualize variables as continua, whereas clinical psychologists tend to treat personality variables as categories. Taxometric procedures developed by Meehl and his colleagues are designed to test whether measurements behave in a categorical or continuous fashion [15]. These procedures have been applied to the study of a variety of clinical syndromes such as depression and to nonclinical variables such as sexual orientation [7].

Quantitative Methods in Testing Alternative Models of Personality Processes

Given that most personality variables cannot be experimentally manipulated, many personality researchers adopt **quasi-experimental** and longitudinal approaches (see **Longitudinal Data Analysis**) to study personality processes. The most common way of modeling such data is to use **multiple linear regression**. Multiple regression is a statistical procedure for modeling the influence of two or more (possibly correlated) variables on an outcome. It is widely used in personality research because of its flexibility (e.g., its ability to handle both categorical and continuous predictor variables and its ability to model multiplicative terms).

One of the most important developments in the use of regression for studying personality processes was

Baron and Kenny's (1986) conceptualization of **moderation** and **mediation** [1]. A variable *moderates* the relationship between two other variables when it statistically interacts with one of them to influence the other. For example, personality researchers discovered that the influence of adverse life events on depressive symptoms is moderated by cognitive vulnerability such that people who tend to make negative attributions about their experiences are more likely to develop depressive symptoms following a negative life event (e.g., failing an exam) than people who do not make such attributions [6]. Hypotheses about moderation are tested by evaluating the interaction term in a multiple regression analysis. A variable *mediates* the association between two other variables when it provides a causal pathway through which the impact of one variable is transmitted to another. Mediational processes are tested by examining whether or not the estimated effect of one variable on another is diminished when the conjectured mediator is included in the regression equation. For example, Sandstorm and Cramer [12] demonstrated that the moderate association between social status (e.g., the extent one is preferred by one's peers) and the use of psychological defense mechanisms after an interpersonal rejection is substantially reduced when changes in stress are statistically controlled. This suggests that social status has its effect on psychological defenses via the amount of stress that a rejected person experiences. In sum, the use of simple regression techniques to examine moderation and mediation enables researchers to test alternative models of personality processes.

During the past 20 years, an increasing number of personality psychologists began using **structural equation modeling (SEM)** to formalize and test causal models of personality processes. SEM has been useful in personality research for at least two reasons. First, the process of developing a quantitative model of psychological processes requires researchers to state their assumptions clearly. Moreover, once those assumptions are formalized, it is possible to derive quantitative predications that can be empirically tested. Second, SEM provides researchers with an improved, if imperfect, way to separate the measurement model (i.e., the hypothesis about how latent variables are manifested via behavior or self-report) from the causal processes of interest (i.e., the causal influences among the latent variables).

One of the most widely used applications of SEM in personality is in behavior genetic research with samples of twins (*see Twin Designs*). Structural equations specify the causal relationships among genetic sources of variation, phenotypic variation, and both shared and nonshared nongenetic sources of variation. By specifying models and estimating parameters with behavioral genetic data, researchers made progress in testing alternative models of the causes of individual differences [9]. Structural equations are also used in longitudinal research (*see Longitudinal Data Analysis*) to model and test alternative hypotheses about the way that personality variables influence specific outcomes (e.g., job satisfaction) over time [4].

Hierarchical linear modeling (HLM) (*see Hierarchical Models*) is used to model personality data that can be analyzed across multiple levels (e.g., within persons or within groups of persons). For example, in ‘diary’ research, researchers may assess peoples’ moods multiple times over several weeks. Data gathered in this fashion are hierarchical because the daily observations are nested within individuals. As such, it is possible to study the factors that influence variation in mood within a person, as well as the factors that influence mood between people. In HLM, the within-person parameters and the between-person parameters are estimated simultaneously, thereby providing an efficient way to model complex psychological processes (*see Linear Multilevel Models*).

The Future of Quantitative Methods in Personality Research

As with many areas of behavioral research, the statistical methods used by personality researchers tend to lag behind the quantitative state of the art. To demonstrate this point, we constructed a snapshot of the quantitative methods in contemporary personality research by reviewing 259 articles from the 2000 to 2002 issues of the *Journal of Personality* and the *Personality Processes and Individual Differences* section of the *Journal of Personality and Social Psychology*. Table 1 identifies the frequency of statistical methods used. As can be seen, although newer and potentially valuable methods such as SEM, HLM, and IRT are used in personality research, they are greatly overshadowed by towers of the quantitative past such as ANOVA. It is our hope that future researchers will

Table 1 Quantitative methods used in contemporary personality research

Technique	Frequency
Correlation (zero-order)	175
ANOVA	90
<i>t</i> Test	87
Multiple regression	72
Factor analysis	47
Structural equation modeling/path analysis	41
Mediation/moderation	20
Chi-square	19
ANCOVA	16
Hierarchical linear modeling and related techniques	14
MANOVA	13
Profile similarity and Q-sorts	9
Growth curve analysis	3
Multidimensional scaling	3
Item response theory	3
Taxometrics	2

Note: Frequencies refer to the number of articles that used a specific quantitative method. Some of the 259 articles used more than one method.

explore the benefits of newer quantitative methods for understanding the nature of personality.

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] Burt, C. (1937). Methods of factor analysis with and without successive approximation, *British Journal of Educational Psychology* **7**, 172–195.
- [3] Cattell, R.B. (1945). The description of personality: principles and findings in a factor analysis, *American Journal of Psychology* **58**, 69–90.
- [4] Fraley, R.C. (2002). Attachment stability from infancy to adulthood: meta-analysis and dynamic modeling of developmental mechanisms, *Personality and Social Psychology Review* **6**, 123–151.
- [5] Fraley, R.C., Waller, N.G. & Brennan, K.G. (2000). An item response theory analysis of self-report measures of adult attachment, *Journal of Personality and Social Psychology* **78**, 350–365.
- [6] Hankin, B.L. & Abramson, L.Y. (2001). Development of gender differences in depression: an elaborated cognitive vulnerability-transactional stress theory, *Psychological Bulletin* **127**, 773–796.

4 Quantitative Methods in Personality Research

- [7] Haslam, N. & Kim, H.C. (2002). Categories and continua: a review of taxometric research, *Genetic, Social, and General Psychology Monographs* **128**, 271–320.
- [8] John, O.P. & Srivastava, S. (1999). The big five trait taxonomy: history, measurement, and theoretical perspectives, in *Handbook of Personality: Theory and Research*, 2nd Edition, L.A. Pervin & O.P. John, eds, Guilford Press, New York, pp. 102–138.
- [9] Loehlin, J.C., Neiderhiser, J.M. & Reiss, D. (2003). The behavior genetics of personality and the NEAD study, *Journal of Research in Personality* **37**, 373–387.
- [10] Reise, S.P. & Waller, N.G. (1993). Traitendness and the assessment of response pattern scalability, *Journal of Personality and Social Psychology* **65**, 143–151.
- [11] Robins, R.W., John, O.P. & Caspi, A. (1998). The typological approach to studying personality, in *Methods and Models for Studying the Individual*, R.B. Cairns, L. Bergman & J. Kagan, eds, Sage Publications, Thousand Oaks, pp. 135–160.
- [12] Sandstrom, M.J. & Cramer, P. (2003). Girls' use of defense mechanisms following peer rejection, *Journal of Personality* **71**, 605–627.
- [13] Stephenson, W. (1952). Some observations on Q technique, *Psychological Bulletin* **49**, 483–498.
- [14] Tupes, E.C. & Christal, R.C. (1961). Recurrent personality factors based on trait ratings, Technical Report No. ASD-TR-61-97, U.S. Air Force, Lackland Air Force Base.
- [15] Waller, N.G. & Meehl, P.E. (1998). *Multivariate Taxometric Procedures: Distinguishing types from Continua*, Sage Publications, Newbury Park.
- [16] Wiggins, J.S., ed. (1996). *The Five-Factor Model of Personality: Theoretical Perspectives*. Guilford Press, New York.
- [17] York, K.L. & John, O.P. (1992). The four faces of eve: a typological analysis of women's personality at midlife, *Journal of Personality and Social Psychology* **63**, 494–508.

R. CHRIS FRALEY AND MICHAEL J. MARKS

Quartiles

DAVID CLARK-CARTER

Volume 3, pp. 1641–1641

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quartiles

The first quartile (often shown as Q_1) is the number below which lie a quarter of the scores in the set, and the third quartile (Q_3) is the number that has three-quarters of the numbers in the set below it. The second quartile (Q_2) is the median. The first quartile is the same as the 25th **percentile** and the third quartile is the same as the 75th percentile.

As with medians, the ease of calculating quartiles depends on whether the sample can be straightforwardly divided into quarters. The equation for finding the rank of any given quartile is:

$$Q_x = \frac{x \times (n + 1)}{4}, \quad (1)$$

where x is the quartile we are calculating and n is the sample size (*see* **Quantiles** for a general formula).

As an example, if there are 11 numbers in the set and we want to know the rank of the first quartile, then it is $1 \times (11 + 1)/4 = 3$. We can then find the third number in the ordered set and it will lie on the first quartile; the third quartile will have the rank 9.

On the other hand, if the above equation does not produce an exact rank, that is, when $n + 1$ is not divisible by 4, then the value of the quartile will be found between the two values whose ranks lie on either side of the quartile that is sought. For example, in the set 2, 4, 6, 9, 10, 11, 14, 16, 17, there are 9

numbers. $(n + 1)/4 = 2.5$, which does not give us a rank in the set. Therefore, the first quartile lies between the second and third numbers in the set, that is, between 4 and 6. Taking the mean of the two numbers produces a first quartile of 5.

Interpolation can be used to find the first and third quartiles when data are grouped into frequencies for given ranges, for example age ranges, in a similar way to that used to find the median in such circumstances.

Quartiles have long been used in graphical displays of data, for example, by **Galton** in 1875 [1]. More recently, their use has been advocated by those arguing for **exploratory data analysis (EDA)** [2], and they are important elements of **box plots**. Quartiles are also the basis of measures of spread such as the **interquartile range** (or midspread) and the semi-interquartile range (or quartile deviation).

References

- [1] Galton, F. (1875). Statistics by intercomparison, with remarks on the law of frequency of error, *Philosophical Magazine* 4th series, **49**, 33–46. Cited in Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Belknap Press, Cambridge.
- [2] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

DAVID CLARK-CARTER

Quasi-experimental Designs

WILLIAM R. SHADISH AND JASON K. LUELLEN

Volume 3, pp. 1641–1644

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quasi-experimental Designs

Campbell and Stanley [1] described a theory of quasi-experiments that was revised and expanded in Cook and Campbell [2] and Shadish, Cook, and Campbell [4]. Quasi-experiments, like all experiments, test hypotheses about the effects of manipulable treatments. However, quasi-experiments lack the process of random assignment of study units¹ to conditions. Rather, some nonrandom assignment mechanism is used. For example, units might be allowed to choose treatment for themselves or treatment might be assigned on the basis of need. Many other nonrandom mechanisms are possible (*see* **Randomization**).

As background, there are three basic requirements for establishing a causal relationship between a treatment and an effect. First, the treatment must precede the effect. This temporal sequence is met when the experimenter or someone else first introduces the treatment, with the outcome observed afterward. Second, the treatment must be related to the effect. This is usually measured by one of many statistical procedures. Third, aside from treatment, no plausible alternative explanations for the effect must exist. This third requirement is often facilitated by random assignment, which ensures that alternative explanations are no more likely in the treatment than in the control group. Lacking random assignment in quasi-experiments, it is harder to rule out these alternative explanations.

Alternative explanations for the observed effect are called *threats* to **internal validity** [1]. They include:

1. *Ambiguous temporal precedence*: Lack of clarity about which variable occurred first may yield confusion about which variable is cause and which is effect.
2. *Selection*: Systematic differences over conditions in respondent characteristics that could also cause the observed effect.
3. *History*: Events occurring concurrently with treatment that could cause the observed effect.
4. *Maturation*: Naturally occurring changes over time that could mimic a treatment effect.
5. *Regression*: When units are selected for their extreme scores, they will often have less extreme

scores on other measures, which can mimic a treatment effect.

6. *Attrition*: Loss of respondents to measurement can produce artifactual effects if reasons for attrition vary by conditions.
7. *Testing*: Exposure to a test can affect scores on subsequent exposures to that test, which could mimic a treatment effect.
8. *Instrumentation*: The nature of a measure may change over time or conditions in a way that could mimic a treatment effect.
9. *Additive and interactive effects of threats to internal validity*: The impact of a threat can be added to, or may depend on the level of, another threat.

When designing quasi-experiments, researchers should attempt to identify which of these threats to causal inference is plausible and reduce their plausibility with design elements. Some common designs are presented shortly. Note, however, that simply choosing one of these designs does not guarantee the validity of inferences. Further, the choice of design might improve internal validity but simultaneously hamper external, construct, or statistical conclusion validity (*see* **External Validity**) [4].

Some Basic Quasi-experimental Designs

In a common notation system used to present these designs, treatments are represented by the symbol X , observations or measurements by O , and assignment mechanisms by NR for nonrandom assignment or C for cutoff-based assignment. The symbols are ordered from left to right following the temporal sequence of events. Subscripts denote variations in measures or treatments. The descriptions of these designs are brief and represent but a few of the designs presented in [4].

The One-group Posttest Only Design

This simple design involves one posttest observation on respondents (O_1) who experienced a treatment (X). We diagram it as:

$$X \quad O_1$$

This is a very weak design. Aside from ruling out ambiguous temporal precedence, nearly all other threats to internal validity are usually plausible.

The Nonequivalent Control Group Design With Pretest and Posttest

This common quasi-experiment adds a pretest and control group to the preceding design:

$$\begin{array}{ccccccc} NR & O_1 & X & O_2 & & & \\ \hline NR & O_1 & & O_2 & & & \end{array}$$

The additional design elements prove very useful. The pretest helps determine whether people changed from before to after treatment. The control group provides an estimate of the counterfactual (i.e., what would have happened to the same participants in the absence of treatment). The joint use of a pretest and a control group serves several purposes. First, one can measure the size of the treatment effect as the difference between the treatment and control outcomes. Second, by comparing pretest observations, we can explore potential selection bias. This design allows for various statistical adjustments for pretest group differences when found. Selection is the most plausible threat to this design, but other threats (e.g., history) often apply too (*see Nonequivalent Group Design*).

Interrupted Time Series Quasi-experiments

The **interrupted time-series design** can be powerful for assessing treatment impact. It uses repeated observations of the same variable over time. The researcher looks for an interruption in the slope of the time series at the point when treatment was implemented. A basic time-series design with one treatment group and 10 observations might be diagrammed as:

$$O_1 \ O_2 \ O_3 \ O_4 \ O_5 \ X \ O_6 \ O_7 \ O_8 \ O_9 \ O_{10}$$

However, it is preferable to have 100 observations or more, allowing use of certain statistical procedures to obtain better estimates of the treatment effect. History is the most likely threat to internal validity with this design.

The Regression Discontinuity Design

For the **regression discontinuity design**, treatment assignment is based on a cutoff score on an assignment variable measured prior to treatment. A simple

two-group version of this design where those scoring on one side of the cutoff are assigned to treatment and those scoring on the other side are assigned to control is diagrammed as:

$$\begin{array}{cccc} O_A & C & X & O \\ O_A & C & & O \end{array}$$

where the subscript *A* denotes a pretreatment measure of the assignment variable. Common assignment variables are measures of need or merit.

The logic of this design can be illustrated by fitting a regression line to a scatterplot of the relationship between assignment variable scores (horizontal axis) and outcome scores (vertical axis). When a treatment has no effect, the regression line should be continuous. However, when a treatment has an effect, the regression line will show a discontinuity at the cutoff score corresponding to the size of the treatment effect. Differential attrition poses a threat to internal validity with this design; and failure to model the shape of the regression line properly can result in biased results.

Conclusion

Quasi-experiments are simply a collection of design elements aimed at reducing threats to causal inference. Reynolds and West [3] combined design elements in an exemplary way. They assessed the effects of a campaign to sell lottery tickets using numerous design elements, including untreated matched controls, multiple pretests and posttests, nonequivalent dependent variables, and removed and repeated treatments. Space does not permit us to discuss a number of other design elements and the ways in which they reduce threats to validity (but see [4]). Finally, recent developments in the analysis of quasi-experiments have been made including **propensity score** analysis, hidden bias analysis, and selection bias modeling (see [4], Appendix 5.1).

Note

1. Units are often persons but may be animals, plots of land, schools, time periods, cities, classrooms, or teachers, among others.

References

- [1] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.

- [2] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton-Mifflin, Boston.
- [3] Reynolds, K.D. & West, S.G. (1987). A multiplist strategy for strengthening nonequivalent control group designs, *Evaluation Review* **11**, 691–714.
- [4] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.

WILLIAM R. SHADISH AND JASON K. LUELLEN

Quasi-independence

SCOTT L. HERSHBERGER

Volume 3, pp. 1644–1647

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quasi-independence

Independence Model

One of the most basic questions that can be answered by a **contingency table** is whether the row and column variables are *independent*; that is, whether a statistical association is present between the variables [5]. For example, assume that two variables, X and Y (see Table 1), each with two outcomes, are measured on a sample of respondents.

In this 2×2 table, n_{11} is the number of observations with levels $X = 1$ and $Y = 1$, n_{12} is the number of observations with levels $X = 1$ and $Y = 2$, n_{21} is the number of observations with levels $X = 2$ and $Y = 1$, n_{22} is the number of observations with levels $X = 2$ and $Y = 2$, and a $\cdot$ in a subscript indicates summation over the variable that corresponds to that position.

The cells of this 2×2 table can also be expressed as probabilities (see Table 2).

If X and Y are independent, then it is true that

$$p(X = j \text{ and } Y = k) = p(X = j) \times p(Y = k) \quad (1)$$

for all i and j . A test of the independence of X and Y would the following null and alternative hypotheses:

$$H_0 : p_{jk} = p_{j\cdot} p_{\cdot k} \quad (2)$$

Table 1 Observed frequencies in a 2×2 contingency table

		Y		Total
		1	2	
X	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Table 2 Observed probabilities in a 2×2 contingency table

		Y		Total
		1	2	
X	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot}$

versus

$$H_1 : p_{jk} \neq p_{j\cdot} p_{\cdot k} \quad (3)$$

When the data are organized in a $J \times K$ contingency table, often a simple test, such as the *Pearson χ^2 test of association*, will suffice to accept or reject the null hypothesis of independence [1]. The χ^2 test of association examines the closeness of the observed cell frequencies (n) to the cell frequencies expected ($\hat{n}$) under the null hypothesis – the cell frequencies that would be observed if X and Y are independent. Under the null hypothesis of independence, the expected frequencies are calculated as

$$\hat{n}_{jk} = n_{\cdot\cdot} \hat{p}_{j\cdot} \hat{p}_{\cdot k} = \frac{n_{j\cdot} n_{\cdot k}}{n_{\cdot\cdot}} \quad (4)$$

The χ^2 test of association is defined by

$$\chi^2 = \sum_j \sum_k \frac{(\hat{n}_{jk} - n_{jk})^2}{\hat{n}_{jk}} \quad (5)$$

with degrees of freedom equal to $(J - 1) \times (K - 1)$.

As an example of a test of independence, consider the 5×2 table of gambling frequency and male sexual orientation adapted from Hershberger and Bogaert [3] (see Table 3).

The expected frequency of each cell is given in parentheses. For example, the expected frequency of male homosexuals who reported ‘little’ gambling is 39.56:

$$\hat{n}_{31} = \frac{n_{3\cdot} n_{\cdot 1}}{n_{\cdot\cdot}} = \frac{(345)(1051)}{9166} = 39.56. \quad (6)$$

The χ^2 test of association statistic is 59.24, $df = 4$, $p < .0001$, indicating that the null hypothesis of independence should be rejected: A relationship

Table 3 Gambling frequency and male sexual orientation

Gambling frequency	Sexual orientation		Total
	Homosexual	Heterosexual	
None	714 (679.82)	5215 (5249.20)	5929
Rare	227 (291.90)	2319 (2254.10)	2546
Little	33 (39.56)	312 (305.44)	345
Some	74 (38.30)	260 (295.70)	334
Much	3 (1.38)	9 (10.62)	12
Total	1051	8115	9166

2 Quasi-independence

does exist between male sexual orientation and gambling frequency.

As an alternative to the χ^2 test of association, we can test the independence of two variables using **loglinear modeling** [6]. The loglinear modeling approach is especially useful for examining the association among more than two variables in contingency tables of three or more dimensions. A loglinear model of independence may be developed from the formula provided earlier for calculating expected cell frequencies in a two dimensional table:

$$\hat{n}_{jk} = n_{..} p_{j.} p_{.k}. \quad (7)$$

Taking the natural logarithm of this product transforms it into a loglinear model of independence:

$$\log \hat{n}_{jk} = \log n_{..} + \log p_{j.} + \log p_{.k}. \quad (8)$$

Thus, $\log \hat{n}_{jk}$ depends additively on a term based on total sample size ($\log n_{..}$), a term based on the marginal probability of row j ($\log p_{j.}$), and a term based on the marginal probability of column k ($\log p_{.k}$). A more common representation of the loglinear model of independence for a two dimensional table is

$$\log \hat{n}_{jk} = \lambda + \lambda_j X + \lambda_k Y. \quad (9)$$

For a **2 × 2 table**, the loglinear model requires the estimation of three parameters: a ‘grand mean’ represented by λ , a ‘main effect’ for variable X ($\lambda_j X$) and a ‘main effect’ for variable Y ($\lambda_k Y$). The number of parameters estimated depends on the number of row and column levels of the variables. In general, if variable X has J levels, then $J - 1$ parameters are estimated for the X main effect, and if Y has K levels, then $K - 1$ parameters are estimated for the Y main effect. For example, for the 5×2 table shown above, six parameters are estimated:

$$\lambda, \lambda_1 X, \lambda_2 X, \lambda_3 X, \lambda_4 X, \lambda_1 Y.$$

In addition, in order to obtain unique estimates of the parameters, constraints must be placed on the parameter values. A common set of constraints is

$$\sum_j \lambda_j X = 0 \quad (10)$$

and

$$\sum_k \lambda_k Y = 0. \quad (11)$$

Table 4 Gambling frequency and male sexual orientation: natural logarithms of the expected frequencies

Gambling frequency	Sexual orientation	
	Homosexual	Heterosexual
None	6.52	8.57
Rare	5.68	7.72
Little	3.68	5.72
Some	3.65	5.69
Much	1.10	2.20

Note that the independence model for the 5×2 table, has four degrees of freedom, obtained from the difference between the number of cells (10) and the number of parameters estimated (6).

Returning to the 5×2 table of sexual orientation and gambling frequency, the natural logarithms of the expected frequencies are as shown in Table 4.

The parameter estimates are

$$\lambda = 4.99$$

$$\lambda_1 X = 2.55 \quad \lambda_2 X = 1.71 \quad \lambda_3 X = -.29 \quad \lambda_4 X = -.32$$

$$\lambda_1 Y = -1.02. \quad (12)$$

To illustrate how the parameter estimates are related to the expected frequencies, consider the expected frequency for cell (1, 1):

$$6.52 = \lambda + \lambda_1 X + \lambda_1 Y = 4.99 + 2.55 - 1.02. \quad (13)$$

The *likelihood ratio statistic* can be used to test the fit of the independence model:

$$G^2 = 2 \sum_j \sum_k n_{jk} \log \left(\frac{n_{jk}}{\hat{n}_{jk}} \right) = 52.97. \quad (14)$$

The likelihood ratio statistic is asymptotically chi-square distributed. Therefore, with $G^2 = 52.97$, $df = 4$, $p < .0001$, the independence model is rejected. Note that if we incorporate a term representing the association between X and Y , $\lambda_{jk} XY$, four additional degrees of freedom are used, thus leading to a *saturated model* [2] with $df = 0$ (which is not a directly testable model). However, from our knowledge that the model without $\lambda_{jk} XY$ does not fit the data, we know that the association between X and Y must be significant. Additionally, although the fit of the saturated model as whole is not

testable, Pearson χ^2 goodness-of fit statistics are computable for the main effect of X , frequency of gambling ($\chi^2 = 2843.17$, $df = 4$, $p < .0001$), implying significant differences in frequency of gambling; for the main effect of Y , sexual orientation ($\chi^2 = 158.37$, $df = 1$, $p < .0001$), implying that homosexual and heterosexual men differ in frequency; and most importantly, for the association between frequency of gambling and sexual orientation ($\chi^2 = 56.73$, $df = 4$, $p < .0001$), implying that homosexual and heterosexual men differ in how frequently they gamble.

Quasi-independence Model

Whether by design or accident, there may be no observations in one or more cells of a contingency table. We can distinguish between two situations in which **incomplete contingency tables** can be expected [4]:

1. *Structural zeros*. On the basis of our knowledge of the population, we do *not* expect one or more combinations of the factor levels to be observed in a sample. By design, we have one or more empty cells (*see Structural Zeros*).
2. *Sampling zeros*. Although in the population all possible combinations of factor levels occur, we do not observe one or more of these combinations in our sample. By accident, we have one or more empty cells.

While sampling zeros occur from deficient sample sizes, too many factors, or too many factor levels, structural zeros occur when it is theoretically impossible for a cell to have any observations. For example, let us assume we have two factors, sex (male or female) and breast cancer (yes or no). While it is medically possible to have observations in the cell representing males who have breast cancer (male $\times$ yes), the rareness of males who have breast cancer in the population may result in no such cases appearing in our sample. On the other hand, let us say we sample both sex and the frequency of different types of cancers. While the cell representing males who have prostate cancer will have observations, it is impossible to have any observations in the cell representing females who have prostate cancer. Sampling and structural zeros should not be analytically treated the same.

While sampling zeros should contribute to the estimation of the model parameters, structural zeros should not.

An independence model that is fit to a table with one or more structural zeros is called a *quasi-independence model*. The loglinear model of quasi-independence is

$$\log \hat{n}_{jk} = \lambda + \lambda_j X + \lambda_k Y + s_{jk} I, \quad (15)$$

where I is an indicator variable that equals 1 when a cell has a structural zero and equals 0 when the cell has a nonzero number of observations [1]. Note that structural zeros affect the degrees of freedom of the Pearson goodness-of-fit statistic and the likelihood ratio statistic: In $J \times K$ table, there are only $JK - s$ observations, where s refers to the number of cells with structural zeros.

As an example of a situation where structural zeros *and* sampling zeros could occur, consider a study that examined the frequency of different types of cancers among men and women (Table 5).

Note that our indicator variable s takes on the value of 1 for women with prostate cancer and for men with ovarian cancer (both cells are examples of structural zeros; that is, ‘impossible’ observations). However, s does take on the value of 0 even though there are no men with breast cancer because conceivably such men could have appeared in our sample (this cell is an example of a sampling zero; that is, although rare, it is a ‘possible’ observation). The likelihood ratio statistic G^2 for the quasi-independence model is 24.87, $df = 3$, $p < .0001$, leading to the rejection of the model. Thus, even taking into account the impossibility of certain cancers appearing in men or women,

Table 5 Frequency of the different types of cancer among men and women

Type of cancer	Sex		Total
	Male	Female	
Throat	25	20	45
Lung	40	35	75
Prostate	45	structural 0	45
Ovarian	structural 0	10	10
Breast	sampling 0	20	20
Stomach	10	7	17
Total	120	92	212

4 Quasi-independence

there is a relationship between type of cancer and sex.

The importance of treating sampling zeros from structural zeros different analytically is emphasized if we fit the quasi-independence model without distinguishing the difference between the two types of zeros. In this case, the likelihood ratio statistic is 19, $df = 2$, $p = .91$, leading us to conclude erroneously that no association exists between type of cancer and sex. Also note that by incorrectly treating the category of men with breast cancer as a structural zero, we have reduced the degrees of freedom from 3 to 2.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, New York.
- [2] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, 2nd Edition, Chapman & Hall, Boca Raton.
- [3] Hershberger, S.L. & Bogaert, A.F. (in press). Male and female sexual orientation differences in gambling, *Personality and Individual Differences*.
- [4] Koehler, K. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables, *Journal of the American Statistical Association* **81**, 483–493.
- [5] von Eye, A. (2002). *Configural Frequency Analysis*, Lawrence Erlbaum, Mahwah.
- [6] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum, Hillsdale.

SCOTT L. HERSHBERGER

Quasi-symmetry in Contingency Tables

SCOTT L. HERSHBERGER

Volume 3, pp. 1647–1650

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quasi-symmetry in Contingency Tables

Symmetry Model for Nominal Data

The *symmetry model* for a square $I \times I$ **contingency table** assumes that the expected cell probabilities (π) of cell i, j is equal to that of cell j, i . When there is symmetry, $\pi_{i.} = \sum_j \pi_{ij} = \sum_j \pi_{ji} = \pi_{.j}$, a condition which also implies *marginal homogeneity* (see **Marginal Independence**) [1]. Cell probabilities satisfy marginal homogeneity if $\pi_{i.} = \pi_{.j}$, in which the row and column probabilities for the same classification are equal. We can see the relationship between marginal homogeneity and symmetry by noting that the probabilities in a square table satisfy symmetry if $\pi_{ij} = \pi_{ji}$ for all pairs of cells. Thus, a square table is symmetric when the elements on opposite sides of the main diagonal are equal. In **2 × 2 tables**, tables that have symmetry must also have marginal homogeneity. However, when in larger tables, marginal homogeneity can occur without symmetry [2].

Example of Symmetry Model for Nominal Data

Written as a **log-linear model** for the expected cell frequencies (e), the symmetry model is

$$\log e_{ij} = \lambda + \lambda_i X + \lambda_j Y + \lambda_{ij} XY, \quad (1)$$

in which the constraints

$$\begin{aligned} \lambda_i X &= \lambda_i Y, \\ \lambda_{ij} XY &= \lambda_{ji} XY \end{aligned} \quad (2)$$

are imposed. The fit if the symmetry model is based on

$$e_{ij} = e_{ji} = \frac{n_{ij} + n_{ji}}{2}; \quad (3)$$

that is, the expected entry in cell i, j is the average of the entries on either side of the diagonal. The diagonal entries themselves are not relevant to the symmetry hypothesis and are ignored in fitting the model [9].

We use as an example for fitting the symmetry model data for which we wish to conduct **multidimensional scaling**. Multidimensional scaling models that are based on Euclidean distance assume that the distance from object a_i to object a_j is the same as that from a_j to object a_i [5]. If symmetry is confirmed, a multidimensional scaling model based on Euclidean distance can be applied to the data; if symmetry is not confirmed, then a multidimensional scaling model not based on Euclidean distance (e.g., city block metric) should be considered.

Consider the data in Table 1 from [8], in which subjects who did not know Morse code listened to pairs of 10 signals consisting of a series of dots and dashes and were required to state whether the two signals they heard were the same or different.

Each number in the table is the percentage of subjects who responded that the row and column signals were the same. In the experiment, the row signal always preceded the column signal.

In order to fit the symmetry model, a symmetry variable is created, which takes on the same value for cells (1, 2) and (2, 1), another value for cells (1, 3) and (3, 1), and so forth. Therefore, the symmetry model has $[I(I-1)/2] = [10(10-1)/2] = 45$ degrees of freedom for goodness-of-fit. The χ^2 goodness-of-fit statistic for the symmetry model is 192.735, which at 45 degrees of freedom, suggests that the symmetry model cannot be accepted. The log likelihood $G^2 = 9314.055$.

Bowker's test of symmetry can also be used to test the symmetry model [3]. For Bowker's test, the null hypothesis is that the probabilities in a square table satisfy symmetry or that for the cell probabilities, $p_{ij} = p_{ji}$ for all pairs of table cells. Bowker's test

Table 1 Morse code data

Digit	1	2	3	4	5	6	7	8	9	0
1 (-----)	84	63	13	8	10	8	19	32	57	55
2 (..----	62	89	54	20	5	14	20	21	16	11
3 (...---	18	64	86	31	23	41	16	17	8	10
4 (....--)	5	26	44	89	42	44	32	10	3	3
5 (.....)	14	10	30	69	90	42	24	10	6	5
6 (-....)	15	14	26	24	17	86	69	14	5	14
7 (---...)	22	29	18	15	12	61	85	70	20	13
8 (----.)	42	29	16	16	9	30	60	89	61	26
9 (-----)	57	39	9	12	4	11	42	56	91	78
0 (-----)	50	26	9	11	5	22	17	52	81	84

2 Quasi-symmetry in Contingency Tables

of symmetry is

$$B = \sum_{i < j} \sum_j \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}. \quad (4)$$

The test is asymptotically chi-square distributed for large samples, with $I(I - 1)/2$ degrees of freedom under the null hypothesis of symmetry. For these data, $B = 108.222$, $p < .0001$ at 45 degrees of freedom. Thus, Bowker's test also rejects the symmetry null hypothesis. When $I = 2$, Bowker's test is identical to McNemar's test [7], which is calculated as

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}. \quad (5)$$

Quasi-symmetry Model for Nominal Data

The symmetry model rarely fits the data well because of the imposition of marginal homogeneity. A generalization of symmetry is the quasi-symmetry model, which permits marginal heterogeneity [4]. Marginal heterogeneity is obtained by permitting the log-linear main effect terms to differ. Under symmetry, the log-linear model is

$$\log e_{ij} = \lambda + \lambda_i X + \lambda_j Y + \lambda_{ij} XY, \quad (6)$$

with the constraints

$$\begin{aligned} \lambda_i X &= \lambda_i Y, \\ \lambda_{ij} XY &= \lambda_{ji} XY. \end{aligned} \quad (7)$$

Now under quasi-symmetry, the constraint $\lambda_i X = \lambda_i Y$ is removed but the constraint $\lambda_{ij} XY = \lambda_{ji} XY$ is retained. Removing the constraint $\lambda_i X = \lambda_i Y$ permits marginal heterogeneity.

The quasi-symmetry model can be best understood by considering the **odds ratios** implied by the model. The quasi-symmetry model implies that the odds ratios on one side of the main diagonal are identical to corresponding ones on the other side of the diagonal [1]. That is, for $i \neq i'$ and $j \neq j'$,

$$\alpha = \frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}} = \frac{\pi_{ji}\pi_{j'i'}}{\pi_{ji'}\pi_{j'i}}. \quad (8)$$

Adjusting for differing marginal distributions, there is a symmetric association pattern in the table. Like the symmetry model, in order to fit the quasi-symmetry model, a variable is created which takes

on the same value for cells (1, 2) and (2, 1), another value for cells (1, 3) and (3, 1), and so forth. This variable takes up $I(I - 1)/2$ degrees of freedom. In addition, in order to allow for marginal heterogeneity, the constraint $\lambda_i X = \lambda_i Y$ is removed, which results in estimating $\lambda_i - 1$ parameters. Therefore, the quasi-symmetry model has

$$I^2 - \left[\frac{I(I - 1)}{2} + (I - 1) \right] = \frac{(I - 1)(I - 2)}{2} \quad (9)$$

degrees of freedom

for goodness-of-fit.

Example of Quasi-symmetry Model for Nominal Data

We return to the Morse code data presented earlier. As we did for the symmetry model, we specify a symmetry variable which takes on the same value for corresponding cells above and below the main diagonal. In addition, we incorporate the Morse code variable (the row and column factor) in order to allow for marginal heterogeneity. The model has $[(10 - 1)(10 - 2)/2] = 36$ degrees of freedom. The χ^2 goodness-of-fit statistic is 53.492, $p < .05$, which suggests that the quasi-symmetry model is rejected. The log likelihood $G^2 = 9382.907$.

We can test for marginal homogeneity directly by computing twice the difference in log likelihoods between the symmetry and quasi-symmetry models, a difference that is asymptotically χ^2 on $I - 1$ degrees of freedom. Twice the difference in log likelihoods is $2(9382.907 - 9314.055) = 137.704$, $df = 9$, suggesting that marginal homogeneity is not present.

Quasi-symmetry Model for Ordinal Data

A quasi-symmetry model for ordinal data (*see Scales of Measurement*) can be specified as

$$\log e_{ij} = \lambda + \lambda_i X + \lambda_j Y + \beta u_j + \lambda_{ij} XY, \quad (10)$$

where u_j denotes ordered scores for the row and column categories. As in the nominal quasi-symmetry model,

$$\lambda_{ij} XY = \lambda_{ji} XY, \quad (11)$$

but

$$\lambda_i X \neq \lambda_i Y \quad (12)$$

to allow for marginal heterogeneity. In the ordinal quasi-symmetry model,

$$\lambda_j Y - \lambda_j X = \beta u_j, \quad (13)$$

which means that the difference in the two main effects has a linear trend across the response categories, and that this difference is a constant multiple β [1]. Setting $\beta = 0$ produces the symmetry model. Thus, because an additional parameter β is estimated in the ordinal quasi-symmetry model, the ordinal quasi-symmetry model has one less degree of freedom than the symmetry model, or $df = [I(I - 1)/2] - 1$.

Example of Quasi-symmetry Model for Ordinal Data

The data in Table 2 are from [6], representing the movement in occupational status from father to son, where the status variable is ordered.

The ordinal quasi-symmetry model has $[7(7 - 1)/2] - 1 = 20$ df. The χ^2 goodness-of-fit statistic is 29.584, $p = .077$ ($G^2 = 14044.363$), suggesting that the null hypothesis of quasi-symmetry can be accepted. Removing the constant multiple β parameter from the model, results in a test of the symmetry model, with $\chi^2 = 38.007$, $df = 21$, $p < .05$ ($G^2 = 14040.007$). Although quasi-symmetry can be accepted for these data, symmetry should be rejected.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.

Table 2 Movement in occupational status from father to son

Father's status	Son's status						
	1	2	3	4	5	6	7
1	50	19	26	8	18	6	2
2	16	40	34	18	31	8	3
3	12	35	65	66	123	23	21
4	11	20	58	110	223	64	32
5	14	36	114	185	714	258	189
6	0	6	19	40	179	143	71
7	0	3	14	32	141	91	106

[2] Bhapkar, V.P. (1979). On tests of marginal symmetry and quasi-symmetry in two- and three-dimensional contingency tables, *Biometrics* **35**, 417–426.

[3] Bowker, A.H. (1948). Bowker's test for symmetry, *Journal of the American Statistical Association* **43**, 572–574.

[4] Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. [Contributions to the statistical analysis of correlation matrices], *Annales de la Faculté des sciences de l'Université de Toulouse pour les sciences mathématiques et les sciences physiques* **29**, 77–182.

[5] Cox, T.F. & Cox, M.A.A. (1994). *Multidimensional Scaling*, Chapman & Hall, London.

[6] Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories, *Journal of the American Statistical Association* **74**, 537–552.

[7] McNemar, Q. (1947). Note on the sampling error of the difference between two correlated proportions, *Psychometrika* **12**, 153–157.

[8] Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired associate learning tasks, *Journal of Experimental Psychology* **53**, 94–101.

[9] Wickens, T.D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*, Lawrence Erlbaum, Hillsdale.

SCOTT L. HERSHBERGER

Quetelet, Adolphe

DIANA FABER

Volume 3, pp. 1650–1651

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Quetelet, Adolphe

Born: February 22, 1796, in Ghent, Belgium.

Died: February 17, 1874, in Brussels.

Quételet was born nearly fifty years after **Pierre-Simon Laplace**, the French mathematician and statistician. Partly thanks to Laplace, the subject of mathematical statistics, and in particular the laws of probability, became increasingly recognized as useful in their application to social phenomena. Quételet came to statistics through the exact science of astronomy, and some of his energy and ambition was channeled into his efforts to encourage the foundation of observatories. In the Belgian uprising in 1830 against French control of the country, the Observatory was used as a defence of Brussels; life became difficult for scientists in the capital and some of them joined the military defence. The Belgian revolution led to the defeat of France and its control of the country, but the scientific reputation of France remained paramount and the cultural ties between the two countries survived, to some extent, intact.

In 1819, Quételet gained the first doctorate in science at the University of Ghent, on the subject of a theory of conical sections. He then taught mathematics at the Athaeneum in Brussels. In 1820, he was elected to the Académie des Sciences et Belles-Lettres de Bruxelles. In 1823, he travelled to Paris where he spent three months. Here he learned from Laplace, in an informal way, some observational astronomy and probability theory. He also learned the method of **least squares**, as a way of reducing astronomical observations. Soon after he founded the journal '*Correspondence Mathématique et Physique*' to which he contributed many articles between the years 1825 and 1839. In 1828, Quételet became the founder of the national Observatory and received the title of Royal astronomer and meteorologist. He acted as perpetual secretary of the Royal Academy from 1834 until his death in 1874. Quételet's avowed ambition was to become known as the 'Newton of statistics'. In the body of his

work, he left proposals for the application of statistics to human affairs on which his achievement could be judged.

The first published work of Quételet in 1826 was '*Astronomie élémentaire*'. The second was published as a memoir of the Académie des Sciences in the following year, and reflected his interest in the application of statistics to social phenomena – '*Recherches sur la population, les naissances, les décès, les prisons, etc*'. A work on probability followed in 1828, '*Instructions populaires sur le calcul des probabilités*'. In 1835, '*Sur l'homme et le développement de ses facultés, ou l'essai de physique sociale*' introduced the notion of 'social physics' by analogy with physical and mathematical science. He had already proposed the notion of '*mécanique sociale*' as a correlate to '*mécanique céleste*'. Both terms pointed to the application of scientific method of analyzing social data. In 1846, Quételet published '*Lettres*', which discussed the theory of probabilities as applied to moral and political sciences. Statistical laws, and how they are applied to society, was the subject of a work in 1848: '*Du système social et des lois qui les régissent*'. In 1853 appeared a second work on probability: '*Théorie des probabilités*'. Quételet likened periodic phenomena and regularities as found in nature to changes and periodicity in human societies. He saw in the forces of mankind and social upheavals an analogue to the perturbations in nature. The name of Quételet has become strongly associated with his concept of the average man ('*l'homme moyen*'), constructed from his statistics on human traits and actions. It has been widely used since, both as a statistical abstraction and in some cases as a guide, useful or otherwise for social judgments.

Further Reading

- Porter, T.M. (1986). *The Rise of Statistical Thinking 1820–1900*, Princeton University Press, Princeton.
- Stigler, S.M. (1986). *The History of Statistics: the Measurement of Uncertainty Before 1900*, The Belknap Press of Harvard University Press, Harvard.

DIANA FABER

Encyclopedia of Statistics in

BEHAVIORAL SCIENCE

Editors in Chief **Brian Everitt** **David Howell**

VOLUME

4

r-z

VOLUME 4

R & Q Analysis.	1653-1655	Regression Artifacts.	1723-1725
R Squared, Adjusted R Squared.	1655-1657	Regression Discontinuity Design.	1725-1727
Random Effects and Fixed Effects Fallacy.	1657-1661	Regression Model Coding for the Analysis of Variance.	1727-1729
Random Effects in Multivariate Linear Models: Prediction.	1661-1665	Regression Models.	1729-1731
Random Forests.	1665-1667	Regression to the Mean.	1731-1732
Random Walks.	1667-1669	Relative Risk.	1732-1733
Randomization.	1669-1674	Reliability: Definitions and Estimation.	1733-1735
Randomization Based Tests.	1674-1681	Repeated Measures Analysis of Variance.	1735-1741
Randomized Block Design: Nonparametric Analyses.	1681-1686	Replicability of Results.	1741-1743
Randomized Block Designs.	1686-1687	Reproduced Matrix.	1743-1744
Randomized Response Technique.	1687	Resentful Demoralization.	1744-1746
Range.	1687-1688	Residual Plot.	1746-1750
Rank Based Inference.	1688-1691	Residuals.	1750-1754
Rasch Modeling.	1691-1698	Residuals in Structural Equation, Factor Analysis, and Path Analysis Models.	1754-1756
Rasch Models for Ordered Response Categories.	1698-1707	Resistant Line Fit.	1756-1758
Rater Agreement.	1707-1712	Retrospective Studies.	1758-1759
Rater Agreement - Kappa.	1712-1714	Reversal Design.	1759-1760
Rater Agreement - Weighted Kappa.	1714-1715	Risk Perception.	1760-1764
Rater Bias Models.	1716-1717	Robust Statistics for Multivariate Methods.	1764-1768
Reactivity.	1717-1718	Robust Testing Procedures.	1768-1769
Receiver Operating Characteristics Curves.	1718-1721	Robustness of Standard Tests.	1769-1770
Recursive Models.	1722-1723		

Runs Test.	1771	Shepard Diagram.	1830
Sample Size and Power Calculation.	1773-1775	Sibling Interaction Effects.	1831-1832
Sampling Distributions.	1775-1779	Sign Test.	1832-1833
Sampling Issues in Categorical Data.	1779-1782	Signal Detection Theory.	1833-1836
Saturated Model.	1782-1784	Signed Ranks Test.	1837-1838
Savage, Leonard Jimmie.	1784-1785	Simple Random Assignment.	1838-1840
Scales of Measurement.	1785-1787	Simple Random Sampling.	1840-1841
Scaling Asymmetric Matrices.	1787-1790	Simple V Composite Tests.	1841-1842
Scaling of Preferential Choice.	1790-1794	Simulation Methods for Categorical Variables.	1843-1849
Scatterplot Matrices.	1794-1795	Simultaneous Confidence Interval.	1849-1850
Scatterplot Smoothers.	1796-1798	Single and Double-Blind Procedures.	1854-1855
Scatterplots.	1798-1799	Single-Case Designs.	1850-1854
Scheffe, Henry.	1799-1800	Skewness.	1855-1856
Second order Factor Analysis: Confirmatory.	1800-1803	Slicing Inverse Regression.	1856-1863
Selection Study (Mouse Genetics).	1803-1804	Snedecor, George Waddell.	1863-1864
Sensitivity Analysis.	1805-1809	Social Interaction Models.	1864-1865
Sensitivity Analysis in Observational Studies.	1809-1814	Social Networks.	1866-1871
Sequential Decision Making.	1815-1819	Social Psychology.	1871-1874
Sequential Testing.	1819-1820	Social Validity.	1875-1876
Setting Performance Standards - Issues, Methods.	1820-1823	Software for Behavioral Genetics.	1876-1880
Sex-Limitation Models.	1823-1826	Software for Statistical Analyses.	1880-1885
Shannon, Claude.	1827	Spearman, Charles Edward.	1886-1887
Shared Environment.	1828-1830	Spearman's Rho.	1887-1888

Sphericity Test.	1888-1890	Structural Equation Modeling:Nonstandard Cases.	1931-1934
Standard Deviation.	1891	Structural Zeros.	1958-1959
Standard Error.	1891-1892	Subjective Probability and Human Judgement.	1960-1967
Standardized Regression Coefficients.	1892	Summary Measure Analysis of Longitudinal Data.	1968-1969
Stanine Scores.	1893	Survey Questionnaire Design.	1969-1977
Star and Profile Plots.	1893-1894	Survey Sampling Procedures.	1977-1980
State Dependence.	1895	Survival Analysis.	1980-1987
Statistical Models.	1895-1897	Symmetry Plot.	1989-1990
Stem and Leaf Plot.	1897-1898	Symmetry: Distribution Free Tests for.	1987-1988
Stephenson, William.	1898-1900	Tau-Equivalent and Congeneric Measurements.	1991-1994
Stevens, S S.	1900-1902	Teaching Statistics to Psychologists.	1994-1997
Stratification.	1902-1905	Teaching Statistics: Sources.	1997-1999
Structural Equation Modeling and Test Validation.	1951-1958	Telephone Surveys.	1999-2001
Structural Equation Modeling: Categorical Variables.	1905-1910	Test Bias Detection.	2001-2007
Structural Equation Modeling: Checking Substantive Plausibility.	1910-1912	Test Construction.	2007-2011
Structural Equation Modeling: Mixture Models.	1922-1927	Test Construction: Automated.	2011-2014
Structural Equation Modeling: Multilevel.	1927-1931	Test Dimensionality: Assessment of.	2014-2021
Structural Equation Modeling: Nontraditional Alternatives.	1934-1941	Test Translation.	2021-2026
Structural Equation Modeling: Overview.	1941-1947	Tetrachoric Correlation.	2027-2028
Structural Equation Modeling: Software.	1947-1951	Theil Slope Estimate.	2028-2030
Structural Equation Modeling:Latent Growth Curve Analysis.	1912-1922	Thomson, Godfrey Hilton.	2030-2031
		Three dimensional (3D) Scatterplots.	2031-2032

Three-Mode Component and Scaling Methods.	2032-2044	Utility Theory.	2098-2101
Thurstone, Louis Leon.	2045-2046	Validity Theory and Applications.	2103-2107
Time Series Analysis.	2046-2055	Variable Selection.	2107-2110
Tolerance and Variance Inflation Factor.	2055-2056	Variance.	2110
Transformation.	2056-2058	Variance Components.	2110-2113
Tree Models.	2059-2060	Walsh Averages.	2115-2116
Trellis Graphics.	2060-2063	Wiener, Norbert.	2116-2117
Trend Tests for Counts and Proportions.	2063-2066	Wilcoxon, Frank.	2117-2118
Trimmed Means.	2066-2067	Wilcoxon-Mann-Whitney Test.	2118-2121
T-Scores.	2067	Winsorized Robust Measures.	2121-2122
Tukey Quick Test.	2069-2070	Within Case Designs: Distribution Free Methods.	2122-2126
Tukey, John Wilder.	2067-2069	Yates' Correction.	2127-2129
Tversky, Amos.	2070-2071	Yates, Frank.	2129-2130
Twin Designs.	2071-2074	Yule, George Udny.	2130-2131
Twins Reared Apart Design.	2074-2076	Z-Scores.	2131-2132
Two by Two Contingency Tables.	2076-2081		
Two-mode Clustering.	2081-2086		
Two-way Factorial: Distribution-Free Methods.	2086-2089		
Type I, Type II and Type III Sums of Squares.	2089-2092		
Ultrametric Inequality.	2093-2094		
Ultrametric Trees.	2094-2095		
Unidimensional Scaling.	2095-2097		
Urban, F M.	2097-2098		

R & Q Analysis

J. EDWARD JACKSON

Volume 4, pp. 1653–1655

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

R & Q Analysis

A number of criteria exist for performing **principal component analysis** on a data matrix. These criteria are referred to as *R*-, *Q*-, *N*-, *M*- and *P*-analysis. The first four of these criteria all involve deviations about column means, row means, or both. The properties of these criteria were given by Okamoto [7]. *P*-analysis involves the raw data directly.

In principal component analysis, one generally begins with an $n \times p$ data matrix X representing n observations on p variables. Some of the criteria will be illustrated by numerical examples, all using the following data matrix of $n = 4$ observations (rows) on $p = 3$ variables (columns):

$$X = \begin{array}{c} \begin{array}{ccc} & \text{(variables)} & \\ \begin{array}{l} \text{variable means} \\ 0 \quad 0 \quad 0 \end{array} & \begin{bmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{bmatrix} & \begin{array}{l} \\ \text{(observations)} \end{array} \end{array} \quad (1)$$

This matrix will have a rank of three. The variable means have been subtracted to simplify the computations. This example is taken from [6, Chapter 11], which includes more detail in the operations that are to follow.

R-analysis

In principal component analysis, one generally forms some type of $p \times p$ dispersion matrix of the variables from the $n \times p$ data matrix X , usually a covariance matrix (see **Correlation and Covariance Matrices**) or its related correlation matrix. A set of linear transformations, utilizing the **eigenvectors** of this matrix, is found which will transform the original correlated variables into a new set of variables. These new variables are uncorrelated and are called *principal components*. The values of the transformed data are called *principal component scores*. Further analysis may be carried out on these scores (see **Principal Components and Extensions**). (A subset of these transformed variables associated with the larger **eigenvalues** is often retained for this analysis.) This procedure is sometimes referred to as *R-analysis*,

and is the most common application of principal component analysis. Similar procedures may also be carried out in some **factor analysis** models.

For example, consider an *R*-analysis of matrix X . Rather than use either covariance or correlation matrices, which would require different divisors for the different examples, the examples will use sums of squares and cross-products matrices to keep the units the same. Then, for *R*-analysis, this matrix for the variables becomes:

$$X'X = \begin{bmatrix} 8 & 8 & 6 \\ 8 & 10 & 7 \\ 6 & 7 & 6 \end{bmatrix} \quad (2)$$

whose eigenvalues are $l_1 = 22.282$, $l_2 = 1.000$ and $l_3 = 0.718$. The fact that there are three positive eigenvalues indicates that $X'X$ has a rank of three. The unit (i.e., $U'U = I$) eigenvectors for $X'X$ are:

$$U = \begin{bmatrix} -0.574 & 0.816 & -0.066 \\ -0.654 & -0.408 & 0.636 \\ -0.493 & -0.408 & -0.768 \end{bmatrix} \quad (3)$$

Making a diagonal matrix of the reciprocals of the square roots of the eigenvalues, we have:

$$L^{-0.5} = \begin{bmatrix} 0.212 & 0 & 0 \\ 0 & 1.000 & 0 \\ 0 & 0 & 1.180 \end{bmatrix} \quad (4)$$

and the principal component scores $Y = XUL^{-0.5}$ become:

$$Y = \begin{bmatrix} 0.625 & -0.408 & -0.439 \\ -0.139 & -0.408 & 0.751 \\ -0.729 & 0 & -0.467 \\ 0.243 & 0.816 & 0.156 \end{bmatrix} \quad (5)$$

where each row of this matrix gives the three principal component scores for the corresponding data row in X .

Q-analysis

In *Q-analysis*, this process is reversed, and one studies the relationships among the observations rather than the variables. Uses of *Q-analysis* include the clustering of the individuals in the data set (see **Hierarchical Clustering**). Some **multidimensional scaling** techniques are an extension of *Q-analysis*, and are often used where the data are not homogeneous and require segmentation [4].

2 R & Q Analysis

In Q-analysis, an $n \times n$ covariance or correlation matrix will be formed for the observations and the eigenvectors, and principal component scores obtained from these. Generally, $n > p$ so that covariance or correlation matrices will not have full rank, and there will be a minimum of $n - p$ zero eigenvalues.

Using the same data matrix from the preceding section, the corresponding sums of squares and cross-products matrix become:

$$XX' = \begin{bmatrix} 9 & -2 & -10 & 3 \\ -2 & 1 & 2 & -1 \\ -10 & 2 & 12 & -4 \\ 3 & -1 & -4 & 2 \end{bmatrix} \quad (6)$$

with eigenvalues $l_1 = 22.282$, $l_2 = 1.000$, $l_3 = 0.718$, and $l_4 = 0$. The first three eigenvalues are identical to those in the Q-analysis. The significance of the fourth eigenvalue being zero is because XX' contains no more information than does $X'X$, and, hence, only has a rank of three.

Although one can obtain four eigenvectors from this matrix, the fourth one is not used as it has no length. The first three eigenvectors are:

$$U^* = \begin{bmatrix} 0.625 & -0.408 & -0.439 \\ -0.139 & -0.408 & 0.751 \\ -0.729 & 0 & -0.467 \\ 0.243 & 0.816 & 0.156 \end{bmatrix} \quad (7)$$

Note that this is the same as the matrix Y of principal scores obtained in the R-analysis above. If one obtains the principal component scores using these eigenvectors, (i.e., $Y^* = X'U^*L^{-0.5}$), it will be found that these principal component scores will be equal to the eigenvectors U of the R-analysis. Therefore, $Y^* = U$, and $Y = U^*$.

N-analysis (Singular Value Decomposition)

With proper scaling or normalization, as has been used in these examples, the eigenvectors of R-analysis become the principal component scores of Q-analysis, and vice versa. These relationships can be extended to *N-analysis* or the *singular value decomposition* [1, 5]. Here, the eigenvalues and vectors as well as the principal component scores for either R- or Q-analysis may be determined directly from the data matrix, namely:

$$X = YL^{0.5}U' = U^*L^{0.5}Y^* \quad (8)$$

The practical implication of these relationships is that the eigenvalues, eigenvectors, and principal component scores can all be obtained from the data matrix directly in a single operation. In addition, using the relationships above,

$$X = U^*L^{0.5}U' \quad (9)$$

This relationship is employed in *dual-scaling* techniques, where both variables and observations are being presented simultaneously. Examples of such a technique are the **biplot** [2] and MDPREF [4], which was designed for use with preference data (see **Scaling of Preferential Choice**). The graphical presentation of both of these techniques portrays both the variables and the observations on the same plot, one as vectors and the other as points projected against these vectors. These are not to be confused with the so-called 'point-point' plots, which use a different algorithm [6, Section 10.7].

Related Techniques

In addition to R-, Q-, and N-analysis, there are two more criteria, which, though more specialized, should be included for completeness. One of these, *M-analysis*, is used for a data matrix that has been corrected for both its column and row means (so-called *double-centering*). This technique has been used for the two-way **analysis of variance** where there is no estimate of error other than that included in the interaction term. The interaction sum of squares may be obtained directly from double-centered data. M-analysis may then be employed on these data to detect instances of nonadditivity, and/or obtain a better estimate of the true inherent variability [6, Section 13.7]. A version of M-analysis used in multidimensional scaling is a method known as *principal coordinates* [3, 8, 9].

In the antithesis of M-analysis, the original data are not corrected for either variable or observation means. This is referred to as *P-analysis*. In this case, the covariance or correlation matrix is replaced by a matrix made up of the raw sums of squares and cross-products of the data. This is referred to as a *product* or *second moment* matrix and, does not involve deviations about either row or column means. The method of principal components may be carried out on this matrix as well, but some of the usual properties such as rank require slight

modifications. This technique is useful for certain additive models, and, for this reason, many of the published applications appear to be in the field of chemistry, particularly with regard to Beer's Law. For some examples, see [6, Section 3.4].

References

- [1] Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika* **1**, 211–218.
- [2] Gabriel, K.R. (1971). The biplot-graphic display of matrices with application to principal component analysis, *Biometrika* **58**, 453–467.
- [3] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338.
- [4] Green, P.E., Carmone Jr, F.J. & Smith, S.M. (1989). *Multidimensional Scaling*, Allyn and Bacon, Needam Heights.
- [5] Householder, A.S. & Young, G. (1938). Matrix approximations and latent roots, *American Mathematical Monthly* **45**, 165–171.
- [6] Jackson, J.E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, New York.
- [7] Okamoto, M. (1972). Four techniques of principal component analysis, *Journal of the Japanese Statistical Society* **2**, 63–69.
- [8] Torgerson, W.S. (1952). Multidimensional scaling I: theory and method, *Psychometrika* **17**, 401–419.
- [9] Torgerson, W.S. (1958). *Theory and Methods of Scaling*, John Wiley & Sons, New York.

(See also **Multivariate Analysis: Overview**)

J. EDWARD JACKSON

***R*-squared, Adjusted *R*-squared**

JEREMY MILES

Volume 4, pp. 1655–1657

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

R-squared, Adjusted R-squared

Many statistical techniques are carried out in order to predict or explain variation in a measure – these include univariate techniques such as linear regression (*see* **Multiple Linear Regression**) and **analysis of variance**, and multivariate techniques, such as multilevel models (*see* **Linear Multilevel Models**), **factor analysis**, and **structural equation models**. A measure of the proportion of variance accounted for in a variable is given by *R*-squared (*see* **Effect Size Measures**).

The variation in an outcome variable (*y*) is represented by the sum of squared deviations from the mean, referred to as the total sum of squares (SS_{total}):

$$SS_{\text{total}} = \sum (y - \bar{y})^2 \quad (1)$$

(Note that dividing this value by $N - 1$ gives the variance.)

General linear models (which include regression and ANOVA) work by using least squares estimators; that is, they find parameter estimates and thereby predicted values that account for as much of the variance in the outcome variable as possible – the difference between the predicted value and the actual score for each individual is the **residual**. The sum of squared residuals is the error sum of squares, also known as the within groups sum of squares or residual sum of squares (SS_{error} , SS_{within} , or SS_{residual}). The variation that has been explained by the model is the difference between the total sum of squares and the residual sum of squares, and is called the *between groups sum of squares* or the *regression sum of squares* (SS_{between} or $SS_{\text{regression}}$).

R-squared is given by:

$$R^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} \quad (2)$$

In a standardized regression equation, where the correlations between variables are known, R^2 is given by:

$$R^2 = b_1 r_{yx_1} + b_2 r_{yx_2} + \dots + b_k r_{yx_k}, \quad (3)$$

where b_1 represents the standardized regression of *y* on x_1 , and r_{yx} represents the correlation between *y* and x .

Where the correlation matrix is known, the formula:

$$R_{i.123..k}^2 = 1 - \frac{1}{\mathbf{R}_{ii}^{-1}} \quad (4)$$

may be used, although this involves the inversion of the matrix $\mathbf{R}$, and should really only be attempted by computer (or by those with considerable time on their hands). $\mathbf{R}^{-1}$ is the inverse of the correlation matrix of all variables.

In the simple case of a regression with one predictor, the square of the correlation coefficient (*see* **Pearson Product Moment Correlation**) is equal to *R*-squared. However, this interpretation of *R* does not generalize to the case of multiple regression. A second way of considering *R* is to consider it as the correlation between the values of the outcome predicted by the regression equation and the actual values of the outcome. For this reason, *R* is sometimes considered to indicate the ‘fit’ of the model.

Cohen [1] has provided conventional descriptions of effect sizes for *R*-squared (as well as for other effect size statistics). He defines a small effect as being R^2 equal to 0.02, a medium effect as $R^2 = 0.13$, and a large effect as being $R^2 = 0.26$.

R^2 is a sample estimate of the proportion of variance explained in the outcome variables, and is biased upwards, relative to the population proportion of variance explained. To explain this, imagine we are in the unfortunate situation of having collected random numbers rather than real data (fortunately, we do not need to actually collect any data because we can generate these with a computer). The true (population) correlation of each variable with the outcome is equal to zero; however, thanks to sampling variation, it is very unlikely that any one correlation in our sample will equal zero – although the correlations will be distributed around zero. We have two variables that may be correlated negatively or positively, but to find R^2 we square them, and therefore they all become positive. Every time we add a variable, R^2 will increase; it will never decrease. If we have enough variables, we will find that R^2 is equal to 1.00 – we will have explained all of the variance in our sample, but this will of course tell us nothing about the population. In the long run, values of R^2 in our sample will tend to be higher than values of R^2 in the population (this does not mean that R^2 is always higher in the sample than in the population). In order to correct for this, we use adjusted R^2 ,

2 *R*-squared, Adjusted *R*-squared

calculated using:

$$\text{Adj. } R^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}, \quad (5)$$

where N is the sample size, and k is the number of predictor variables in the analysis. Smaller values for N , and larger values for k , lead to greater downward adjustment of R^2 . In samples taken from a population where the population value of R^2 is 0, the sample R^2 will always be greater than 0. Adjusted R^2 is centered on 0, and hence can become negative; but R^2 is a proportion of variance, and a variance can never be negative (it is the sum of squares) – a negative variance estimate therefore does not make sense and this must be an underestimate.

A useful source of further information is [2].

References

- [1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Mahwah, NJ.
- [2] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah, NJ.

JEREMY MILES

Random Effects and Fixed Effects Fallacy

PHILIP T. SMITH

Volume 4, pp. 1657–1661

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Random Effects and Fixed Effects Fallacy

Introduction

In most psychological experiments, the factors investigated consist of *fixed effects*, that is, all possible levels of each factor are included in the experiment. Clear examples of fixed factors include *Sex* (if both *male* and *female* are included) and *Interference* (in an experiment that manipulated distraction during a task, and included such conditions as *no interference*, *verbal interference*, and *visual interference*). In contrast, a *random effect* is where the levels of a factor included in the experiment do not exhaust the possible levels of the factor, but consist of a random sample from a population of levels. In most psychological experiments, there is one random effect, *Subjects*: the experimenter does not claim to have tested all the subjects who might have undertaken the task, but hopes that the conclusions from any statistical test apply not only to the people tested but also to the population from which they have been drawn.

Analysis of factorial experiments where there are several fixed-effects factors and one random-effects factor (usually subjects) is the core of an introductory course on **analysis of variance**, and methods and results for all simple designs are well understood. The situation is less clear if there are two random-effects factors in the same experiment, and some researchers have argued that this is the case in experiments involving materials drawn from language. Two artificial datasets have been constructed, shown in Table 1, to illustrate the problem (ignore for the moment the variable *AoA* in Table 1(b); this will be discussed later).

An experimenter is interested in word frequency effects in a categorization task. He selects three high-frequency words, w_1 to w_3 , and three low-frequency words, w_4 to w_6 . Four subjects, s_1 to s_4 , make decisions for all the words. Their decision times are recorded and shown as ‘RT’ in Table 1. Thus, this is a repeated measures design with the factor *Words* nested within the factor *Frequency*. This is a common design in psycholinguistic experiments, which has been chosen because it is used in many discussions of this topic (e.g., [1, 6, 7]). The actual examples are for illustrative purposes only: It is certainly not being

Table 1 Two artificial data sets

(a) Small variance between words				(b) Large variance between words				
<i>S</i>	<i>W</i>	<i>Freq</i>	RT	<i>S</i>	<i>W</i>	<i>Freq</i>	<i>AoA</i>	RT
S_1	w_1	<i>hi</i>	10	s_1	w_1	<i>hi</i>	1	9
S_1	w_2	<i>hi</i>	11	s_1	w_2	<i>hi</i>	3	11
S_1	w_3	<i>hi</i>	12	s_1	w_3	<i>hi</i>	4	13
S_1	w_4	<i>lo</i>	13	s_1	w_4	<i>lo</i>	2	12
S_1	w_5	<i>lo</i>	14	s_1	w_5	<i>lo</i>	4	14
S_1	w_6	<i>lo</i>	15	s_1	w_6	<i>lo</i>	6	16
S_2	w_1	<i>hi</i>	10	s_2	w_1	<i>hi</i>	1	9
S_2	w_2	<i>hi</i>	12	s_2	w_2	<i>hi</i>	3	12
S_2	w_3	<i>hi</i>	12	s_2	w_3	<i>hi</i>	4	13
S_2	w_4	<i>lo</i>	14	s_2	w_4	<i>lo</i>	2	13
S_2	w_5	<i>lo</i>	14	s_2	w_5	<i>lo</i>	4	14
S_2	w_6	<i>lo</i>	15	s_2	w_6	<i>lo</i>	6	16
S_3	w_1	<i>hi</i>	11	s_3	w_1	<i>hi</i>	1	10
S_3	w_2	<i>hi</i>	11	s_3	w_2	<i>hi</i>	3	11
S_3	w_3	<i>hi</i>	12	s_3	w_3	<i>hi</i>	4	13
S_3	w_4	<i>lo</i>	13	s_3	w_4	<i>lo</i>	2	12
S_3	w_5	<i>lo</i>	13	s_3	w_5	<i>lo</i>	4	13
S_3	w_6	<i>lo</i>	15	s_3	w_6	<i>lo</i>	6	16
S_4	w_1	<i>hi</i>	10	s_4	w_1	<i>hi</i>	1	9
S_4	w_2	<i>hi</i>	10	s_4	w_2	<i>hi</i>	3	10
S_4	w_3	<i>hi</i>	11	s_4	w_3	<i>hi</i>	4	12
S_4	w_4	<i>lo</i>	13	s_4	w_4	<i>lo</i>	2	12
S_4	w_5	<i>lo</i>	15	s_4	w_5	<i>lo</i>	4	15
S_4	w_6	<i>lo</i>	15	s_4	w_6	<i>lo</i>	6	16

suggested that it is appropriate to design experiments using such small numbers of subjects and stimuli.

One possible analysis is to treat *Frequency* and *Words* as fixed-effect factors. Such an analysis uses the corresponding interactions with *Subjects* as the appropriate error terms. Analyzed in this way, *Frequency* turns out to be significant ($F(1, 3) = 80.53$, $P < 0.01$) for both datasets in Table 1. There is something disturbing about obtaining the same results for both datasets: it is true that the means for *Frequency* are the same in both sets (hi *Frequency* has a mean RT of 11.00, lo *Frequency* has a mean RT of 14.08, in both sets), but the effect looks more consistent in dataset (a), where all the hi *Frequency* words have lower means than all the lo *Frequency* words, than in dataset (b), where there is much more variation. In an immensely influential paper, Herb Clark [1] suggested that the analyses we have just described are invalid, because *Words* is being treated as a fixed effect: other words could have been selected that meet our selection criteria (in the present case, to be of hi or lo *Frequency*) so *Words* should be

2 Random Effects and Fixed Effects Fallacy

treated as a random effect. Treating *Words* as a fixed effect is, according to Clark, the *Language-as-Fixed-Effect Fallacy*.

Statistical Methods for Dealing with Two Random Effects in the Same Experiment

F_1 and F_2

Treating *Words* as a fixed effect, as we did in the previous paragraph, is equivalent to averaging across *Words*, and carrying out an ANOVA based purely on *Subjects* as a random effect. This is known as a *by-subjects* analysis and the F values derived from it usually carry the suffix ‘1’; so, the above analyses have shown that $F_1(1, 3) = 80.53$, $P < 0.01$. An alternative analysis would be, for each word, to average across subjects and carry out an ANOVA based purely on *Words* as a random effect. This is known as a *by-materials* analysis (the phrase *by-items* is sometimes used). The F values derived from this analysis usually carry the suffix ‘2’. In the present case, the dataset (a) by-materials analysis yields $F_2(1, 4) = 21.39$, $P = 0.01$, which is significant, whereas dataset (b) by-materials analysis yields $F_2(1, 4) = 4.33$, $P = 0.106$, clearly nonsignificant. This accords with our informal inspection of Table 1, which shows a more consistent frequency effect for dataset (a) than for dataset (b).

It is sometimes said that F_1 assesses the extent to which the experimental results may generalize to new samples of subjects, and that F_2 assesses the extent to which the results will generalize to new samples of words. These statements are not quite accurate: neither F_1 nor F_2 are pure assessments of the presence of an effect. The standard procedure in an ANOVA is to estimate the variance due to an effect, via its mean square (MS), and compare this mean square with other mean squares in the analysis to assess significance. Using formulas, to be found in many textbooks (e.g. [13, 14]), the analysis of the Table 1 data as a three-factor experiment where *Freq* and *W* are treated as fixed effects and *S* is treated as a random effect yields the following equations:

$$E(MS_{\text{Freq}}) = \sigma_e^2 + q\sigma_{\text{Freq} \times S}^2 + nq\sigma_{\text{Freq}}^2 \quad (1)$$

$$E(MS_{W(\text{Freq})}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 + n\sigma_{W(\text{Freq})}^2 \quad (2)$$

$$E(MS_S) = \sigma_e^2 + pq\sigma_S^2 \quad (3)$$

$$E(MS_{\text{Freq} \times S}) = \sigma_e^2 + q\sigma_{\text{Freq} \times S}^2 \quad (4)$$

$$E(MS_{W(\text{Freq}) \times S}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 \quad (5)$$

E means expected or theoretical value, σ_A^2 refers the variance attributable to A , e is random error, n is the number of *Subjects*, p the number of levels of *Frequency*, and q the number of *Words*. The researcher rarely needs to know the precise detail of these equations, but we include them to make an important point about the choice of error terms in hypothesis testing: (1), $E(MS_{\text{Freq}})$, differs from (4) only in having a term referring to the variance in the data attributable to the *Frequency* factor, so we can test for whether *Frequency* makes a nonzero contribution to the variance by comparing (1) with (4), more precisely by dividing the estimate of variance described in (1) (MS_{Freq}) by the estimate of variance described in (4) ($MS_{\text{Freq} \times S}$). In other words, $MS_{\text{Freq} \times S}$ is the appropriate error term for testing for the effect of *Frequency*.

If *Frequency* is treated as a fixed effect, but *Words* and *Subjects* are treated as random effects, the variance equations change, as shown below.

$$E(MS_{\text{Freq}}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 + q\sigma_{\text{Freq} \times S}^2 + n\sigma_{W(\text{Freq})}^2 + nq\sigma_{\text{Freq}}^2 \quad (6)$$

$$E(MS_{W(\text{Freq})}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 + n\sigma_{W(\text{Freq})}^2 \quad (7)$$

$$E(MS_S) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 + pq\sigma_S^2 \quad (8)$$

$$E(MS_{\text{Freq} \times S}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 + q\sigma_{\text{Freq} \times S}^2 \quad (9)$$

$$E(MS_{W(\text{Freq}) \times S}) = \sigma_e^2 + \sigma_{W(\text{Freq}) \times S}^2 \quad (10)$$

The most important change is the difference between (1) and (6). Unlike (1), (6) contains terms involving the factor *Words*. Equation (6) is telling us that some of the variance in calculating MS_{Freq} is due to the possible variation of the effect for different selections of words (in contrast, (1) assumes all relevant words that have been included in the experiment, so different selections are not possible). This means that F_1 (derived from dividing (6) by (9)) is contaminated by *Words*: a significant F_1 could arise from a fortuitous selection of words. Similarly, F_2 (derived by dividing (6) by (7)) is contaminated by *Subjects*: a significant F_2 could arise from a fortuitous selection of subjects. By themselves, F_1 and F_2 are insufficient to solve the problem. (This is not to say that significant F_1 or F_2 are never worth reporting: for example,

practical constraints when testing very young children or patients might mean the number of stimuli that can be used is too small to permit an F_2 test of reasonable power: here F_1 would be worth reporting, though the researcher needs to accept that generalization to other word sets has yet to be established.)

Quasi- F Ratios

There are ratios that can be derived from (6) to (10), which have the desired property that the numerator differs from the denominator by only a term involving σ_{Freq}^2 . Two possibilities are

$$F' = \frac{(MS_{\text{Freq}} + MS_{W(\text{Freq}) \times S})}{(MS_{W(\text{Freq})} + MS_{\text{Freq} \times S})} \quad (11)$$

and

$$F'' = \frac{MS_{\text{Freq}}}{(MS_{W(\text{Freq})} + MS_{\text{Freq} \times S} - MS_{W(\text{Freq}) \times S})} \quad (12)$$

These F s are called *Quasi- F ratios*, reflecting the fact that they are similar to standard F ratios, but, because they are not the simple ratio of two mean squares, their distribution is only approximated by the standard F distribution. Winer [13, pp. 377–378] and Winer et al. [14, pp. 374–377], give the basic formulas for degrees of freedom in (11) and (12), derived from Satterthwaite [10]. It is doubtful that the reader will ever need to calculate such expressions by hand, so these are not given here. SPSS (Version 12) uses (12) to calculate quasi- F ratios: for example, if the data in Table 1 are entered into SPSS in exactly the form shown in the table, and if *Freq* is entered as a Fixed Factor and *S* and *W* are entered as Random Factors, and Type I SS are used, then SPSS suggests there is a significant effect of *Frequency* for dataset (a) ($F(1, 4.916) = 18.42$, $P < 0.01$), but not for dataset (b) ($F(1, 4.249) = 4.20$, $P = 0.106$). If *S* and *W* are truly random effects, then this is the correct statistical method. Many authorities, for example, [1, 6, 13], prefer (11) to (12) since (12) may on occasion lead to a negative number (see [5]).

Min F'

The method outlined in section ‘Quasi- F Ratios’ can be cumbersome: for any experiment with realistic

numbers of subjects and items, data entry into SPSS or similar packages can be very time-consuming, and if there are missing data (quite common in reaction-time experiments), additional corrections need to be made. A short cut is to calculate min F' , which, as its name suggests, is an estimate of F' that falls slightly below true F' . The formula is as follows:

$$\min F' = \frac{F_1 \cdot F_2}{(F_1 + F_2)} \quad (13)$$

The degrees of freedom for the numerator of the F ratio remains unchanged ($p - 1$), and the degrees of freedom for the denominator is given by

$$\text{df} = \frac{(F_1^2 + F_2^2)}{(F_1^2/\text{df}_1 + F_2^2/\text{df}_2)} \quad (14)$$

where df_1 is the error degrees of freedom for F_1 and df_2 is the error degrees of freedom for F_2 . For the data in Table 1, dataset (a) has $\min F'(1, 5.86) = 16.90$, $P < 0.01$, and dataset (b) has $\min F'(1, 4.42) = 4.11$, $P = 0.11$, all values being close to the true F' values shown in the previous section.

Best practice, then, is that when you conduct an ANOVA with two random-effects factors, use (11) or (12) if you can, but if you cannot, (13) and (14) provide an adequate approximation.

Critique

Clark’s paper had an enormous impact. From 1975 onward, researchers publishing in leading psycholinguistic journals, such as *Journal of Verbal Learning and Verbal Behavior* (now known as *Journal of Memory and Language*) accepted that something needed to be done in addition to a *by-subjects* analysis, and at first min F' was the preferred solution. Nowadays, min F' is hardly ever reported, but it is very common to report F_1 and F_2 , concluding that the overall result is significant if both F_1 and F_2 are significant. As Raaijmakers et al. [8] have correctly pointed out, this latter practice is wrong: simulations [4] have shown that using the simultaneous significance of F_1 and F_2 to reject the null hypothesis of no effect can lead to serious inflation of Type I errors. It has also been claimed [4] and [12] that min F' is too conservative, but this conservatism is quite small and the procedure quite robust to modest violations of the standard ANOVA assumptions [7, 9].

One reason for $\min F'$'s falling into disuse is its absence from textbooks psychology researchers are likely to read: Only one graduate level textbook has been found, by Allen Edwards [3], that gives a full treatment to the topic. Jackson & Brashers [6] give a very useful short overview, though their description of calculations using statistical packages is inevitably out of date. Another reason for not calculating $\min F'$ is that we have become so cosseted by statistics packages that do all our calculations for us, that we are not prepared to work out $\min F'$ and its fiddly degrees of freedom by hand. (If you belong to this camp, there is a website (www.pallier.org/ressources/MinF/compminf.htm) that will work out $\min F'$, its degrees of freedom and its significance for you.)

The main area of contention in the application of these statistical methods is whether materials should be treated as a *random* effect. This point was picked up by early critics of Clark [2, 12]: researchers do not select words at random, and indeed often go to considerable lengths to select words with appropriate properties ('It has often seemed to me that workers in this field counterbalance and constrain word lists to such an extreme that there may in fact be no other lists possible within the current English language.' [2, p.262]). Counterbalancing (for example, arranging that half the subjects receive word set *A* in condition C_1 and word set *B* in condition C_2 , and the other half of the subjects receive word set *B* in condition C_1 and word set *A* in condition C_2) would enable a by-subjects analysis to be carried out uncontaminated by effects of materials [8]. Counterbalancing, however, is not possible when the effect of interest involves intrinsic differences between words, as in the examples in Table 1: different words must be used if we want to examine word frequency effects.

Constraining word lists, that is selecting sets of words that are matched on variables we know to influence the task they are to be used in, is often a sensible procedure: it makes little sense to select words that differ widely in their frequency of occurrence in the language when we know that frequency often has a substantial influence on performance. The trouble with such procedures is that matching can never be perfect because there are too many variables that influence performance. One danger of using a constrained word set, which appears to give good results in an experiment, is that the experimenter, and others

who wish to replicate or extend his or her work, are tempted to use the same set of words in subsequent experiments. Such a procedure may be capitalizing on some as yet undetected idiosyncratic feature of the word set, and new sets of words should be used wherever possible. A further drawback is that results from constrained word sets can be generalized only to other word sets that have been constrained in a similar manner.

An alternative to using highly constrained lists is to include influential variables in the statistical model used to analyze the data (e.g., [11]). For example, another variable known to influence performance with words is *Age of Acquisition* (*AoA*). A researcher dissatisfied with the large variance displayed by different words in Table 1(b), and believing *AoA* was not adequately controlled, might add *AoA* as a covariate to the analysis, still treating *Subjects* and *Words* as random effects. This now transforms the previously nonsignificant quasi-*F* ratio for *Frequency* to a significant one ($F(1, 3.470) = 18.80, P < 0.05$).

A final remark is that many of the comments about treating *Words* as a random effect apply to treating *Subjects* as a random effect. In psycholinguistic experiments, we frequently reject subjects of low IQ or whose first language is not English, and, when we are testing older populations, we generally deal with a self-selected sample of above average individuals. In an area such as morphology, which is often not taught formally in schools, there may be considerable individual differences in the way morphemically complex words are represented. All of these examples suggest that attempts to model an individual *subject's* knowledge and abilities, for example, via covariates in **analyses of covariance**, could be just as important as modeling the distinct properties of individual *words*.

References

- [1] Clark, H.H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research, *Journal of Verbal Learning and Verbal Behavior* **12**, 335–359.
- [2] Clark, H.H., Cohen, J., Keith Smith, J.E. & Keppel, G. (1976). Discussion of Wike and Church's comments, *Journal of Verbal Learning and Verbal Behavior* **15**, 257–266.
- [3] Edwards, A.L. (1985). *Experimental Design in Psychological Research*, 2nd Edition, Harper & Row, New York.

-
- [4] Forster, K.I. & Dickinson, R.G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1 , F_2 , F' , and min F' , *Journal of Verbal Learning and Verbal Behavior* **15**, 135–142.
 - [5] Gaylor, D.W. & Hopper, F.N. (1969). Estimating the degrees of freedom for linear combinations of means squares by Satterthwaite's formula, *Technometrics* **11**, 691–706.
 - [6] Jackson, S. & Brashers, D.E. (1994). *Random Factors in ANOVA*, Sage Publications, Thousand Oaks.
 - [7] Maxwell, S.F. & Bray, J.H. (1986). Robustness of the quasi F statistic to violations of sphericity, *Psychological Bulletin* **99**, 416–421.
 - [8] Raaijmakers, J.G.W., Schrijnemakers, J.M.C. & Gremmen, F. (1999). How to deal with "The Language-as-Fixed-Effect Fallacy": common misconceptions and alternative solutions, *Journal of Memory and Language* **41**, 416–426.
 - [9] Santa, J.L., Miller, J.J. & Shaw, M.L. (1979). Using quasi F to prevent alpha inflation due to stimulus variation, *Psychological Bulletin* **86**, 37–46.
 - [10] Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**, 110–114.
 - [11] Smith, P.T. (1988). How to conduct experiments with morphologically complex words, *Linguistics* **26**, 699–714.
 - [12] Wike, E.L. & Church, J.D. (1976). Comments on Clark's "The Language-as-Fixed-Effect Fallacy", *Journal of Verbal Learning and Verbal Behavior* **15**, 249–255.
 - [13] Winer, B.J. (1971). *Statistical Principles in Experimental Design*, 2nd Edition, McGraw-Hill Kogakusha, Tokyo.
 - [14] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Design*, 3rd Edition, McGraw-Hill, New York.

PHILIP T. SMITH

Random Effects in Multivariate Linear Models: Prediction

NICHOLAS T. LONGFORD

Volume 4, pp. 1661–1665

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Random Effects in Multivariate Linear Models: Prediction

Random effects are a standard device for representing the differences among observational or experimental units that are or can be perceived as having been drawn from a well-defined population. Such units may be subjects (individuals), their organizations (businesses, classrooms, families, administrative units, or teams), or settings within individuals (time periods, academic subjects, sets of tasks, and the like). The units are associated with random effects because they are incidental to the principal goal of the analysis – to make inferences about an *a priori* specified population of individuals, geographical areas, conditions, or other factors (contexts).

In modeling, random effects have the advantage of parsimonious representation, that a large number of quantities are summarized by a few parameters that describe their distribution. When the random effects have a (univariate) normal distribution, it is described completely by a single variance; the mean is usually absorbed in the regression part of the model. The fixed-effects counterparts of these models are **analysis of covariance** (ANCOVA) models, in which each effect is represented by one or a set of parameters.

In the resampling perspective, fixed effects are used for factors that are not altered in hypothetical replications. Typically, factors with few levels (categories), such as experimental conditions or treatments, which are the focus of the inference, are regarded as fixed. In contrast, a different set of random effects is realized in each hypothetical replication; the replications share only the distribution of the effects. A logical inconsistency arises when the analyzed sample is an enumeration. For example, when the districts of a country are associated with random effects, a replication would yield a different set of district-level effects. Yet, a more natural replication, considered in sampling theory in particular, keeps the effects fixed for each district – the same set of districts would be realized. This conflict is resolved by a reference to a superpopulation, arguing that inferences are desired for a domain *like* the one analyzed,

and in each replication a different domain is realized with a different division into districts.

A constructive way of addressing the issue of fixed *versus* random effects is by admitting that incorrect models may be useful for inference. That is, the effects are fixed, but it is advantageous to treat them in inference as random. Apart from a compact description of the collection of units, by their (estimated) distribution or its parameters, random effects enable estimation of unit-specific quantities that is more efficient than in the **maximum likelihood** or **least squares** for standard fixed effects ANCOVA. The efficiency is achieved by *borrowing strength* across the units [9]. Its theoretical antecedent is the work on shrinkage estimation [5] and its application to small-area statistics [3].

Borrowing strength can be motivated by the following general example. If the units are similar, then the pooled (domain-related) estimator of the quantity of interest θ may be more efficient for the corresponding unit-related quantity θ_j , because the squared bias $(\theta_j - \theta)^2$ is much smaller than the sampling variance $\text{var}(\hat{\theta}_j)$ of the unbiased estimator of θ_j . Instead of selecting the domain estimator $\hat{\theta}$ or the unbiased large-variance estimator $\hat{\theta}_j$, these two estimators are combined,

$$\tilde{\theta}_j = (1 - b_j)\hat{\theta} + b_j\hat{\theta}_j, \quad (1)$$

with a constant b_j , or its estimator $\hat{b}_j$, for which the combination has some optimal properties, such as minimum mean squared error (MSE). The combination (*composition*) $\tilde{\theta}_j$ can be interpreted as exploiting the similarity of the units. The gains are quite dramatic when the units are similar and $\text{var}(\hat{\theta}_j) \gg \text{var}(\hat{\theta})$. That occurs when there are many (aggregate-level) units j and most of them are represented in the dataset by only a few observations each.

Inference about the individual units is usually secondary to studying the population as a whole. Nevertheless, interest in units on their own may arise as a result of an inspection of the data or their analysis that aimed originally at some population features. Model diagnostics are a notable example of this.

Estimation of random effects is usually referred to as *prediction*, to avoid the terminological conflict of ‘estimating random variables’, a contradiction in terms if taken literally. The task of prediction is to define a function of the data that is, in a well-defined

2 Random Effects in Multivariate Linear Models: Prediction

sense, as close to the target as possible. With random effects, the target does not appear to be stationary.

In fact, *the realizations* of random variables, or quantities that are fixed across replications but regarded as random in the model, are estimated. Alternatively, the prediction can be described as estimating the quantity of interest *given* that it is fixed in the replications; the corresponding quantities for the other units are assumed to vary across replications. The properties of such an estimator (predictor) should be assessed *conditionally* on the realized value of the target.

Random-effects models involving normality and linearity are greatly preferred because of their analytical tractability, easier interpretation, and conceptual proximity to ordinary regression (*see Multiple Linear Regression*) and ANCOVA. We discuss first the prediction of random effects with the model

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \delta_j + \boldsymbol{\varepsilon}_j, \quad (2)$$

where $\mathbf{y}_j$ is the $n_j \times 1$ vector of (univariate) outcomes for (aggregate or level-2) unit $j = 1, \dots, N_2$, $\mathbf{X}_j$ is the regression design matrix for unit j , $\boldsymbol{\beta}$ the vector of regression parameters, δ_j the random effect for unit j , and $\boldsymbol{\varepsilon}_j$ the vector of its elementary-level (residual) terms, or deviations (*see Variance Components*). The random terms δ_j and $\boldsymbol{\varepsilon}_j$ are mutually independent, with respective centered normal distributions $\mathcal{N}(0, \sigma_2^2)$ and $\mathcal{N}(0, \sigma^2 \mathbf{I}_{n_j})$; $\mathbf{I}$ is the identity matrix of the size given in the subscript.

The model in (2) can be interpreted as a set of related regressions for the level-2 units. They have all the regression coefficients in common, except for the intercept $\beta_0 + \delta_j$. The regressions are parallel. The obvious generalization allows any regression coefficients to vary, in analogy with introducing group-by-covariate interactions in ANCOVA. Thus, a subset of the covariates in $\mathbf{X}$ is *associated with variation*. The corresponding submatrix of $\mathbf{X}_j$ is denoted by $\mathbf{Z}_j$ (its dimensions are $n_j \times r$), and the model is

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j, \quad (3)$$

where $\boldsymbol{\delta}_j \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{\Sigma})$, independently ($\mathbf{0}_r$ is the $r \times 1$ column vector of zeros), [4] and [7]. In agreement with the ANCOVA conventions, $\mathbf{Z}_j$ usually contains the intercept column $\mathbf{1}_{n_j}$. Variables that are constant within groups j can be included in $\mathbf{Z}$, but the interpretation in terms of varying regressions does not

apply to them because the within-group regressions are not identified for them.

The random effects $\boldsymbol{\delta}_j$ are estimated from their conditional expectations given the outcomes $\mathbf{y}$. The matrices $\mathbf{X}_j$ and $\mathbf{Z}_j$ are assumed to be known, or are conditioned on, even when they depend on the sampling or the data-generation process. Assuming that the parameters $\boldsymbol{\beta}$, σ^2 and those involved in $\boldsymbol{\Sigma}$ are known, the conditional distribution of $\boldsymbol{\delta}_j$ is normal,

$$(\boldsymbol{\delta}_j | \mathbf{y}, \boldsymbol{\theta}) \sim \mathcal{N} \left\{ \frac{1}{\sigma^2} \boldsymbol{\Sigma} \mathbf{G}_j^{-1} \mathbf{Z}_j^\top \mathbf{e}_j, \boldsymbol{\Sigma} \mathbf{G}_j^{-1} \right\}, \quad (4)$$

where $\mathbf{G}_j = \mathbf{I}_r + \sigma^{-2} \mathbf{Z}_j^\top \mathbf{Z}_j \boldsymbol{\Sigma}$ and $\mathbf{e}_j = \mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}$. The vector $\boldsymbol{\delta}_j$ is predicted by its (naively) estimated conditional expectation. The univariate version of this estimator, for the model in (2), corresponding to $\mathbf{Z}_j = \mathbf{1}_{n_j}$, is

$$\tilde{\delta}_j = \frac{n_j \hat{\omega}}{1 + n_j \hat{\omega}} \bar{e}_j, \quad (5)$$

where $\hat{\omega}$ is an estimate of the variance ratio $\omega = \sigma_2^2 / \sigma^2$ (σ_2^2 is the univariate version of $\boldsymbol{\Sigma}$) and $\bar{e}_j = (\mathbf{y}_j - \mathbf{X}_j \hat{\boldsymbol{\beta}})^\top \mathbf{1}_{n_j} / n_j$ is the average residual in unit j . Full or restricted maximum likelihood estimation (MLE) can be applied for $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ and $\hat{\boldsymbol{\Sigma}}$ or $\hat{\sigma}_2^2$. In general, restricted MLE is preferred because the estimators of σ^2 and σ_2^2 are unbiased. As the absence of bias is not maintained by nonlinear transformations, this preference has a poor foundation for predicting $\boldsymbol{\delta}_j$. Thus, even the restricted MLE of ω , the ratio of two unbiased estimators, $\hat{\omega} = \hat{\sigma}_2^2 / \hat{\sigma}^2$, is biased. The bias of $\hat{\omega}$ can be corrected, but it does not lead to an unbiased estimator $\tilde{\delta}_j$, because $1/(1 + n_j \omega)$ is estimated with bias.

These arguments should not be interpreted as claiming superiority of full MLE over restricted MLE, merely that no bias is not the right goal to aim for. Absence of bias is not a suitable criterion for estimation in general; minimum MSE, combining bias and sampling variance, is more appropriate. Prediction of random effects is an outstanding example of successful (efficient) biased estimation. Bias, its presence and magnitude, depend on the resampling (replication) perspective adopted. Reference [10] presents a viewpoint in which the estimators we consider are unbiased. In fact, the terminology ‘best linear unbiased predictor’ (BLUP) is commonly used, and is

appropriate when different units are realized in replications. Indeed, $E(\tilde{\delta}_j - \delta_j | \omega) = 0$ when the expectation is taken both over sampling within units j and over the population of units j , because $E(\tilde{\delta}_j) = E(\delta_j) = 0$. In contrast,

$$E(\tilde{\delta}_j | \delta_j; \omega) = \frac{n_j \omega}{1 + n_j \omega} \delta_j, \quad (6)$$

so $\tilde{\delta}_j$ is *conditionally* biased. The conditional properties of $\tilde{\delta}_j$ are usually more relevant. The return, sometimes quite generous, for the bias is reduced sampling variance.

When δ_j are regarded as fixed effects, their least-squares estimator (and also MLE) is $\hat{\delta}_j = \bar{e}_j$. As $\tilde{\delta}_j = q_j \hat{\delta}_j$, where $q_j = n_j \hat{\omega} / (1 + n_j \hat{\omega}) < 1$, $\tilde{\delta}_j$ can be interpreted as a *shrinkage* estimator and q_j , or more appropriately $1 - q_j = 1 / (1 + n_j \hat{\omega})$, as a shrinkage coefficient. The coefficient q_j is an increasing function of both the sample size n_j and $\hat{\omega}$; more shrinkage takes place (q_j is smaller) for units with smaller sample sizes and when $\hat{\omega}$ is smaller. That is, for units with small samples, greater weight is assigned to the overall domain (its average residual $\bar{e} = (n_1 \bar{e}_1 + \dots + n_{N_2} \bar{e}_{N_2}) / (n_1 + \dots + n_{N_2})$ vanishes either completely or approximately) – the resulting bias is preferred to the substantial sampling variance of $\bar{e}_j$. Small ω indicates that the units are very similar, so the average residuals $\bar{e}_j$ differ from zero mainly as a result of sampling variation; shrinkage then enables estimation more efficient than by $\bar{e}_j$. The same principles apply to multivariate random effects, although the discussion is not as simple and the motivation less obvious.

Shrinkage estimation is a form of empirical Bayes estimation (see **Bayesian Statistics**). In Bayes estimation, a prior distribution is imposed on the model parameters, in our case, the random effects δ_j . In empirical Bayes estimation, the prior distribution is derived (estimated) from the same data to which it is subsequently applied; see [8] for examples in educational measurement. Thus, the prior distribution of δ_j is $\mathcal{N}(0, \sigma_2^2)$, and the posterior, with σ_2^2 and other model parameters replaced by their estimates, is $\mathcal{N}(\tilde{\delta}_j, \hat{\sigma}_2^2 / (1 + n_j \hat{\omega}))$.

Somewhat loosely, $\hat{\sigma}_2^2 / (1 + n_j \hat{\omega})$ is quoted as the sampling variance of $\tilde{\delta}_j$, and its square root as the standard error. This is incorrect on several counts. First, these are *estimators* of the sampling variance or standard error. Next, they estimate the sampling

variance for a particular replication scheme (with δ_j as a random effect) assuming that the model parameters β , σ^2 , and σ_2^2 are known. For large-scale data, the uncertainty about β and σ^2 can be ignored, because their estimation is based on many degrees of freedom. However, σ_2^2 is estimated with at most N_2 degrees of freedom, one for each level-2 unit, and N_2 is much smaller than the elementary-level sample size $n = n_1 + \dots + n_{N_2}$. Two factors complicate the analytical treatment of this problem; $\tilde{\delta}_j$ is a nonlinear function of ω and the precision of $\hat{\omega}$ depends on the (unknown) ω . The next element of ‘incorrectness’ of using $\hat{\sigma}_2^2 / (1 + n_j \hat{\omega})$ is that it refers to an ‘average’ unit j with the sample size n_j . Conditioning on the sample size is a common practice, even when the sampling design does not guarantee a fixed sample size n_j for the unit. But the sample size can be regarded as auxiliary information, so the conditioning on it is justified. However, $\tilde{\delta}_j$ is biased, so we should be concerned with its MSE:

$$\begin{aligned} \text{MSE}(\tilde{\delta}_j; \delta_j) &= E\{(\tilde{\delta}_j - \delta_j)^2 | \delta_j\} \\ &= \frac{(n_j \omega)^2}{(1 + n_j \omega)^2} \frac{\sigma^2}{n_j} + \frac{\delta_j^2}{(1 + n_j \omega)^2}, \end{aligned}$$

assuming that β , σ^2 , and ω are known and the sample-average residual $\bar{e}$ vanishes. Rather inconveniently, the MSE depends on the target δ_j itself. The conditional variance is obtained by replacing δ_j^2 with its expectation σ_2^2 .

Thus, $\hat{\sigma}_2^2 / (1 + n_j \hat{\omega})$ is an estimator of the *expected* MSE (eMSE), where the expectation is taken over the distribution of the random effects $\delta_{j'}$, $j' = 1, \dots, N_2$. It underestimates $\text{eMSE}(\tilde{\delta}_j; \delta_j)$ because some elements of uncertainty are ignored. As averaging is applied, it is not a particularly good estimator of $\text{MSE}(\tilde{\delta}_j; \delta_j)$. It is sometimes referred to as the comparative standard error [4]. The MSE can be estimated more efficiently by bootstrap [2] or by framing the problem in terms of incomplete information [1], and representing the uncertainty by plausible values of the unknown parameters, using the principles of **multiple imputation**, [11] and [12]. Approximations by various expansions are not very effective because they depend on the variances that have to be estimated.

The normal distribution setting is unusual by its analytical tractability, facilitated by the property that the normality and homoscedasticity are maintained

4 Random Effects in Multivariate Linear Models: Prediction

by conditioning. These advantages are foregone with **generalized mixed linear models**. They are an extension of **generalized linear models** that parallels the extension of linear regression to random coefficient models:

$$g\{E(\mathbf{y}_j|\delta_j)\} = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\delta_j, \quad (7)$$

where g is a monotone function, called the *link function*, and the assumptions about all the other terms are the same as for the random coefficient model that corresponds to normality and identity link (see **Generalized Linear Mixed Models**). The conditional distribution of $\mathbf{y}_j$ given δ_j has to be specified; extensive theory is developed for the case when this distribution belongs to the exponential family (see **Generalized Linear Models (GLM)**). The realization of δ_j is estimated, in analogy with BLUP in the normality case, by estimating its conditional expectation given the data and parameter estimates. In general, the integral in this expectation is not tractable, and we have to resort to numerical approximations. These are computationally manageable for one- or two-dimensional δ_j , especially if the number of units j in the domain, or for which estimation of δ_j is desired, is not excessive. Some approximations avoid the integration altogether, but are not very precise, especially when the between-unit variance σ_2^2 (or $\boldsymbol{\Sigma}$) is substantial. The key to such methods is an analytical approximation to the (marginal) likelihood, based on Laplace transformation or quasilielihood. These methods have been developed to their apparently logical conclusion in the h -likelihood [6]. Fitting models by h -likelihood involves no integration, the random effects can be predicted without any extensive computing, and more recent work by the authors is concerned with joint modeling of location and variation structures and detailed diagnostics.

In principle, any model can be extended to its random-effects version by assuming that a separate model applies to each (aggregate) unit, and specifying how the parameters vary across the units. No variation (identical within-unit models) is a special case in such a model formulation. Modeling is then concerned with the associations in an average or typical

unit, and with variation within and across the units. Unit-level random effects represent the deviation of the model for a given unit from the average unit. Units can differ in all aspects imaginable, including their level of variation, so random effects need not be associated only with regression or location, but can be considered also for variation and any other model features.

References

- [1] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* **39**, 1–38.
- [2] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [3] Fay, R.E. & Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269–277.
- [4] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Edition, Edward Arnold, London.
- [5] James, W. & Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 361–379.
- [6] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society* **58**, 619–678.
- [7] Longford, N.T. (1993). *Random Coefficient Models*, Oxford University Press, Oxford.
- [8] Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*, Springer-Verlag, New York.
- [9] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications, *Journal of the American Statistical Association* **78**, 55–65.
- [10] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects, *Statistical Science* **6**, 15–51.
- [11] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley and Sons, New York.
- [12] Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489.

NICHOLAS T. LONGFORD

Random Forests

ADELE CUTLER

Volume 4, pp. 1665–1667

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Random Forests

Random forests were introduced by Leo Breiman in 2001 [1], and can be thought of as **bagging classification and regression trees (CART)**, for which each node is split using a random subset of the variables, and not pruning. More explicitly, we select a bootstrap sample (*see Bootstrap Inference*) from the data and fit a binary decision tree to the bootstrap sample. To fit the tree, we split nodes by randomly choosing a small number of variables and finding the best split on these variables only. For example, in a classification problem for which we have, say, 100 input variables, we might choose 10 variables at random, independently, each time a split is to be made. For every distinct split on these 10 variables, we compute some measure of node purity, such as the *gini index* [2], and we select the split

that optimizes this measure. Cases on each side of the split form new nodes in the tree, and the splitting procedure is repeated until all the nodes are pure. We typically grow the tree until it is large, with no pruning, and then combine the trees as with bagging (averaging for regression and voting for classification).

To illustrate, we use the R [8] function *randomForest* to fit a classifier to the data in Figure 1. The classification boundary and the data are given in Figure 1(a). In Figures 1(b), 1(c), and 1(d), the shading intensity indicates the weighted vote for class 1. As more trees are included, the nonlinear boundary is estimated more accurately.

Studies (e.g., [4]) show that random forests are about as accurate as *support vector machines* [6] and **boosting** [3], but unlike these competitors, random forests are interpretable using several quantities that we can compute from the forest.

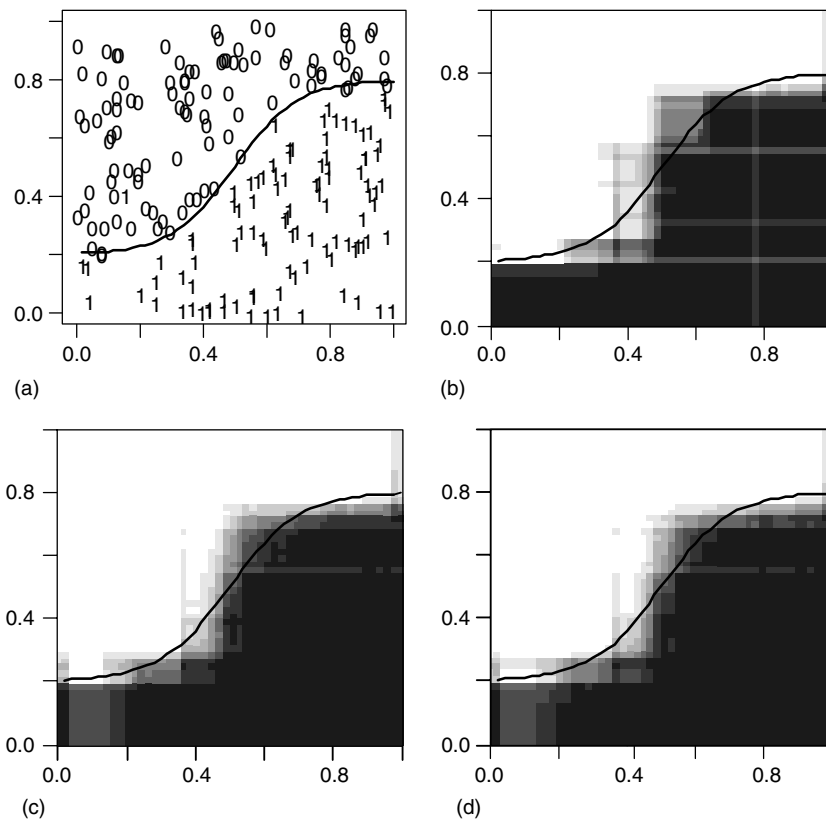


Figure 1 (a) Data and underlying function; (b) random forests, 10 trees; (c) random forests, 100 trees; and (d) random forests, 400 trees

2 Random Forests

The first such quantity is *variable importance*. We compute variable importance by considering the cases that are left out of a bootstrap sample ('out-of-bag'). If we are interested in the importance of variable 3, for example, we randomly permute variable 3 in the out-of-bag data. Then, using the tree that we obtained from the bootstrap sample, we subtract the prediction accuracy for the permuted out-of-bag data from that for the original out-of-bag data. If variable 3 is important, the permuted out-of-bag data will have lower prediction accuracy than the original out-of-bag data, so the difference will be positive. This measure of variable importance for variable 3 is averaged over all the bootstrap samples, and the procedure is repeated for each of the other input variables.

A second important quantity for interpreting random forests is the proximity matrix (see **Proximity Measures**). The proximity between any two cases is computed by looking at how often they end up in the same terminal node. These quantities, suitably standardized, can be used in a proximity-based clustering (see **Hierarchical Clustering**) or **multidimensional scaling** procedure to give insight about the data structure. For example, we might pick out subgroups of cases that almost always stay together in the trees, or **outliers** that are almost always alone in a terminal node.

Random forests can be used in a clustering context by thinking of the observed data as class 1, creating a

synthetic second class, and using the random forests' classifier. The synthetic second class is created by randomly permuting the values of each input variable. The proximities from random forests can be used in a proximity-based clustering procedure.

More details on random forests can be obtained from <http://stat-www.berkeley.edu/users/breiman/RandomForests>, along with freely available software.

References

- [1] Breiman, L. (2001). Random Forests, *Machine Learning* **45**(1), 5–32.
- [2] Breiman, L., Friedman, J.H., Olshen, R. & Stone, C. (1983). *Classification and Regression Trees*, Wadsworth.
- [3] Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm, in *Machine Learning: Proceedings of the Thirteenth International Morgan Kauffman, San Francisco, CA. Conference*, 148–156.
- [4] Meyer, D., Leisch, F., Hornik, K. (2002). Adaptive information systems and modeling in economics and management science, Report No 78., Benchmarking support vector machines. <http://www.wu-wien.ac.at/am/>
- [5] R Development Core Team, (2004). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna. www.R-project.org.
- [6] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer.

ADELE CUTLER

Random Walks

DONALD LAMING

Volume 4, pp. 1667–1669

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Random Walks

Suppose there is an accident on a motorway that reduces traffic past that point to a cautious one-vehicle-at-a-time on the hard shoulder. A mile or so in advance, the traffic is channelled into two lanes and, as you reach that two-lane restriction, you find yourself level with a Rolls-Royce. Thereafter, sometimes the Rolls-Royce edges forward a length, sometimes it is your turn, and Figure 1 shows how the relative position of the two cars develops. To your chagrin, the Rolls-Royce has crept ahead; will you ever catch up?

A random walk is the cumulative sum of a series of independent and identically distributed random variables, $\sum_1^n X_i$, and Figure 1 is a simple example (as also is the forward progress of either car). As a vehicle somewhere ahead edges past the scene of the accident, you or the Rolls-Royce (but not both) can move forward one car length – one step in the random walk. Assuming that the two lanes feed equally and at random past the accident, then the relative positions of the two cars is analogous to the difference in the numbers of heads and tails in a sequence of tosses of a fair coin. If the sequence continues for long enough, it is certain that the numbers of heads and tails will, at some point, be equal, but the mean wait is infinite. That is to say, you will most probably pass the point of restriction before you draw level with the Rolls-Royce.

Each step in Figure 1 could, of course, be itself the sum of a number of independent and identically distributed random variables. Suppose I let a drop of

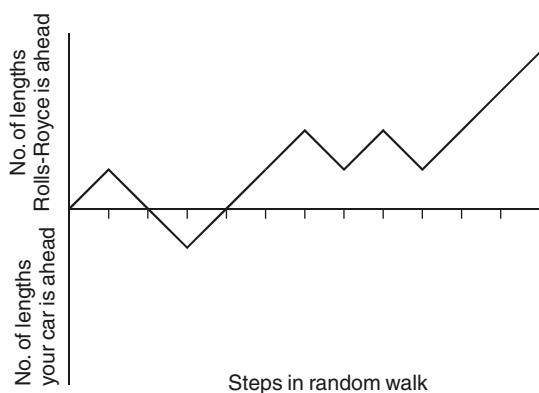


Figure 1 Relative position of two cars on a motorway

black ink fall into a glass of water. The ink slowly diffuses throughout the water, driven by Brownian motion. Suppose Figure 1 represents the drift of a notional particle of black ink on the left–right axis. Each step can then be split into an arbitrary number of substeps. If the substeps are independent and identically distributed, then the random walk is actually a random process, unfolding in continuous time. But such a decomposition (into an arbitrary number of independent and identically distributed substeps) is possible only if the distribution of each step is *infinitely divisible*. Amongst well-known probability distributions, the normal, the Poisson, and the gamma (or chi-squared) (see **Catalogue of Probability Density Functions**) distributions are infinitely divisible. In addition, a compound Poisson distribution (in which each Poisson event is itself a random variable) is infinitely divisible with respect to its Poisson parameter, so the class of infinitely divisible distributions is very broad. But amongst these possibilities only the normal (or Wiener) process is continuous with respect to the spatial dimension; all the other random processes contain jumps.

Interest in random walks began in the 18th century with gamblers wanting to know the chances of their being ruined. Suppose the game is ‘absolutely fair’ (see **Martingales**), so that the probabilities of winning and losing are equal. The paths in Figure 2 trace out different gamblers’ cumulative wins and losses. If a gambler should ever lose his or her entire fortune, he or she will have nothing left to gamble with, and this is represented by his or her time line (path) in Figure 2 descending to the axis at 0. This poses the following question: What is the probability that the random walk will ever fall below a certain specified value (the gambler’s entire fortune)? In more detail, how does the random walk behave if we delete all those gamblers who are ruined from the point of their bankruptcy onwards (broken lines in Figure 2)? For the binomial walk of Figure 1, this is a simple problem. Clearly, any walk that strays below the horizontal boundary must be struck out from that point onwards. But we must also delete those continuations of such a random walk that happen to rise upwards from the boundary as well as those that continue below. This may be achieved by introducing a mirror-image source (dotted lines in Figure 2) below the boundary. The fortunes of a gambler who has escaped ruin may be represented by the difference

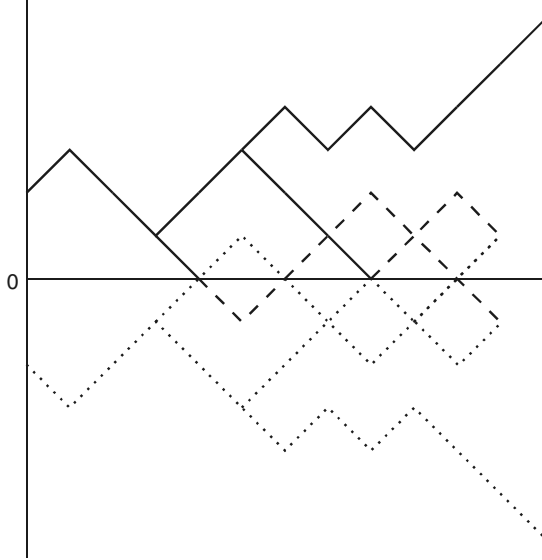


Figure 2 Calculation of a random walk with one absorbing barrier by deletion of an image process

between these two random processes, the original and the mirror image, above the horizontal boundary [2].

The mirror-image technique works equally for random processes in continuous time; and a simple modification to the argument adjusts it to the case where the random walk drifts up or down (the gambler is playing with skilled card sharps and loses more often than he wins) [1, p. 50]. If the basic random process is a (normal) Wiener process, the time taken to reach the boundary (to be ruined) is given by the *Wald distribution*

$$f(t) = \left[\frac{a}{\sqrt{(2\pi\sigma^2 t^3)}} \right] \exp \left\{ \frac{-(a - \mu t)^2}{2\sigma^2 t} \right\}, \quad (1)$$

where a is the distance to the boundary (the gambler's fortune) and μ and σ^2 the rates at which the mean and variance of the random process increase per unit time. Schwartz [5] has used this distribution, convolved with an exponential to provide the characteristic long tail, as a model for simple reaction times.

Random walks have also been proposed as models for two-choice reaction times [6]. There are now two boundaries placed on either side of a starting point (Figure 3), one corresponding to each response. The response depends on which boundary is reached

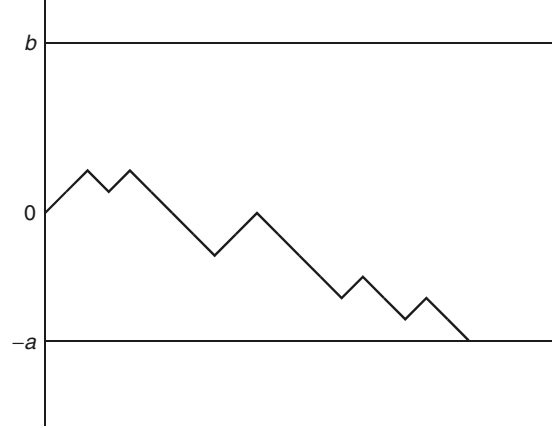


Figure 3 Random walk with two absorbing barriers

first; the reaction time is, of course, the time taken to reach that boundary. So this model bids to determine both choice of response and latency from a common process.

The distribution of reaction time now depends on the location of two absorbing boundaries as well as the statistical properties of the processes representing the two alternative stimuli. Unfortunately, the simple argument using mirror-image sources is not practicable now; it generates an infinite series of sources stretching out beyond both boundaries. But the response probabilities and the moments of the reaction time distributions can be readily obtained from Wald's identity [7, p. 160]. Let $\varphi(\omega)$ be the characteristic function of the random process per unit time. Let Z be the terminal value of the process on one or the other boundary. Then

$$E \left\{ \frac{\exp(Z\omega)}{[\varphi(\omega)]^t} \right\} = 1. \quad (2)$$

Even so, the development of this model is not simple except in two special cases.

There are two distinct random processes involved, representing the two alternative stimuli, A and B. The special cases are distinguished by the relationship between these two processes.

1. Suppose that $f_B(x) = e^x f_A(x)$. The variable x is then the probability ratio between the two alternatives, and the reaction time model may be interpreted as a sequential probability ratio test between the two stimuli [3].

2. Suppose instead that $f_B(x) = f_A(-x)$. The two processes are now mirror images of each other [4].

The nature of reaction times is such that a random walk has an intuitive appeal as a possible model; this is especially so with two-choice reaction times in which both probabilities of error and reaction times derive from a common source. But the relationship of experimental data to model predictions has not provided great grounds for confidence; it has typically been disappointing.

References

- [1] Bartlett, M.S. (1955). *An Introduction to Stochastic Processes*, Cambridge University Press, Cambridge.
- [2] Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy, *Reviews of Modern Physics* **15**, 1.
- [3] Laming, D. (1968). *Information Theory of Choice-Reaction Times*, Academic Press, London.
- [4] Link, S.W. (1975). The relative judgment theory of two choice response time, *Journal of Mathematical Psychology* **12**, 114–135.
- [5] Schwartz, W. (2001). The ex-Wald distribution as a descriptive model of response times, *Behavior Research Methods, Instruments & Computers* **33**, 457–469.
- [6] Stone, M. (1960). Models for choice-reaction time, *Psychometrika* **25**, 251–260.
- [7] Wald, A. (1947). *Sequential Analysis*, Wiley, New York.

DONALD LAMING

Randomization

ROBERT J. BOIK

Volume 4, pp. 1669–1674

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Randomization

Introduction

Randomization is the intentional use of a random process either in the design phase (prerandomization) or in the analysis phase (postrandomization) of an investigation. Prerandomization includes random selection, random assignment, and randomized response methods. Postrandomization includes randomized decision rules and randomization-based inference methods such as permutation tests and **bootstrap methods**. The focus of this article is on random selection, random assignment, and their relationship to randomization-based inference. Definitions of simple random assignment and selection along with a brief discussion of the origins of randomization are given in this section. Justifications for and criticisms of randomization are given in the next section.

Simple random selection refers to any process that selects a sample of size n without replacement from a population of size $N > n$, such that each of the $N!/([n!(N-n)!])$ possible samples is equally likely to be selected. Simple random assignment refers to any process that assigns one of t treatments to each of N subjects, such that each of the $N!/(n_1!n_2!, \dots, n_t!)$ possible assignments in which treatment j is assigned to n_j subjects is equally likely. A method for performing random assignment, as well as a warning about a faulty assignment method, is described in [26]. For convenience, the term randomization will be used in this article to refer either to random selection or to random assignment. Details on **randomized response** methods can be found in [67]. Randomized decision rules are described in [11, §1.5] and [40, §9.1].

The prevailing use of random assignment in experimentation owes much to **R. A. Fisher**, who in 1926 [17], apparently was the first to use the term randomization [7]. Random assignment, however, had been used much earlier, particularly in behavioral science research. Richet [49] used random assignment in an 1884 telepathy experiment in which subjects guessed the suit of a card. The Society for Psychical Research in London was receptive to using random assignment and by 1912 it was being used in parapsychological research at American universities. Random assignment also was used as early

as 1885 in psychophysics experiments [42], but the procedure was not as readily accepted here as it was in parapsychological research. Fisher made random assignment a formal component of experimental design, and he introduced the method of permutation tests (*see* **Permutation Based Inference**) [18]. Pitman [44–46] provided a theoretical framework for permutation tests and extended them to tests on correlations and to **analysis of variance**.

Random sampling also has origins within social science research. Kiaer [34] proposed in 1897 that a representative (purposive) sample rather than a census be used to gather data about an existing population. The proposal met substantial opposition, in part because the method lacked a theoretical framework. Bowley, in 1906 [3], showed how the **central limit theorem** could be used to assess the accuracy of population estimates based on simple random samples. Work on the theory of stratified random sampling (*see* **Stratification**) had begun by 1923 [64], but as late as 1926 random sampling and purposive sampling were still treated on equal grounds [4]. It was **Neyman** [41] who provided the theoretical framework for random sampling and set the course for future research. In this landmark paper, Neyman introduced the randomization-based sampling distribution, described the theory of stratified random sampling with optimal allocation, and developed the theory of **confidence intervals**. Additional discussions on the history of randomization can be found in [16, 25, 29, 43, 47, 48, 57, 59], and [61].

Why Randomize?

It might seem unnecessary even to ask the question posed by the title of this section. For most behavioral scientists, the issue was resolved in the sophomore-level experimental psychology course. It is apparent, however, that not all statisticians take this course. At least two articles with the title ‘Why Randomize’ are widely cited [23, 32]. A third article asked ‘Must We Randomize Our Experiment?’ and answered – ‘sometimes’ [2]. A fourth asked ‘Experimental Randomization: Who Needs It?’ and answered – ‘nobody’ [27]. Additional articles that have addressed the question posed in this section include [5, 6, 9, 24, 36, 38, 55, 60–63, 65], and [66]. Arguments for randomization as well as selected criticisms are summarized next.

To Ensure that Linear Estimators are Unbiased and Consistent. A statistic is unbiased for estimating a parameter if the mean of its **sampling distribution** is equal to the value of the parameter, regardless of which value the parameter might have. A statistic is consistent for a parameter if the statistic converges in probability to the value of the parameter as sample size increases.

A sampling distribution for a statistic can be generated by means of postrandomization, provided that prerandomization was employed for data collection. Sampling distributions generated in this manner are called *randomization distributions*. In an observational study, random sampling is sufficient to ensure that sample means are unbiased and consistent for population means. Random sampling also ensures that the empirical cumulative distribution function is unbiased and consistent for the population cumulative distribution function and this, in turn, is the basis of the bootstrap. Likewise, random assignment together with unit-treatment additivity ensures that differences between means of treatment groups are unbiased and consistent estimators of the true treatment differences, even if important explanatory variables have been omitted from the model.

Unbiasedness is generally thought of as a desirable property, but an estimator may be quite useful without being unbiased. First, biased estimators are sometimes superior to unbiased estimators with respect to mean squared error (variance plus squared bias). Second, unbiasedness can be criticized as an artificial advantage because it is based on averaging over treatment assignments or subject selections that could have been but were not observed [5]. Averaging over data that were not observed violates the likelihood principle, which states that inferences should be based solely on the likelihood function given the observed data. A brief introduction to the issues regarding the likelihood principle can be found in [50] (see **Maximum Likelihood Estimation**).

To Justify Randomization-based Inference. One of the major contributions of Neyman [41] was to introduce the randomization distribution for survey sampling. Randomization distributions provide a basis for assessing the accuracy of an estimator (e.g., standard error) as well as a framework for constructing confidence intervals.

Randomization distributions based on designed experiments are particularly useful for testing sharp

null hypotheses. For example, suppose that treatment and control conditions are randomly assigned to subjects and that administration of the treatment would have an additive effect, say δ , for each subject. The permutation test, based on the randomization distribution of treatment versus control means, provides an exact test of the hypothesis $\delta = 0$. Furthermore, a confidence interval for δ can be obtained as the set of all values δ_0 for which $H_0: \delta = \delta_0$ is not rejected using the permutation test. In general, randomization plus subject-treatment additivity eliminates the need to know the exact process that generated the data.

It has been suggested that permutation tests are meaningful even when treatments are not randomly assigned [9, 12–15, 19, 39, 54]. The resulting P values might have descriptive value but without random assignment they do not have inferential value. In particular, they cannot be used to make inferences about causation [24].

Randomization-based inference has been criticized because it violates the conditionality principle [1, 10, 27]. This principle states that inference should be made conditional on the values of ancillary statistics; that is, statistics whose distributions do not depend on the parameter of interest. The outcome of randomization is ancillary. Accordingly, to obey the conditionality principle, inference must be made conditional on the observed treatment assignment or sample selection. Postrandomizations do not yield additional information. This criticism of randomization-based inference loses much of its force if the model that generated the data is unknown. Having a valid inference procedure in the absence of distributional knowledge is appealing and appears to outweigh the cost of violating the conditionality principle.

To Justify Normal-theory Tests. Kempthorne [30, 31] showed that if treatments are randomly assigned to subjects and unit-treatment additivity holds, then the conventional F Test is justified even in the absence of normality. The randomization distribution of the test statistic under H_0 is closely approximated by the central F distribution. Accordingly, the conventional F Test can be viewed as an approximation to the randomization test. In addition, this result implies that the choice of the linear model is not *ad hoc*, but follows from randomization together with unit-treatment additivity.

To Protect Against Subjective Biases of the Investigator. Randomization ensures that treatment assignment is not affected by conscious or subconscious biases of the experimenter. This justification has been criticized on the grounds that an investigator who cannot be trusted without randomization does not become trustworthy by using randomization [27]. The issue, however, is not only about trust. Even the most trustworthy experimenter could have an unintentional influence on the outcome of the study. The existence of unintentional experimenter effects is well documented [53] and there is little reason to believe that purposive selection or purposive assignment would be immune from such effects.

To Elucidate Causation. It has been argued that the scientific (as opposed to statistical) purpose of randomization is to elucidate causation [33, 62]. Accurate inferences about causality are difficult to make because experimental subjects are heterogeneous and this variability can lead to bias. Random assignment guards against pretreatment differences between groups on recorded variables (overt bias) as well as on unobserved variables (hidden bias).

In a designed experiment, the causal effect of one treatment relative to a second treatment for a specific subject can be defined as the difference between the responses under the two treatments. Most often in practice, however, only one treatment can be administered to a specific subject. In the counterfactual approach (*see Counterfactual Reasoning*), the causal effect is taken to be the difference between potential responses that would be observed under the two treatments, assuming subject-treatment additivity and no **carryover effects** [28, 52, 55, 56]. These treatment effects cannot be observed, but random assignment of treatments is sufficient to ensure that differences between the sample means of the treatment groups are unbiased and consistent for the true causal effects. The counterfactual approach has been criticized on the grounds that assumptions such as subject-treatment additivity cannot be empirically verified [8, 20, 63].

‘Randomization is rather like insurance [22].’ It protects one against biases, but it does not guarantee that treatment groups will be free of pretreatment differences. It guarantees only that over the long run, average pretreatment differences are zero. Nonetheless, even after a bad random assignment, it is unlikely that treatment contrasts will be completely

confounded with pretreatment differences. Accordingly, if treatment groups are found to differ on important variables after random assignment, then covariate adjustment still can be used (*see Analysis of Covariance*). This role of reducing the probability of confounding is not limited to frequentist analyses; it also is relevant in Bayesian analyses [37].

Under certain conditions, causality can be inferred without random assignment. In particular, if experimental units are homogeneous, then random assignment is unnecessary [55]. Also, random assignment is unnecessary (but may still be useful) under covariate sufficiency. Covariate sufficiency is said to exist if all covariates that affect the response are observed [63]. Under covariate sufficiency, hidden bias is nonexistent and adjustment for differences among the observed covariates is sufficient to remove overt bias, even if treatments are not randomly assigned. Causal inferences from structural equation models fitted to observational data as in [58] implicitly require covariate sufficiency. Without this condition, inferences are limited to ruling out causal patterns that are inconsistent with the observed data. Causal inference in observational studies without covariance sufficiency is substantially more difficult and is sensitive to model misspecification [68, 69].

Furthermore, random assignment is unnecessary for making causal inferences from experiments whenever treatments are assigned solely on the basis of observed covariates, even if the exact assignment mechanism is unknown [21]. The conditional probability of treatment assignment given the observed covariates is known as the **propensity score** [52]. If treatments are assigned solely on the basis of observed covariates, then adjustment for differences among the propensity scores is sufficient to remove bias. One way to ensure that treatment assignment is solely a function of observed covariates is to randomly assign treatments to subjects, possibly after blocking on one or more covariates.

To Enhance Robustness of Inferences. Proponents of optimal experimental design and sampling recommend that treatments be purposively assigned and that subjects be purposively selected using rational judgments rather than random processes. The advantage of purposive assignment and selection is that they can yield estimators that are more efficient than those based on randomization [35, 51]. If the presumed model is not correct, however, then the

resulting inferences may be faulty. Randomization guards against making incorrect inferences due to model misspecification.

For example, consider the problem of constructing a regression function for a response, Y , given a single explanatory variable, X . If the regression function is known to be linear, then the variance of the least squares slope estimator is minimized by selecting observations for which half of the X s are at the maximum and half of the X s are at the minimum value (see **Optimal Design for Categorical Variables**). If the true regression function is not linear, however, then the resulting inference will be incorrect and the investigator will be unable to perform diagnostic checks on the model. In contrast, if (X, Y) pairs are randomly selected then standard regression diagnostic plots can be used to detect model misspecification and to guide selection of a more appropriate model.

References

- [1] Basu, D. (1980). Randomization analysis of experimental data: the fisher randomization test, (with discussion), *Journal of the American Statistical Association* **75**, 575–595.
- [2] Box, G.E.P. (1990). Must we randomize our experiment? *Journal of Quality Engineering* **2**, 497–502.
- [3] Bowley, A.L. (1906). Address to the economic science and statistics section of the British Association for the Advancement of Science, 1906, *Journal of the Royal Statistical Society* **69**, 540–558.
- [4] Bowley, A.L. (1926). Measurement of the precision attained in sampling, *Bulletin of the International Statistics Institute* **22** (Supplement to Liv. 1), 6–62.
- [5] Bunke, H. & Bunke, O. (1978). Randomization. Pro and contra, *Mathematische Operationsforschung und Statistik, Series Statistics* **9**, 607–623.
- [6] Cox, D.R. (1986). Some general aspects of the theory of statistics, *International Statistical Review* **54**, 117–126.
- [7] David, H.A. (1995). First (?) occurrence of common terms in mathematical statistics, *The American Statistician* **49**, 121–133.
- [8] Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion), *Journal of the American Statistical Association* **95**, 407–448.
- [9] Easterling, R.G. (1975). Randomization and statistical inference, *Communications in Statistics* **4**, 723–735.
- [10] Efron, B. (1978). Controversies in the foundations of statistics, *The American Mathematical Monthly* **85**, 231–246.
- [11] Ferguson, T.S. (1967). *Mathematical Statistics A Decision Theoretic Approach*, Academic Press, New York.
- [12] Finch, P.D. (1979). Description and analogy in the practice of statistics, *Biometrika* **67**, 195–208.
- [13] Finch, P.D. (1981). A descriptive interpretation of standard error, *Australian Journal of Statistics* **23**, 296–299.
- [14] Finch, P.D. (1982). Comments on the role of randomization in model-based inference, *Australian Journal of Statistics* **24**, 146–147.
- [15] Finch, P. (1986). Randomization – I, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 516–519.
- [16] Fienberg, S.E. & Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting, *International Statistical Review* **55**, 75–96.
- [17] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* **33**, 700–725.
- [18] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [19] Freedman, D.A. & Lane, D. (1983). Significance testing in a nonstochastic setting, *Journal of Business and Economic Statistics* **1**, 292–298.
- [20] Gadbury, G.L. (2001). Randomization inference and bias of standard errors, *The American Statistician* **55**, 310–313.
- [21] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC, Boca Raton.
- [22] Gibson, B. (2000). How many Rs are there in statistics? *Teaching Statistics* **22**, 22–25.
- [23] Greenberg, B.G. (1951). Why randomize? *Biometrics* **7**, 309–322.
- [24] Greenland, S. (1990). Randomization, statistics, and casual inference, *Epidemiology* **1**, 421–429.
- [25] Hacking, I. (1988). Telepathy: origins of randomization in experimental design, *Isis* **79**, 427–451.
- [26] Hader, R.J. (1973). An improper method of randomization in experimental design, *The American Statistician* **27**, 82–84.
- [27] Harville, D.A. (1975). Experimental randomization: who needs it? *The American Statistician* **29**, 27–31.
- [28] Holland, P.W. (1986). Statistics and causal inference (with discussion), *Journal of the American Statistical Association* **81**, 945–970.
- [29] Holschuh, N. (1980). Randomization and design: I, in *R. A. Fisher: An Appreciation*, S.E. Fienberg & D.V. Hinkley, eds, Springer Verlag, New York, pp. 35–45.
- [30] Kempthorne O. (1952). *The Design and Analysis of Experiments*, New York: John Wiley & Sons.
- [31] Kempthorne O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.
- [32] Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference* **1**, 1–25.
- [33] Kempthorne, O. (1986). Randomization – II, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 519–525.
- [34] Kiaer, A.N. (1976). *The Representative Method of Statistical Surveys*, Central Bureau of Statistics of Norway, Oslo. English translation of 1897 edition which was

- issued as No. 4 of The Norwegian Academy of Science and Letters, The Historical, Philosophical Section.
- [35] Kiefer, J. (1959). Optimal experimental designs, *Journal of the Royal Statistical Society, Series B* **21**, 272–319.
- [36] Kish, L. (1989). Randomization for statistical designs, *Statistica* **49**, 299–303.
- [37] Lindley, D.V. & Novic, M. (1981). The role of exchangeability in inference, *Annals of Statistics* **9**, 45–58.
- [38] Mayo, O. (1987). Comments on randomization and the Design of Experiments by P. Urbach, *Philosophy of Science* **54**, 592–596.
- [39] Mier, P. (1986). Damned liars and expert witnesses, *Journal of the American Statistical Association* **81**, 269–276.
- [40] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Edition, McGraw-Hill, New York.
- [41] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion), *Journal of the Royal Statistical Society* **97**, 558–625.
- [42] Peirce, C.S. & Jastrow, J. (1885). On small differences of sensation, *Memoirs of the National Academy of Sciences* **3**, 75–83.
- [43] Picard, R. (1980). Randomization and design: II, in R. A. Fisher: *An Appreciation*, S.E. Fienberg & D.V. Hinkley, eds, Springer Verlag, New York, pp. 46–58.
- [44] Pitman, E.G. (1937a). Significance tests which may be applied to samples from any populations, *Journal of the Royal Statistical Society Supplement* **4**, 119–130.
- [45] Pitman, E.G. (1937b). Significance tests which may be applied to samples from any populations: II. The correlation coefficient test, *Journal of the Royal Statistical Society Supplement* **4**, 225–232.
- [46] Pitman, E.G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test, *Biometrika* **29**, 322–335.
- [47] Preece, D.A. (1990). R. A. Fisher and experimental design: a review, *Biometrics* **46**, 925–935.
- [48] Rao, J.N.K. & Bellhouse, D.R. (1990). History and development of the theoretical foundations of survey based estimation and analysis, *Survey Methodology* **16**, 3–29.
- [49] Richet, C. (1884). La suggestion mentale et le calcul des probabilités, *Revue Philosophique de la France et de l'Étranger* **18**, 609–674.
- [50] Robins, J. & Wasserman, L. (2002). Conditioning, likelihood, and coherence: a review of some foundational concepts, in *Statistics in the 21st Century*, A.E. Raftery, M.A. Tanner & M.T. Wells, eds, Chapman & Hall/CRC, Boca Raton, pp. 431–443.
- [51] Royall, R.M. (1970). On finite population sampling theory under certain linear regression models, *Biometrika* **57**, 377–387.
- [52] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.
- [53] Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York.
- [54] Rouanet, H., Bernard, J.-M. & LeCoutre, B. (1986). Nonprobabilistic statistical inference: a set theoretic approach, *The American Statistician* **40**, 60–65.
- [55] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.
- [56] Rubin, D.B. (1990). Formal modes of statistical inference for causal effects, *Journal of Statistical Planning and Inference* **25**, 279–292.
- [57] Senn, S. (1994). Fisher's game with the devil, *Statistics in Medicine* **13**, 217–230.
- [58] Shipley, B. (2000). *Cause and Correlation in Biology*, Cambridge University Press, Cambridge.
- [59] Smith, T.M.F. (1976). The foundation of survey sampling: a review, *Journal of the Royal Statistical Society, Series A* **139**, 183–204.
- [60] Smith, T.M.F. (1983). On the validity of inferences from non-random samples, *Journal of the Royal Statistical Society, Series A* **146**, 394–403.
- [61] Smith, T.M.F. (1997). Social surveys and social science, *The Canadian Journal of Statistics* **25**, 23–44.
- [62] Sprott, D.A. & Farewell, V.T. (1993). Randomization in experimental science, *Statistical Papers* **34**, 89–94.
- [63] Stone, R. (1993). The assumptions on which causal inferences rest, *Journal of the Royal Statistical Society, Series B* **55**, 455–466.
- [64] Tchuprov, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron* **2**, 646–680.
- [65] Thornett, M.L. (1982). The role of randomization in model-based inference, *Australian Journal of Statistics* **24**, 137–145.
- [66] Urbach, P. (1985). Randomization and the design of experiments, *Philosophy of Science* **52**, 256–273.
- [67] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.
- [68] Winship, C. & Mare, R.D. (1992). Models for sample selection bias, *Annual Review of Sociology* **18**, 327–350.
- [69] Winship, C. & Morgan, S.L. (1999). The estimation of causal effects from observational data, *Annual Review of Sociology* **25**, 659–706.

ROBERT J. BOIK

Randomization Based Tests

JOHN LUDBROOK

Volume 4, pp. 1674–1681

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Randomization Based Tests

Introduction

My brief is to provide an overview of the randomization model of statistical inference, and of the statistical tests that are appropriate to that model. I may have exceeded that brief. In the first instance, I have felt obliged to present and discuss one of its major rivals: the population model of inference. In the second instance, I have not confined my description to tests on continuous or rank-ordered data under the randomization model. It has seemed to me that the so-called ‘exact’ tests on categorical data should also be considered under this model for they, too, involve permutation (*see* **Exact Methods for Categorical Data**).

I shall try to differentiate between the population and randomization models of statistical inference. I shall not consider the Bayesian model, partly (at least) because it is not popular, and partly because I have to confess that I do not fully understand its application to real-life experimentation (*see* **Bayesian Statistics**).

The Population Model of Statistical Inference

This is sometimes known as the classical model. It was first articulated in stringent theoretical terms by **Neyman** and **Pearson** with respect to continuous data [21, 22] (*see* **Neyman–Pearson Inference**). It presupposes, or has as its assumptions, the following: (a) the samples are drawn randomly from defined populations, and (b) the frequency distributions of the populations are mathematically definable. An alternative definition embraces the notion that if a large (infinite) number of samples (with replacement) were to be taken from the population(s) actually sampled in the experiment, then the P value attached to the null hypothesis corresponds to the frequency with which the values of the test statistic (for instance, the difference between means) are equal to or exceed those in the actual experiment. It should be noted that Neyman and Pearson also introduced the notions of Type 1 error (α , or false rejection of the null hypothesis) and Type 2 error (β , or false acceptance

of the null hypothesis) [21, 22]. By extension, this led to the notion of **power** to reject the null hypothesis ($1 - \beta$). Though Neyman and Pearson’s original proposals were concerned with tests of significance (P values), Neyman later introduced the notion of **confidence intervals** (CIs) [20].

It is important to note that statistical inferences under this model, whether they are made by hypothesis-testing (P values) or by estimation (CIs), refer to the parent populations from which the random samples were drawn.

The mathematically definable populations to which the inferences refer are usually normally distributed or are derivatives of the normal (Gaussian) distribution (for instance, t , F , χ^2).

A late entrant to the population model of inference is the technique of **bootstrapping**, invented by Bradley Efron in the late 1970s [5]. Bootstrapping is done, using a fast computer, by random resampling of the samples (because the populations are inaccessible), with replacement. It allows inferences to be made (P values, SEs, CIs) that refer to randomly sampled populations, but with the difference from classical statistical theory that no assumptions need be made about the frequency distribution of the populations.

The historical relationship between the enunciation of the population model of inference and the first descriptions of statistical tests that are valid under this model is rather curious. This is because the tests were described before the model was. ‘Student’ (**W.S. Gosset**) described what came to be known as the t distribution in 1908 [28], later converted into a practical test of significance by Fisher [6]. **R.A. Fisher** gave, in 1923, a detailed account of his use of **analysis of variance (ANOVA)** to evaluate the results of a complex experiment involving 12 varieties of potato, 6 different manures, and 3 replicates in a **randomized block design** [9]. The analysis included the Variety $\times$ Manure interaction. All this, performed with pencil and paper! But it was not until 1928 that Neyman and Pearson expounded the population model of inference [21, 22].

So, what is wrong with the population model of inference? As experimental biologists (not least, behavioral scientists) should know – but rarely admit – we never take random samples (*see* **Randomization**). At best, we take nonrandom samples of the experimental units (humans, animals, or whatever) that are available – ‘samples of

convenience' [16]. The availability may come about because units have presented themselves to a clinic, have responded to a call for volunteers, or can be purchased from animal breeders. We then randomize the experimental units to 'treatment groups', for instance, no treatment (control), placebo, various drugs, or various environmental manipulations. In these circumstances, it is impossible to argue that genuine populations have been randomly sampled. Enter, the randomization model of inference and tests under that model.

The Randomization Model of Statistical Inference

This was enunciated explicitly only about 50 years ago [12], even though statistical tests under this model had been described and performed since the early 1930s. This is undoubtedly because, until computers were available, it might take days, weeks, or even months to analyze the results of a single experiment. Much more recently, this and other models have been critically appraised by Rubin [26]. The main features of the model are that (a) experimental groups are not acquired by random sampling, but by taking a nonrandom sample and allocating its members to two or more 'treatments' by a process of randomization; (b) tests under this model depend on a process of permutation (randomization); (c) the tests do not rely on mathematically defined frequency-distributions; (d) inferences under this model do not refer to populations, but only to the particular experiment; (e) any wider application of these statistical inferences depends on scientific (verbal), not statistical, argument. Good accounts of this model for nonmathematical statisticians are given in monographs [4, 11, 16, 18] (*see* **Permutation Based Inference**).

Arguments in favor of this model have been provided by R.A. Fisher [7]: '...conclusions [from t or F Tests] have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method [randomization].' And by Kempthorne [12], who concluded: 'When one considers the whole problem of experimental inference, that is, of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using

[a] method of inference other than randomization analysis.'

The randomization model of inference and the statistical tests conducted under that model have attracted little attention from theoretical statisticians. Why? My guess is that this is because, to theoretical statisticians, the randomization model is boring. There are some exceptions [2, 25] but, as best I can judge, these authors write of inferences being referable to populations. I argue that because the experimental designs involve randomization, not random sampling, they are wrong.

What statistical tests are appropriate to randomization? For continuous data, the easiest to understand are those designed to test for differences between or among means [15]. In 1935, R.A. Fisher analyzed, by permutation, data from an experiment on matched pairs of plants performed by Charles Darwin [8]. Fisher's goal was to show that breaches of the assumptions for Student's t Test did not affect the outcome. In 1936, Fisher analyzed craniometric data from two independent groups by permutation [7]. In 1933, Eden and Yates reported a much more ambitious analysis of a complex experiment by permutation, to show that analysis of variance (ANOVA) was not greatly affected by breaches of assumptions [3]. As I shall show shortly, the calculations involved in these studies can only be described as heroic. Perhaps, because of this, Fisher repudiated permutation tests in favor of t Tests and ANOVA. He said that their utility '... consists in their being able to supply confirmation. whether rightly or, more often, wrongly, when it is suspected that the simpler tests have been apparently injured by departure from normality' [8]. It is strange that Fisher, the inventor, practitioner, and proselytizer of randomization in experimental design [8], seems not to have made the connection between randomization in design and the use of permutation (randomization) tests to analyze the results. All Fisher's papers can be downloaded free-of-charge from the website www.library.adelaide.edu.au/digitised/fisher/.

Over the period 1937 to 1952, the notion of transforming continuous data into ranks was developed [10, 13, 19, 29] (*see* **Rank Based Inference**). The goal of the authors was to simplify and expedite analysis of variance, and to cater for data that do not fulfill the assumption of normality in the

parent populations. All four sets of authors relied on calculating approximate (asymptotic) P values by way of the χ^2 or z distributions. It is true that **Wilcoxon** [29] and Kruskal and Wallis [13] did produce small tables of exact P values resulting from the permutation of ranks. But it is curious that only Kruskal and Wallis [13] refer to Fisher's idea of exact tests based on the permutation of means [7, 8].

There is a misapprehension, commonplace among investigators but also among some statisticians [26], that the rank-order tests are concerned with differences between medians. This is simply untrue. These tests are concerned with differences in group mean ranks ($\bar{R}_1 - \bar{R}_2$), and it can easily be demonstrated that, because of the method of ranking, this is not the same as differences between medians [1]. It should be said that though Wilcoxon was the first to describe exact rank-order tests, it was Milton Friedman (later a Nobel prize-winner in economics) who first introduced the notion of converting continuous into rank-ordered data in 1937 [10], but he did not embrace the notion of permutation.

Now, the analysis of categorical data by permutation – this, too, was introduced by R.A. Fisher [8]. He conjured up a hypothetical experiment. It was that a colleague reckoned she could distinguish a cup of tea to which the milk had been added first from a cup in which the milk had been added after the tea (Table 3). This well-known ‘thought’ experiment was the basis for what is now known as the Fisher exact test. Subsequently, this exact method for the analysis of categorical data in 2×2 tables of frequency has been extended to $r \times c$ tables of frequencies, both unordered and ordered.

As a piece of history, when **Frank Yates** described his correction for continuity for the χ^2 test on small 2×2 tables, he validated his correction by reference to Fisher's exact test [30].

Statistical Tests Under the Randomization Model of Inference

Differences Between or Among Means

Two Independent Groups of Size n_1 and n_2 . The procedure followed is to exchange the members of the groups in all possible permutations, maintaining the original group sizes (Tables 1 and 2). That is, all

Table 1 Illustration of the process of permutation in the case of two independent, randomized, groups

Permutation	Group 1 $n = 2$	Group 2 $n = 3$
A	a, b	c, d, e
B	a, c	b, d, e
C	a, d	b, c, e
D	a, e	b, c, d
E	b, c	a, d, e
F	b, d	a, c, e
G	b, e	a, c, d
H	c, d	a, b, d
I	c, e	a, b, d
J	d, e	a, b, c

The number of possible permutations (see Formula 2) is $(2 + 3)!/(2)!(3)! = 10$, whether the entries are continuous or ranked data.

Table 2 Numerical datasets corresponding to Table 1

Dataset	Group 1	Group 2	$\bar{x}_2 - \bar{x}_1$	$\bar{R}_2 - \bar{R}_1$
A	1, 3	4, 7, 9	4.67	2.50
B	1, 4	3, 7, 9	3.83	1.67
C	1, 7	3, 4, 9	1.33	0.83
D	1, 9	3, 4, 7	-0.33	0.00
E	3, 4	1, 7, 9	2.17	0.83
F	3, 7	1, 4, 9	-0.33	0.00
G	3, 9	1, 4, 7	-2.00	-0.83
H	4, 7	1, 3, 9	-1.17	-0.83
I	4, 9	1, 3, 7	-2.83	-1.67
J	7, 9	1, 3, 4	-5.33	-2.50

1, 3, 4, 7 and 9 were substituted for a, b, c, d and e in Table 1, and the differences between means ($\bar{x}_2 - \bar{x}_1$) calculated. The ranks 1, 2, 3, 4 and 5 were substituted for a, b, c, d and e in Table 1, and the differences between mean ranks ($\bar{R}_2 - \bar{R}_1$) calculated.

Exact permutation test on difference between means [7]: dataset A, one-sided $P = 0.100$, two-sided $P = 0.200$; dataset B, one-sided $P = 0.200$, two-sided $P = 0.300$; dataset J, one-sided $P = 0.100$, two-sided $P = 0.100$.

Exact permutation test on difference between mean ranks [18, 28]: dataset A, one-sided $P = 0.100$, two-sided $P = 0.200$; dataset B, one-sided $P = 0.200$, two-sided $P = 0.400$; dataset J, one-sided $P = 0.100$, two-sided $P = 0.200$.

possible ways in which the original randomization could have fallen out are listed. Then:

$$P = \frac{\text{No. of permutations in which the difference between group means is equal to or more extreme than that observed}}{\text{All possible permutations}}. \quad (1)$$

4 Randomization Based Tests

For a one-sided P , only differences that are equal to or greater than that observed (that is, with + sign) are used. For a two-sided P , differences, regardless of the sign, that are greater than that observed are used.

The number of all possible permutations increases steeply with group size. The formula for this, for group sizes of n_1 and n_2 , and corresponding group means $\bar{x}_1$ and $\bar{x}_2$, is:

$$\frac{(n_1 + n_2)!}{(n_1)!(n_2)!} \quad (2)$$

This innocuous formula disguises the magnitude of the computational problem. Thus, for $n_1 = n_2 = 5$, the number of all possible permutations is 252. For $n_1 = n_2 = 10$, it is 184 756. And for $n_1 = n_2 = 15$, it is 155 117 520. A solution to the sometimes massive computational problem is to take Monte Carlo random samples (see **Monte Carlo Simulation**) of, say, 10 000 from the many millions of possible permutations. Interestingly, this is what Eden and Yates did in solving the massive problem of their complex analysis of variance – by the ingenious use of cards [3].

Matched Pairs. Given n matched pairs, the number of all possible permutations is

$$2^n. \quad (3)$$

Thus, if $n = 5$, the number of all possible permutations is 32; for $n = 10$, 1024; and for $n = 15$, 32768. The last is what R.A. Fisher computed with pencil and paper in 1935 [8].

k Independent Groups. This corresponds to one-way ANOVA. It is a simple extension of two independent groups, and the number of all possible permutations is given by an extension of formula (2). Usually, instead of using the difference between the means, one uses a simplified F statistic [4, 11, 18].

k Matched Groups. This corresponds to two- or multi-way ANOVA (see **Factorial Designs**). There is no great problem if only the main effects are to be extracted. But it is usually the interactions that are the focus of interest (see **Interaction Effects**). There is no consensus on how best to go about extracting these. Edgington [4], Manly [18], and Good [11] have suggested how to go about this.

In the case of a two-way, factorial, design first advocated by Fisher [9], there seems to be no great problem. Good [11] describes clearly how, first, the main (fixed) effects should be factored out, leaving the two-way interaction for analysis by permutation.

But what about a three-way factorial design? Not uncommon in biomedical experimentation. But this involves no fewer than three two-way interactions, and one three-way interaction. It is the last that might test the null hypothesis of prime interest. Can this interaction be extracted by permutation? Good [11] shows, by rather complex theoretical argument, how this could be done.

Then, what about **repeated-measures designs**? Not an uncommon design in biomedical experimentation. If the order of repeated measurements is randomized, Lunneborg shows how this can be handled by permutation [16, 17]. But, if the order of the repeated measurements is not randomized (for example, time cannot be randomized, nor can ascending dose- or stimulus-response designs), surely, analysis of the results cannot be done under the randomization model of inference?

My pragmatic view is that the more complex the experimental design, the less practicable is a randomization approach. Or, to put it another way, the more complex the experimental design, the closer tests under the classical and randomization models of inference approach each other.

Confidence Intervals (CIs) Under the Randomization Model. It seems to me that CIs are irrelevant to the randomization model. This is because they refer to populations that have been randomly sampled [20], and this is emphatically not the case under the randomization model.

Minimal Group Size and Power in Randomization Tests. Conventional ways of thinking about these have to be abandoned, because they depend on the population model of inference. My practical solution is to calculate the maximum number of possible permutations (formulae 2, 3). It must be at least 20 in order to be able to achieve $P \leq 0.05$.

Differences Between or Among Group Mean Ranks

As indicated above, the computational problem of evaluating these differences by permutation is much

less than that presented by differences between/among means. It is, therefore, rarely necessary to resort to Monte Carlo random sampling.

An enormous number of rank-order tests has been described: the **Wilcoxon-Mann-Whitney test** for two independent groups [19, 29], the Wilcoxon matched pairs, signed-rank, test [29], the **Kruskal-Wallis test** (see **Kruskal-Wallis Test**) on k independent groups [13], and the **Friedman test** on k related groups [4], are well known. But there is a host of other, eponymous, rank-order tests [14]. A word of caution is necessary. Almost all general statistics programs offer these tests, though executed asymptotically by normal or χ^2 approximations rather than by permutation. The problem with the asymptotic versions is that the algorithms used are often not described. How ties are handled is of critical importance. This matter has been addressed recently [1]. An example of the exact Wilcoxon-Mann-Whitney test is given in Table 2.

Exact Tests on Categorical Data

The simplest case is a 2×2 table of frequencies (Table 3). As Fisher described in simple terms [8], and Siegel and Castellan in modern notation [27], the point probability of H_0 is described by:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} \quad (4)$$

However, the probability of H_0 refers to the probability of occurrence of the observed values, plus the probabilities of *more extreme values in the same direction*. And, if a two-sided P is sought, *the same or more extreme values in either direction* must be taken into account. Thus, (4) must be applied to all these tables, and the P values summed. In the example of Table 3, the two-sided P is twice the one-sided value. This is only because of the equality and symmetry of the marginal totals.

There is a multitude of **exact tests** on categorical data for 2×2 tables of frequencies, and for unordered and ordered $r \times c$ tables. But even in the simplest case of 2×2 tables, there are problems about the null hypotheses being tested. It seems to me that the Fisher exact test, and the various formulations of the χ^2 test, are concerned with very vague null hypotheses (H_0). I prefer tests in which H_0 is much more specific: for

Table 3 Results of R.A. Fisher's thought experiment on the ability of a lady to distinguish whether milk or tea was added first to the cup [8]

		Actual design		Row
		Milk first	Tea first	Total
Lady's decisions	Milk first	3	1	4
	Tea first	1	3	4
	Column total	4	4	8

The lady was informed in advance that she would be presented with 8 cups of tea, in 4 of which the tea had been added first, in 4 the milk. The exact probability for the null hypothesis that the lady was unable to distinguish the order is obtained by the sum of the point probabilities (Formula 4) for rearrangements of the Table in which the observed, or more extreme, values would occur. Thus, two-sided $P = 0.22857 + 0.0142857 + 0.22857 + 0.0142857 = 0.48573$.

Table 4 Properties of exact tests on 2×2 tables of frequencies

Test	Null hypothesis	Conditioning
Fisher's	Vague	Both marginal totals fixed
Odds ratio	OR = 1	Column marginal totals fixed
Exact χ^2	Columns and rows independent	Unconditional
Difference between proportions	$p_1 - p_2 = 0$	Unconditional
Ratio between proportions	$p_1/p_2 = 1$	Unconditional

Conditioning means 'fixed in advance by the design of the experiment'. In real-life experiments, both marginal totals can never be fixed in advance. Column totals are usually fixed in advance if they represent treatment groups.

instance, that OR = 1, $p_1 - p_2 = 0$, or $p_1/p_2 = 1$ (Table 4).

There is a further problem with the Fisher exact test. It is a conditional test, in which the column and marginal totals in a 2×2 table are fixed in advance. This was so in Fisher's thought experiment (Table 3), but it is almost inconceivable that this could be achieved in a real-life experiment. In the case that H_0 is that OR = 1, there is no difficulty if

one regards the column totals as group sizes. Tests on proportions, for instance, that $p_1 - p_2 = 0$, or $p_1/p_2 = 1$, are unconditional. But there are complex theoretical and computational difficulties with these.

Acknowledgements

I am most grateful to colleagues who have read and commented on my manuscript, especially my Section Editor, Cliff Lunneborg.

Appendix: Software Packages for Permutation Tests

I list only those programs that are commercially available. I give first place to those that operate under the Microsoft Windows system. There is an excellent recent and comparative review of several of these [23, 24]. The remaining programs operate under DOS. I shall mention only one of the latter, which I have used. The remaining DOS programs require the user to undertake programming, and I do not list these. All the commercial programs have their own datafile systems. I have found dfPowerDBMS/Copy v. 8 (Dataflux Corporation, Cary NC) invaluable in converting any spreadsheet into the appropriate format for almost any statistics program.

Microsoft Windows

StatXact v.6 with Cytel Studio (Cytel Software Corporation, Cambridge MA). StatXact is menu-driven. For differences between/among group means, it caters for two independent groups, matched pairs, and the equivalent of one-way ANOVA. It does not have a routine for two-way ANOVA. For differences between/among group mean ranks, it caters for the Wilcoxon-Mann-Whitney test, the Wilcoxon signed-rank test for matched pairs, the Kruskal-Wallis test for k independent group mean ranks, and the Friedman test for k matched groups. For categorical data, it provides a great number of tests. These include the Fisher exact test, exact χ^2 , a test on $OR = 1$, tests on differences and ratios between proportions; and for larger and more complex tables of frequencies, the Cochran-Armitage test on ordered categorical data.

LogXact (Cytel Software Corporation, Cambridge MA). This deals exclusively with exact logistic

regression analysis for small samples. However, it does not cater for stepwise logistic regression analysis.

SAS v. 8.2 (SAS Institute Inc, Cary NC). SAS has introduced modules for exact tests, developed by the Cytel Software Corporation. These include PROC FREQ, PROC MULTTEST, and PROC NPAR1WAY, PROC UNIVARIATE, and PROC RANK. NPAR1WAY provides permutation tests on the means of two or more independent groups, but not on more than two related groups. PROC RANK caters for a variety of tests on mean ranks, and PROC FREQ a large number of exact tests on tables of frequencies.

Testimate v. 6 (Institute for Data Analysis and Study Planning, Gauting/Munich). This caters for a variety of exact rank-order tests and tests on tables of frequency, but no tests on means.

SPSS (SPSS Inc, Chicago IL). This very popular statistics package has an Exact Tests add-on (leased from StatXact), with routines for a wide range of exact tests on categorical data, some on rank-ordered data, but none on differences between/among means.

DOS Programs

RT v. 2.1 (West Inc., Cheyenne WY). This is Bryan Manly's program, based on his book [18]. One important attribute is that it provides for two-way ANOVA carried out by permutation, in which the interaction can be extracted. However, it has not been developed since 1991, though Bryan Manly hopes that someone will take on the task of translating it onto a Windows platform (personal communication).

References

- [1] Bergmann, R., Ludbrook, J. & Spooren, W.P.M.J. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages, *The American Statistician* **54**, 72–77.
- [2] Boik, R.J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous, *British Journal of Mathematical and Statistical Psychology* **40**, 26–42.

- [3] Eden, T. & Yates, F. (1933). On the validity of Fisher's z test when applied to an actual example of nonnormal data, *Journal of Agricultural Science* **23**, 6–16.
- [4] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.
- [5] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London.
- [6] Fisher, R.A. (1925). Applications of "Student's" distribution, *Metron* **5**, 90–104.
- [7] Fisher, R.A. (1936). The coefficient of racial likeness and the future of craniometry, *Journal of the Royal Anthropological Institute* **66**, 57–63.
- [8] Fisher, R.A. (1971). *The Design of Experiments*, in *Statistical Methods, Experimental Design and Scientific Inference*, 8th Edition, (1st Edition 1935), (1990) J.H. Bennett, ed., Oxford University Press, Oxford.
- [9] Fisher, R.A. & Mackenzie, W.A. (1923). Studies in crop variation. II: the manurial response of different potato varieties, *Journal of Agricultural Science* **13**, 311–320.
- [10] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.
- [11] Good, P.I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd Edition, Springer-Verlag, New York.
- [12] Kempthorne, O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.
- [13] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.
- [14] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, 1st Edition, revised, Prentice-Hall, Upper Saddle River.
- [15] Ludbrook, J. & Dudley, H.A.F. (1998). Why permutation tests are superior to t - and F -tests in biomedical research, *The American Statistician* **52**, 127–132.
- [16] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Lodge.
- [17] Lunneborg, C.E. (2002). Randomized treatment sequence designs: the randomization test as a nonparametric replacement of ANOVA and MANOVA, *Multiple Linear Regression Viewpoints* **28**, 1–9.
- [18] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.
- [19] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.
- [20] Neyman, J. (1941). Fiducial argument and the theory of confidence intervals, *Biometrika* **32**, 128–150.
- [21] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika* **20A**, 175–240.
- [22] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II, *Biometrika* **20A**, 263–294.
- [23] Oster, R.A. (2002a). An examination of statistical software packages for categorical data analysis using exact methods, *The American Statistician* **56**, 235–246.
- [24] Oster, R.A. (2002b). An examination of statistical software packages for categorical data analysis using exact methods – Part II, *The American Statistician* **57**, 201–213.
- [25] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, *Journal of the Royal Statistical Society B* **4**, 119–130.
- [26] Rubin, D.B. (1991). Practical applications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213–1234.
- [27] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw-Hill, New York.
- [28] "Student". (1908). The probable error of a mean, *Biometrika* **6**, 1–25.
- [29] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.
- [30] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Supplement Journal of the Royal Statistical Society* **1**, 217–235.

JOHN LUDBROOK

Randomized Block Design: Nonparametric Analyses

LI LIU AND VANCE W. BERGER

Volume 4, pp. 1681–1686

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Randomized Block Design: Nonparametric Analyses

In a **randomized block design**, there are, in addition to the experimental factor or factors of interest, one or more nuisance factors (*see Nuisance Variables*) influencing the measured responses. We are not interested in evaluating the contributions of these nuisance factors. For example, gender may be relevant in studying smoking cessation, but, in a comparative evaluation of a particular set of smoking cessation techniques, we would not be concerned with assessing the influence of gender per se. However, we would be interested in controlling the gender influence. One can control such nuisance factors by a useful technique called *blocking* (*see Block Random Assignment*), which can reduce or eliminate the contribution these nuisance factors make to the experimental error. The basic idea is to create homogeneous blocks or **strata** in which the levels of the nuisance factors are held constant, while the levels of the experimental factors are allowed to vary. Each estimate of the experimental factor effect within a block is more efficient than estimates across all the samples. When one pools these more efficient estimates across blocks, one should obtain a more efficient estimate than would have been available without blocking [3, 9, 13].

One way to analyze the randomized block design is to use standard parametric **analysis of variance (ANOVA)** methods. However, these methods require the assumption that the experimental errors are normally distributed. If the errors are not normally distributed, or if there are outliers in the data, then parametric analyses may not be valid [2]. In this article, we will present a few distribution-free tests, which do not require the normality assumption. These nonparametric analyses include the Wilcoxon signed rank test (*see Signed Ranks Test*), Friedman's test, the aligned rank test, and Durbin's test.

The Sign Test for the Randomized Complete Block Design with Two Treatments

Consider the boys' shoe-wear data in Table 1, which is from [3]. Two sole materials, A and B, were

Table 1 Boys' shoe-wear. Example: the amount of wear measured for two different materials A and B

Boy	Material A	Material B
1	13.2 (L) ^a	14.0 (R) ^b
2	8.2 (L)	8.8 (R)
3	10.9 (R)	11.2 (L)
4	14.3 (L)	14.2 (R)
5	10.7 (R)	11.8 (L)
6	6.6 (L)	6.4 (R)
7	9.5 (L)	9.8 (R)
8	10.8 (L)	11.3 (R)
9	8.8 (R)	9.3 (L)
10	13.3 (L)	13.6 (R)

^aleft sole; ^bright sole.

randomly assigned to the left sole or right sole for each boy. Each boy is a block, and has one A-soled shoe and one B-soled shoe. The goal was to determine whether or not there was any difference in wearing quality between the two materials, A and B.

The **sign test** (*see Binomial Distribution: Estimating and Testing Parameters*) is a nonparametric test to compare the two treatments in such designs. It uses the signs of the paired differences, $(B - A)$, to construct the test statistic [12]. To perform the sign test, we count the number of positive paired differences, P_+ . Under the null hypothesis of no treatment difference, P_+ has a binomial distribution with parameters $n = 10$ and $p = 0.5$, where n is the number of blocks with nonzero differences. If the sample size is large, then one might use the normal approximation. That is,

$$Z = \frac{P_+ - n/2}{\sqrt{n/4}} \quad (1)$$

is approximately distributed as the standard normal distribution. For the small sample data in Table 1, we have $P_+ = 8$, and the exact binomial P value is 0.109.

The Wilcoxon Signed Rank Test for the Randomized Complete Block Design with Two Treatments

The sign test uses only the signs of the paired differences, but ignores their magnitudes. A more powerful test, called the *Wilcoxon signed rank test*, uses both the signs and the magnitudes of the differences. This signed rank test can also be used to

2 Randomized Block Design: Nonparametric Analyses

analyze the paired comparison designs. The following steps show how to construct the Wilcoxon signed rank test [12]. First, compute the paired differences and drop all the pairs whose paired differences are zero. Second, rank the absolute values of the paired differences across the remaining blocks (pairs) (*see Rank Based Inference*). Third, assign to the resulting ranks the sign of the differences whose absolute value yielded that rank. If there is a tie among the ranks, then use the mid ranks. Fourth, compute the sum of the ranks with positive signs T_+ and the sum of the ranks with negative signs T_- . $T = \min(T_+, T_-)$ is the test statistic. Reject the null hypothesis of no difference if T is small.

If the sample size is small, then the exact distribution of the test statistic should be used. Tables of critical values for T appear in many textbooks, and can be computed in most statistical software packages. If the sample size n (the number of pairs with nonzero differences) is large, then the quantity

$$Z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \quad (2)$$

is approximately distributed as the standard normal. For the boys' shoe-wear data in Table 1, the signed ranks are listed in Table 2. We find that $T = 3$, $Z = -2.5$, and the approximate P value, based on the normal approximation, is 0.0125. Since the sample size is not especially large, we also compute the P value based on the exact distribution, which is 0.0078.

The small P value suggests that the two sole materials were different. Compared to the sign test P value of 0.109, the Wilcoxon signed rank test

Table 2 Signed ranks for Boys' shoe-wear data

Boy	Material A	Material B	Difference (B – A)	Signed
1	13.2 (L) ^a	14.0 (R) ^b	0.8	9
2	8.2 (L)	8.8 (R)	0.6	8
3	10.9 (R)	11.2 (L)	0.3	4
4	14.3 (L)	14.2 (R)	–0.1	–1
5	10.7 (R)	11.8 (L)	1.1	10
6	6.6 (L)	6.4 (R)	–0.2	–2
7	9.5 (L)	9.8 (R)	0.3	4
8	10.8 (L)	11.3 (R)	0.5	6.5
9	8.8 (R)	9.3 (L)	0.5	6.5
10	13.3 (L)	13.6 (R)	0.3	4

^aleft sole; ^bright sole.

P value is much smaller. This is not surprising, since the Wilcoxon signed rank test considers both the signs and the ranks of the differences, and hence it is more powerful (*see Power*).

Friedman's Test for the Randomized Complete Block Design

Friedman's test [5] is a nonparametric method for analyzing randomized complete block designs, and it is based on the ranks of the observations within each block. Consider the example in Table 3, which is an experiment designed to study the effect of four drugs [8]. In this experiment, there are five litters of mice, and four mice from each litter were chosen and randomly assigned to four drugs (A, B, C, and D). The lymphocyte counts (in thousands per cubic millimeter) were measured for each mouse. Here, each litter was a block, and all four drugs were assigned once and only once per block. Since all four treatments (drugs) are compared in each block (litter), this is a randomized complete block design.

To compute the Friedman test statistic, we first rank the observations from different treatments within each block. Assign mid ranks in case of ties. Let b denotes the number of blocks, t denotes the number of treatments, y_{ij} ($i = 1, \dots, b$, $j = 1, \dots, t$) denotes the observation in the i th block and j th treatment, and R_{ij} denotes the rank of y_{ij} within each block. The Friedman test statistic is based on the sum of squared differences of the average rank for each treatment from the overall average rank. That is,

$$F_R = \sum_{j=1}^t \frac{(\bar{R}_j - \bar{R})^2}{\sigma_{\bar{R}_j}^2} \quad (3)$$

where $\bar{R}_j$ is the average rank for treatment j , $\bar{R}$ is the overall average rank, and the denominator is the

Table 3 Lymphocyte counts

Litter	Drugs			
	A	B	C	D
1	7.1	6.7	7.1	6.7
2	6.1	5.1	5.8	5.4
3	6.9	5.9	6.2	5.7
4	5.6	5.1	5.0	5.2
5	6.4	5.8	6.2	5.3

variance of the first term in the numerator. Note that this variance does not depend on the treatment group, but retains the subscript because it is the variance of a quantity indexed by this subscript. Under the null hypothesis, the Friedman test statistic has a chi-square distribution with $(t - 1)$ degrees of freedom. We reject the null hypothesis that there is no difference between treatments if F_R is large.

If there are no ties within blocks, then we have

$$\sigma_{\bar{R}_j}^2 = \frac{t(t+1)}{12b}, \quad (4)$$

and Friedman's test statistic reduces to

$$F_R = \frac{12}{bt(t+1)} \sum_{j=1}^t R_{.j}^2 - 3b(t+1). \quad (5)$$

If there are ties within blocks, then we have

$$\sigma_{\bar{R}_j}^2 = \frac{t(t+1)}{12b} \times D, \quad (6)$$

where

$$D = 1 - \frac{\sum_{i=1}^b T_i}{bt(t^2 - 1)}, \quad T_i = \sum_k (S_k^3 - S_k), \quad (7)$$

k ranges over the number of ties in the i th block, and S_k is the number of observations at the k th tied value.

To compute Friedman's test statistic for the data in Table 3, we first compute the ranks within each block. These are listed in Table 4.

We have

$$\begin{aligned} \sum_{j=1}^t (\bar{R}_j - \bar{R})^2 &= (19.5/5 - 2.5)^2 \\ &+ (8.5/5 - 2.5)^2 + (13.5/5 - 2.5)^2 \\ &+ (8.5/5 - 2.5)^2 = 3.28, \\ D &= 1 - \frac{\sum_{i=1}^b T_i}{bt(t^2 - 1)} = 1 - \frac{12}{5 \times 4 \times (4^2 - 1)} \\ &= 0.96, \\ \sigma_{\bar{R}_j}^2 &= \frac{t(t+1)}{12b} \times D = \frac{4 \times (4+1)}{12 \times 5} \times 0.96 = 0.32. \end{aligned} \quad (8)$$

Table 4 The ranks of the Lymphocyte count data

Litter	Drugs			
	A	B	C	D
1	3.5	1.5	3.5	1.5
2	4	1	3	2
3	4	2	3	1
4	4	2	1	3
5	4	2	3	1
Rank sum	19.5	8.5	13.5	8.5

So $F_R = 3.28/0.32 = 10.25$. Asymptotically, F_R has as its null distribution that of the chi-squared random variable with $(t - 1)$ degrees of freedom. The P value for our obtained test statistic of 10.25, based on the chi-squared distribution with 3 df, is 0.0166, and indicates that the drug effects are different. The exact distribution (*see Exact Methods for Categorical Data*) of the Friedman test statistic can be used for small randomized complete block designs. Odeh et al. [10] provided the critical values of the Friedman test for up to six blocks and six treatments.

Aligned Rank Test for Randomized Complete Block Design

Friedman's test is based on the rankings within blocks, and it has relatively low power when the number of blocks or treatments is small. An alternative is to align the rankings. That is, we subtract from the observation in each block some estimate of the location of the block to make the blocks more comparable. The location-subtracted observations are called *aligned observations*. We rank all the aligned observations from all the blocks instead of ranking them only within each block, and we use the aligned ranks to compute the test statistic. This is called the *aligned rank test* [6]. The aligned rank test statistic is the same as Friedman's test statistic except that the aligned rank test computes the ranks differently.

For the example in Table 3, the aligned rank test statistic is 10.53. We again evaluate this against the chi-squared distribution with $(t - 1) = 3$ df and obtain a P value of 0.0146, slightly smaller than that associated with Friedman's test. For this data, the difference between the aligned rank test and Friedman's test is not very pronounced, but, in some

cases, the more powerful aligned rank test can have a much smaller P value than the Friedman's test.

Durbin's Test for Balanced Incomplete Blocks Design

In a **balanced incomplete block design**, the block size k is smaller than the number of treatments t because it is impractical or impossible to form homogeneous blocks of subjects as large as t . As a result, not all of the treatments can be compared within a block, and this explains the term incomplete. The design is balanced, which means that each pair of treatments is compared in the same number of blocks as every other pair of treatments. And each block contains the same number of subjects and each treatment occurs the same number of times. In such a design, the appropriate subset of k treatments are randomized to the subjects within a particular block.

Table 5 is an example of a balanced incomplete block design [1]. In this study, the measurements are (percentage elongation -300) of specimens of rubber stressed at 400 psi. The blocks are 10 bales of rubber, and two specimens were taken from each bale. Each specimen was assigned to one of the five tests (treatments). We are interested in finding out whether there is any difference among the five tests.

Notice that each pair of treatments occurs together in exactly one bale. Durbin's test [4] can be used to test the treatment difference in a balanced incomplete block design. We first rank the observations within each block, and assign mid ranks for ties. Durbin's

test statistic is

$$D_R = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^t R_{.j}^2 - \frac{3r(t-1)(k+1)}{k-1}, \quad (9)$$

where $t = 5$ is the number of treatments, $k = 2$ is the number of subjects per block ($k < t$), $r = 4$ is the number of times each treatment occurs, and $R_{.j}$ is the sum of the ranks assigned to the j th treatment. Under the null hypothesis that there are no treatment differences, D_R is distributed approximately as chi-squared, with $(t-1)$ degrees of freedom. For the example in Table 5, we have

$$D_R = \frac{12 \times (5-1)}{4 \times 5 \times (2-1) \times (2+1)} \times (7^2 + 6^2 + 4^2 + 8^2 + 5^2) - \frac{3 \times 4 \times (5-1) \times (2+1)}{2-1} = 8 \quad (10)$$

with four degrees of freedom. The P value of 0.0916 indicates that the treatments are somewhat different. Durbin's test reduces to Friedman's test if the number of treatments is the same as the number of units per block.

Cochran–Mantel–Haenszel Row Mean Score Statistic

The Sign test, Friedman's test, the aligned rank test, and Durbin's test are all special cases of the Cochran–Mantel–Haenszel (CMH) (*see Mantel–Haenszel Methods*) row mean score statistic with the ranks as the scores (*see* [11]). Suppose that we have a set of $qs \times r$ contingency tables. Let n_{hij} be the number of subjects in the h th stratum in the i th group and j th response categories, $p_{hi+} = n_{hi+}/n_h$ and $p_{h+j} = n_{h+j}/n_h$. Let

$$\mathbf{n}'_h = (n_{h11}, n_{h12}, \dots, n_{h1r}, \dots, n_{hs1}, \dots, n_{hsr}),$$

$$\mathbf{P}'_{h*+} = (p_{h1+}, \dots, p_{hs+}),$$

$$\mathbf{P}'_{h+*} = (p_{h+1}, \dots, p_{h+r}). \quad (11)$$

Then

$$E(n_{hij}|H_0) = m_{hij} = n_h p_{hi+} p_{h+j},$$

$$E(\mathbf{n}_h|H_0) = \mathbf{m}_h = n_h[\mathbf{P}_{h+*} \otimes \mathbf{P}_{h*+}],$$

Table 5 Measurements (percentage elongation -300) of rubber stressed at 400 psi

Bale	Treatment				
	1	2	3	4	5
1	35	16			
2	20		10		
3	13			26	
4	25				21
5		16	5		
6		21		24	
7		27			16
8			20	37	
9			15		20
10				31	17

$$\text{Var}(\mathbf{n}_h|H_0) = \mathbf{V}_h = \frac{n_h^2}{(n_h - 1)} ((\mathbf{D}_{Ph+*} - \mathbf{P}_{h+*}\mathbf{P}_{h+*}') \otimes (\mathbf{D}_{Ph*+} - \mathbf{P}_{h*+}\mathbf{P}_{h*+}')) \quad (12)$$

where $\otimes$ denotes the left-hand Kronecker product, $\mathbf{D}_{Ph+*}$ and $\mathbf{D}_{Ph*+}$ are diagonal matrices with elements of the vectors $\mathbf{P}_{h+*}$ and $\mathbf{P}_{h*+}$ as the main diagonals. The general CMH statistic [7] is

$$\mathbf{Q}_{CMH} = \mathbf{G}'\mathbf{V}_G^{-1}\mathbf{G}, \quad (13)$$

where

$$\begin{aligned} \mathbf{G} &= \sum_h \mathbf{A}_h(\mathbf{n}_h - \mathbf{m}_h), \\ \mathbf{V}_G &= \sum_h \mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'. \end{aligned} \quad (14)$$

where $\mathbf{A}_h$ is a matrix, and different choices of $\mathbf{A}_h$ provide different statistics, such as the correlation statistic, the general association statistic, and so on. To perform the test based on the mean score statistic, we have $\mathbf{A}_h = \mathbf{a}_h' \otimes [\mathbf{I}_{(s-1)}, \mathbf{0}_{(s-1)}]$, where $\mathbf{a}_h = (a_{h1}, \dots, a_{hr})$ are the scores for the j th response level in the h th stratum [11]. If we use the ranks as the scores, and there is one subject per row and one subject per column in the contingency table of each stratum, then the CMH mean score statistic is identical to Friedman's test. The sign test, aligned rank test, and Durbin's test can also be computed using the CMH mean score statistic with the ranks as the scores.

References

- [1] Bennett, C.A. & Franklin, N.L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*, Wiley, New York.
- [2] Berger, V.W. (2000). Pros and cons of permutation tests, *Statistics in Medicine* **19**, 1319–1328.
- [3] Box, G., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, New York.
- [4] Durbin, J. (1951). Incomplete blocks in ranking experiments, *British Journal of Mathematical and Statistical Psychology* **4**, 85–90.
- [5] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of American Statistical Association* **32**, 675–701.
- [6] Hodges, J.L. & Lehmann, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance, *Annals of Mathematical Statistics* **33**, 482–497.
- [7] Landis, J.R., Heyman, E.R. & Koch, C.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.
- [8] Mead, R., Curnow, R.N. & Hasted, A.M. (1993). *Statistical Methods in Agriculture and Experimental Biology*, Chapman & Hall, London.
- [9] NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>.
- [10] Odeh, R.E., Owen, D.B., Birnbaum, Z.W. & Fisher, L.D. (1977). *Pocket Book of Statistical Tables*, Marcel Dekker, New York.
- [11] Stokes, M.E., Davis, C.S., & Koch, G., (1995). *Categorical Data Analysis Using the SAS System*, SAS Institute, Cary.
- [12] Wayne, W.D. (1978). *Applied Nonparametric Statistics*, Houghton-Mifflin, New York.
- [13] Wu, C.F.J. & Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley & Son, New York.

(See also **Distribution-free Inference, an Overview**)

LI LIU AND VANCE W. BERGER

Randomized Block Designs

SCOTT E. MAXWELL

Volume 4, pp. 1686–1687

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Randomized Block Designs

Sir Ronald A. Fisher's insight that random assignment to treatment groups probabilistically balances all possible confounders and thus allows for a causal inference is one of the most important statistical contributions of the twentieth century (*see* **Randomization**). This contribution undoubtedly explains the popularity of randomized designs in many disciplines, including the behavioral sciences. However, even though **nuisance variables** no longer need to be regarded as possible confounders with random assignment, these variables nevertheless continue to contribute to variance within groups in completely randomized designs. The magnitude of these effects is often substantial in the behavioral sciences, because for many phenomena of interest there are sizable systematic individual differences between individuals. The presence of such individual differences within groups serves to lower **power** and precision because variance due to such differences is attributed to error in completely randomized designs. In contrast, this variance is separately accounted for in the randomized block design (RBD) and thus does not inflate the variance of the error term, thus typically yielding greater power and precision in the RBD.

The basic idea of an RBD is first to form homogeneous blocks of individuals. Then individuals are randomly assigned to treatment groups within each block. Kirk [2] describes four types of possible dependencies: (a) repeated measures, (b) subject matching, (c) identical twins or littermates, and (d) pairs, triplets, and so forth, matched by mutual selection such as spouses, roommates, or business partners. The repeated measures (also known as *within-subjects*) application of the RBD (*see* **Repeated Measures Analysis of Variance**) is especially popular in psychology because it is typically much more efficient than a between-subjects design.

It is incumbent on the researcher to identify a blocking variable that is likely to correlate with the dependent variable of interest in the study (*see* **Matching**). For example, consider a study with 60 research participants comparing three methods of treating depression, where the Beck Depression

Inventory (BDI) will serve as the dependent variable at the end of the study. An ideal blocking variable in this situation would be each individual's BDI score at the beginning of the study prior to treatment. If these scores are available to the researcher, he/she can form 20 blocks of 3 individuals each: the first block would consist of the 3 individuals with the 3 highest BDI scores, the next block would consist of the 3 individuals with the 3 next highest scores, and so forth. Then the researcher would randomly assign 1 person within each block to each of the 3 treatment groups. The random assignment component of the RBD resembles that of the completely randomized design, but the random assignment is said to be restricted in the RBD because it occurs within levels of the blocking variable, in this case, baseline score on the BDI.

The RBD offers two important benefits relative to completely randomized designs. First, the reduction in error variance brought about by blocking typically produces a more powerful test of the treatment effect as well as more precise estimates of effects. In this respect, blocking is similar to **analysis of covariance** [3, 4]. Second, a potential problem in a **completely randomized design** is **covariate** imbalance [1], which occurs when groups differ substantially from one another on a nuisance variable despite random assignment. The RBD minimizes the risk of such imbalance for the blocking variable. It is also important to acknowledge that blocking a variable that turns out not to be related to the dependent variable comes at a cost, because degrees of freedom are lost in the RBD, thus requiring a larger critical value and thereby risking a loss instead of a gain in power and precision.

References

- [1] Altman, D.G. (1998). Adjustment for covariate imbalance, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester, pp. 1000–1005.
- [2] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.
- [3] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Lawrence Erlbaum Associates, Mahwah.

2 Randomized Block Designs

- [4] Maxwell, S.E., Delaney, H.D. & Dill, C.A. (1984). (See also **Block Random Assignment**)
Another look at ANCOVA versus blocking, *Psychological Bulletin* **95**, 136–147.

SCOTT E. MAXWELL

Randomized Response Technique

BRIAN S. EVERITT

Volume 4, pp. 1687–1687

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Randomized Response Technique

The randomized response technique is an approach that aims to get accurate answers to a sensitive question that respondents might be reluctant to answer truthfully, for example, ‘have you ever had an abortion?’ The randomized response technique protects the respondent’s anonymity by offering both the question of interest and an innocuous question that has a known probability (α) of yielding a ‘yes’ response, for example,

1. Flip a coin. Have you ever had an abortion?
2. Flip a coin. Did you get a head?

A random device is then used by the respondent to determine which question to answer. The outcome of the randomizing device is seen only by the respondent, not by the interviewer. Consequently when the interviewer records a ‘yes’ response, it will not be known whether this was a yes to the first or second question [2]. If the probability of the random device posing question one (p) is known, it is possible to estimate the proportion of yes responses to questions one (π), from the overall proportion of yes responses ($P = n_1/n$), where n is the total number

of yes responses in the sample size n .

$$\hat{\Pi} = \frac{P - (1 - p)\alpha}{p} \quad (1)$$

So, for example, if $P = 0.60$, $(360/600)p = 0.80$ and $\alpha = 0.5$, then $\hat{\Pi} = 0.125$. The estimated variance of $\hat{\Pi}$ is

$$\begin{aligned} \text{Var}(\hat{\Pi}) = & \frac{\hat{\Pi}(1 - \hat{\Pi})}{n} \\ & + \frac{(1 - p)^2\alpha(1 - \alpha) + p(1 - p)}{np^2} \quad (2) \end{aligned}$$

For the example here, this gives $\text{Var}(\hat{\Pi}) = 0.0004$.

Further examples of the application of the technique are given in [1].

References

- [1] Chaudhuri, A. & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.
- [2] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.

BRIAN S. EVERITT

Range

DAVID CLARK-CARTER

Volume 4, pp. 1687–1688

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Range

If the smallest score in a set of data is 10 and the largest is 300 then the range is $300 - 10 = 290$. The range has the advantage over just quoting the maximum and minimum values in that it is independent of the part of the scale where those scores occur. Thus, if the largest value were 400 and the smallest 110, then the range would still be 290.

A disadvantage of the range, compared to many other measures of spread, is that it is based solely on the numbers at the two ends of a set. Thus, if

a single value at one extremity of a set of data is widely separated from the rest of the set, then this will have a large effect on the range. For example, the range of the set 5, 7, 9, 11, 13 is 8, whereas if the largest value were 200 instead of 13, then the range would be 195. (Notice that the same range would be produced when an extreme score occurred at the other end of the set: 5, 193, 195, 197, 200.) Hence, the range is better quoted alongside other measures of spread that are less affected by extreme scores, such as the **interquartile range**, rather than on its own.

DAVID CLARK-CARTER

Rank Based Inference

T.P. HETTMANSPERGER, J.W. MCKEAN AND S.J. SHEATHER

Volume 4, pp. 1688–1691

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rank Based Inference

In the behavioral sciences, a researcher often does not wish to assume that the measurements in an experiment came from one or more normal populations. When the population distributions are unknown (but have the same shape), an alternative approach to hypothesis testing can be based on the ranks of the data. For example, data on reaction times are typically skewed and **outliers** or unusual results can exist in situations where results from a new experimental treatment are being compared with those from a well established procedure. When comparing treatment and control groups, the two samples are combined and ranked. Then, rather than use the traditional two-sample t statistic that compares the sample averages, we use the Mann–Whitney–Wilcoxon (MWW) (*see Wilcoxon–Mann–Whitney Test*) statistic that compares the average ranks in the two samples. One advantage is that the null distribution of the MWW statistic does not depend on the common but unspecified shape of the underlying populations. Another advantage is that the MWW statistic is more robust against outliers and gross errors than the t statistic. Finally, the MWW test is more efficient (have greater **power**) than the t Test when the tails of the populations are only slightly heavier than normal and almost as efficient (95.5%) when the populations are normal. Consider also that the data may come directly as ranks. A group of judges may be asked to rank several objects and the researcher may wish to test for differences among the objects. The Friedman rank statistic (*see Friedman’s Test*) is appropriate in this case.

Most of the simple experimental designs (one-sample, two-sample, one-way layout, two-way layout (with one observation per cell), correlation, and regression) have rank tests that serve as alternatives to the traditional tests that are based on least squares methods (sample means). For the designs listed above, alternatives include the Wilcoxon **signed rank test**, the MWW test, the **Kruskal–Wallis test**, the **Friedman test**, Spearman’s rank correlation (*see Spearman’s Rho*), and rank tests for regression coefficients (*see Nonparametric Regression*), respectively. There were just under 1500 citations to the above tests in social science papers as determined by the Social Science Citation Index (1956–2004). The

Kruskal–Wallis (KW) statistic is used for the one-way layout and can be thought of as an extension of the MWW test for two samples. As in the case of the MWW, the data is combined and ranked. Then the average ranks for the treatments are compared to the overall rank average which is $(N + 1)/2$ under the null hypothesis of equal populations, where N is the combined sample size. When the average ranks for the treatments diverge sufficiently from this overall average rank, the null hypothesis of equal treatment populations is rejected. This test corresponds directly to the one-way analysis of variance F Test.

The nonparametric test statistics in the one- and two-sample designs have corresponding estimates associated with them. In the case of the Wilcoxon signed rank statistic, the corresponding estimate of the center of the symmetric distribution (including the mean and median) is the median of the pairwise averages of the data values. This estimate, called the *one-sample Hodges–Lehmann estimate*, combines the robustness of the median with the efficiency of averaging data from a symmetric population. In the two-sample case, the Hodges–Lehmann estimate of the difference in the locations of the two populations is the median of the pairwise differences. In addition, the test statistics can be inverted to provide **confidence intervals** for the location of the symmetric population or for the difference of locations for two populations. The statistical package, Minitab, provides these estimates and confidence intervals along with the tests. Reference [3] is an excellent source for further reading on rank-based methods. Next we discuss estimation and testing in the linear model.

In a general linear model, the least squares approach entails minimizing a sum of squared residuals to produce estimates of the regression coefficients (*see Multiple Linear Regression*). The corresponding F Tests are based on the reduction in sum of squares when passing from the reduced to the full model, where the reduced model reflects the null hypothesis. Rank methods in the linear model follow this same strategy. We replace the least squares criterion by a criterion that is based on the ranks of the residuals. Then we proceed in the same way as a least squares analysis. Good robustness and efficiency properties of the MWW carry over directly to the estimates and tests in the linear model. For a detailed account of this approach, see [3] for an

applied perspective and [2] for the underlying theory with examples.

Computations are always an important issue. Many statistical packages contain rank tests for the simple designs mentioned above but not estimates. Minitab includes estimates and confidence intervals along with rank tests for the simple designs. Minitab also has an undocumented rank regression command `reg` that follows the same syntax as the regular regression command. The website <http://www.stat.wmich.edu/slab/RGLM/> developed by McKean provides a broad range of tests, estimates, standard errors, and data plots for the general linear model. See reference [1] for additional discussion and examples based on the website. Below, we will illustrate the use of the website.

First we sketch the rank-based approach to inference in the linear model, and then we will outline how this approach works in a simple **analysis of covariance**. For $i = 1, \dots, n$, let $e_i = y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta}$ be the i th residual or error term, where $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ are the known regression constants and $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$ are the unknown regression coefficients. The error distribution is assumed to be continuous and have median zero with no further assumptions. We denote by $F(e)$ and $f(e)$ the distribution and density functions of the errors, respectively. We call this the general linear model, since regression, **analysis of variance**, and analysis of covariance can be analyzed with this model. In matrix notation, we have $\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{1}$ is an $n \times 1$ vector of all ones, $\mathbf{X}$ is the $n \times p$ full rank design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, and $\mathbf{e}$ is the $n \times 1$ vector of errors.

In least squares, we minimize the criterion function $\sum e_i^2$ to find the estimates of the regression coefficients. This least squares criterion function is equivalent to $\sum \sum_{i < j} (e_i - e_j)^2$ for inference on $(\beta_1, \dots, \beta_p)$. For rank-based inference, we replace this criterion function by $\sum \sum_{i < j} |e_i - e_j|$. But $\sum \sum_{i < j} |e_i - e_j|$ is proportional to $D(\text{full}) = \sqrt{12} \sum (\text{Rank}(e_i) - (N + 1)/2) e_i$. Hence, rather than a quadratic function of the residuals, we have a linear function of the residuals with the weights determined by the ranks of the residuals. The rank estimate of $\boldsymbol{\beta}$ is found by minimizing $D(\text{full})$, which is a measure of the dispersion of the full model residuals. For testing a null hypothesis, we write $D(\text{reduced})$ for the

dispersion of the residuals with $\boldsymbol{\beta}$ constrained to the null hypothesis. Then $RD = \hat{D}(\text{reduced}) - \hat{D}(\text{full})$ is the reduction in dispersion as a result of fitting the reduced (null) model, where the hat indicates that D has been evaluated at the respective reduced and full model estimates of $\boldsymbol{\beta}$.

Throughout the inference, we need a scaling factor similar to σ^2 , the variance of the error distribution, in least squares. This factor is $\tau = (\sqrt{12} \int f^2(e) de)^{-1}$, where f is the density function of the error distribution. In the case of normal errors, τ is equal to $\sqrt{(\pi/3)}\sigma$. This scaling parameter can be estimated using a density estimate based on the full model residuals.

This leads to the following basic results for inference: the rank-based estimate $\hat{\boldsymbol{\beta}}$ is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\tau^2 \mathbf{X}^T \mathbf{X}$, where $\mathbf{X}$ is the matrix of regression constants. The estimate of β_0 is the median of the residuals, $y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, and it is approximately normally distributed with mean β_0 and variance $(n4f^2(0))^{-1}$ when the design matrix has been centered. Further, for testing $H_0 : M\boldsymbol{\beta} = \mathbf{0}$, where M is a $q \times p$ matrix of full row rank, $F_R = 2RD/q\hat{\tau}$ is approximately distributed as $F(q, n - p - 1)$. We now illustrate this approach on a simple analysis of covariance.

The website <http://www.stat.wmich.edu/slab/RGLM/> was used for computations for this example, and it can be used for a wide variety of models and designs. In this experiment, three advertising media (radio, newspaper, and television) were compared. The experimental units were 15 fast food restaurants located in comparable but different cities, five for each of the media. The response variable y was profits in thousands of dollars. The restaurants were roughly of the same size but had differing levels of food wastage. The percentage of food wastage x was used as a covariate. For example, there was 1.0% in the first restaurant under radio. The data is given in Table 1.

The website provides, along with the significance testing and estimation (with standard errors), data plots, **residual plots**, and standardized residual plots including **histograms**, **boxplots**, and Q-Q plots (*see Probability Plots*). We report here the results of testing for a covariate effect and for a media effect. We find $F_R = 2RD/q\hat{\tau} = 92.6$ with a P value = 0.0001 for the null hypothesis that all media are the same. For the null hypothesis that the coefficient of the covariate x is zero, the P value is effectively

Table 1 Profits in thousands of dollars(y) and Percent food wastage(x)

radio		newspaper		television	
x	y	x	y	x	y
1.0	30	2.1	24	3.4	17
1.4	18	2.6	20	3.9	11
1.9	13	3.1	7	4.3	3
2.5	6	3.6	4	4.7	-6
2.7	3	4.1	-5	5.2	-10

zero. Removing the covariate effect and considering data plots shows that the profits decline from radio to television. Least squares results in a very similar analysis. However, if $y = 6$ in the fourth row of radio is entered incorrectly as $y = 60$, the analysis is essentially the same for the rank tests but the least squares test is no longer significant for either of the null hypotheses, illustrating the relative robustness of the rank tests. In the original data, the coefficient of determination based on the rank approach is 0.90 while it is 0.97 for least squares. This coefficient is often used in data analysis to assess the quality of the fitted model. In the following paragraphs, we discuss robust coefficients of determination.

We now consider the correlation model in which $\mathbf{x}$ is a p -dimensional random vector with distribution function $M(\mathbf{x})$ and density function $m(\mathbf{x})$. We also let $H(\mathbf{x}, y)$ and $h(\mathbf{x}, y)$ denote the joint distribution and density functions of $(\mathbf{x}, y)$, respectively. Then $h(\mathbf{x}, y) = f(y - \beta_0 - \mathbf{x}^T \boldsymbol{\beta})m(\mathbf{x})$. We are interested in a robust measure of the relationship between y and $\mathbf{x}$; that is, a robust measure of association between y and $\mathbf{x}$. As with the traditional measure of determination, our robust measure will be zero if and only if y and $\mathbf{x}$ are independent. Hence, independence becomes the null hypothesis and this translates into $\beta = \mathbf{0}$ so that $h(\mathbf{x}, y) = f(y - \beta_0)m(\mathbf{x})$.

We consider the traditional measure first. Assume without loss of generality that $E(\mathbf{x}) = \mathbf{0}$ and $E(e) = 0$. Let $\text{Var}(e) = \sigma_e^2$ and let $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ denote the variance-covariance matrix of $\mathbf{x}$. Then the traditional coefficient of determination is given by

$$\bar{R}^2 = \frac{\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}{\sigma_e^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}. \quad (1)$$

Note that $\bar{R}^2$ is a measure of association between y and $\mathbf{x}$. It lies between 0 and 1, and it is 0 if and

only if y and $\mathbf{x}$ are independent, since y and $\mathbf{x}$ are independent if and only if $\boldsymbol{\beta} = \mathbf{0}$.

Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ be a random sample from the above correlation model. In order to obtain a consistent estimate of $\bar{R}^2$, treat $\mathbf{x}_i$ as fixed and estimate the parameters by least squares in the full model and then again in the reduced model in which $\boldsymbol{\beta} = \mathbf{0}$. The reduction in sum of squares is $SSR = SST - SSE$, and a consistent estimate of $\bar{R}^2$ is the familiar $R^2 = SSR/SST$. The traditional F statistic is $F = (SSR/p)/MSE$ where $MSE = SSE/(n - p - 1)$. Finally, recall that R^2 can be reexpressed as

$$R^2 = \frac{SSR}{SSR + (n - p - 1)SSE} = \frac{\frac{p}{n-p-1}F}{1 + \frac{p}{n-p-1}F}. \quad (2)$$

We first introduce the robust estimate and then discuss the measure that it estimates (see **Robust Testing Procedures**). The rank test described above for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is $F_R = 2RD/p\hat{\tau}$, and RD is the reduction in dispersion when passing from the reduced to the full model. It is analogous to the reduction in sum of squares in the traditional approach based on least squares. We simply replace F by F_R in the expression above to get

$$\begin{aligned} R_{\text{rank}} &= \frac{\frac{p}{n-p-1}F_R}{1 + \frac{p}{n-p-1}F_R} \\ &= \frac{RD}{RD + (n - p - 1)(\hat{\tau}/2)}. \end{aligned}$$

The statistic R_{rank} is robust because it is a one-to-one function of the robust test statistic F_R . Further, it lies between 0 and 1, having the values 1 for a perfect fit and 0 for a complete lack of fit. These properties make R_{rank} an attractive coefficient of determination for regression problems. Note that R_{rank} is a ratio of scales, while R^2 is a ratio of variances. In general, R_{rank} estimates a different parameter than R^2 . Thus caution is needed in comparing the two statistics. However, like R^2 , R_{rank} can be used for comparison of hierarchical models.

The statistic R_{rank} is a consistent estimate for $\bar{R}_{\text{rank}} = \overline{RD}/(\overline{RD} + \tau/2)$, where

$$\begin{aligned} \overline{RD} &= \int \sqrt{12}(G(y) - \tfrac{1}{2})yg(y)dy \\ &\quad - \int \sqrt{12}(F(y) - \tfrac{1}{2})yf(y)dy, \end{aligned} \quad (3)$$

4 Rank Based Inference

and $G(y)$ and $g(y)$ are the marginal distribution and density functions of y , respectively. This measure is between 0 and 1 and is 0 if and only if y and $\mathbf{x}$ are independent. This is the robust coefficient of determination. In general, $\bar{R}_{\text{rank}}$ and $\bar{R}^2$ differ, they are one-to-one functions of each other when $(\mathbf{x}, y)$ has a multivariate normal distribution. Define

$$\bar{R}_{\text{rank}}^* = 1 - \left[\frac{1 - \bar{R}_{\text{rank}}}{1 - \bar{R}_{\text{rank}}(1 - \pi/6)} \right]^2. \quad (4)$$

Then, under multivariate normality, $\bar{R}_{\text{rank}}^* = \bar{R}^2$. The corresponding statistic is R_{rank}^* which is defined in terms of R_{rank} . Now, under multivariate normality (see **Catalogue of Probability Density Functions**), R_{rank}^* and R^2 estimate the same quantity.

References

- [1] Abebe, A., Crimin, K., McKean, J.W., Haas, J. & Vidmar, T. (2001). Rank-based procedures for linear models: applications to pharmaceutical science data, *Drug Information Journal* **35**, 947–971.
- [2] Hettmansperger, T.P. & McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*, Arnold Publishing, London.
- [3] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, John Wiley, New York.

T.P. HETTMANSPERGER, J.W. MCKEAN AND
S.J. SHEATHER

Rasch Modeling

GERHARD H. FISCHER

Volume 4, pp. 1691–1698

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rasch Modeling

History of the Rasch Model

In 1952, the Danish mathematician Georg Rasch (1901–1980), then consultant for a project of the Ministry of Social Affairs, introduced a multiplicative Poisson model for the analysis of reading errors of school children. He considered the number of errors made by testee S_v in text I_i as a realization of a Poisson variable with parameter λ_{vi} (see **Catalogue of Probability Density Functions**) measuring the testee's 'proneness' to errors when reading that particular text. He then split λ_{vi} into a factor pertaining to the testee, S_v 's reading ability θ_v , and a factor pertaining to the text, the difficulty δ_i of text I_i . To him as a mathematician it was immediate, in virtue of a well-known theorem about Poisson variables, to draw the following conclusion: if a testee had read two texts I_i and I_j , the probability of observing k_{vi} and k_{vj} errors in these two texts, conditional on the total sum of $k_{vi} + k_{vj} = k_v$ errors, had to follow a Binomial distribution characterized by k_v and parameter $\pi = 1/(1 + \delta_j/\delta_i)$ (see **Catalogue of Probability Density Functions**).

This opened Rasch's eyes for a novel approach to measurement in behavioral science: since parameter π no longer depended on the testee parameter θ_v , data from all testees who had read the same two texts could be pooled to obtain a (**maximum likelihood**) estimate (see **Maximum Likelihood Estimation**) of π , from which in turn an estimate of the quotient of the text difficulties, δ_i/δ_j , was obtainable. This enabled a measurement of the *relative* difficulties of the texts – in a specific sense – independently of the sample of testees, and this procedure could of course be extended to the simultaneous comparison of more than two texts. He considered such comparisons of text difficulties as *objective* because the result was generalizable over particular children tested, whatever their individual reading abilities might have been. Later Rasch [19] denoted functions that made such a comparison feasible *comparators*, and comparisons carried out in that manner *specifically objective*.

Instrumental for this was the splitting of parameter λ into a product of a testee's ability and the text's difficulty. When Rasch in 1953 analyzed intelligence test data for the Danish army, he decided to carry the same principle over to the area of test analysis. First,

he sought for a suitable *item response function* (IRF) $P(+|S_v, I_i) = f(\xi_v, \epsilon_i)$, where '+' denoted a correct response, ξ_v the testee's ability, and ϵ_i the easiness of test item I_i ; as a particularly simple algebraic function mapping the positive reals on the semiopen interval $[0, 1)$, he chose $f = x/(1 + x)$. Then he conceived x to be a product of the testee's ability ξ_v and the item's easiness ϵ_i , namely, $x = \xi_v \epsilon_i$, with $\xi_v \geq 0$ and $\epsilon_i \geq 0$. This model is now generally denoted the 'Rasch Model' (RM), but is usually reparameterized by taking the logarithms rather than Rasch's original item and person parameters: $\theta_v = \ln(\xi_v)$ as testee *ability* and $\beta_i = -\ln(\epsilon_i)$ as item *difficulty*.

Definition and Some Basic Properties of the RM

The RM for dichotomous responses (denoted '+' and '-') is defined as

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \quad (1)$$

where X_{vi} is the response variable with realization $x_{vi} = 1$ if testee S_v gives response '+' to I_i , and $x_{vi} = 0$ if S_v 's response to I_i is '-'; $-\infty < \theta_v < \infty$ is the latent ability of S_v , and $-\infty < \beta_i < \infty$ the difficulty of item I_i . All item responses are assumed to be 'locally independent', that is, the probability of any response pattern in a test of length k is the product of k probabilities (1) or complements thereof. Parameters θ_v and β_i are defined only up to an arbitrary additive normalization constant c ; the latter is usually specified by either setting one item parameter to zero (e.g., $\beta_1 = 0$), or setting $\sum_i \beta_i = 0$.

The form of (1) implies that all IRFs are parallel curves. This means that, if the RM is to fit a set of data, all items must have (approximately) *equal discrimination*.

The probability of a complete $(n \times k)$ item score matrix X as a function of the unknown parameters (i.e., the so-called *likelihood function*) can be shown to depend only on the marginal sums of X , namely, on the *raw scores* $r_v = \sum_i x_{vi}$ and the item marginals $x_{.i} = \sum_v x_{vi}$. This implies that the r_{vi} and the $x_{.i}$ contain all the relevant information in the data with respect to the parameters (i.e., are jointly *sufficient statistics*-see **Estimation**), whereas the individual response patterns yield no additional information about the parameters. This is a remarkable asset of the

RM: throughout a century of psychological testing, the number-correct (or raw) score in intelligence and other attainment tests has been employed as a summary of a testee's test achievement; if the RM holds for a particular test, this fact yields a rigorous justification for the use of the raw scores. Therefore, ascertaining whether the RM fits a given set of test data is an enterprise of considerable practical importance.

Another noteworthy property of the RM is that the conditional probability (or likelihood) of an item score matrix X , given all raw scores r_v , is a function of the item parameters only. (This parallels Rasch's earlier observation about the conditional distribution of the number of reading errors in two texts, given the testee's total number of reading errors.) The conditional probability can thus serve as a comparator for the item parameters, enabling specifically objective comparisons of the item difficulties. (Symmetrically, specifically objective comparisons between persons are also possible, but carrying them out directly is impractical under most realistic conditions. The way to compare persons is to first estimate all item parameters and then to estimate the person parameters by considering the item parameters as given constants.)

These two favorable properties of the RM – sufficiency of the raw scores and of the item marginals, specifically objective comparisons between items and between testees – raises the question of whether other models exist that share these properties with the RM. Within a framework of locally independent items with continuous, strictly monotone IRFs with lower limits zero (i.e., no guessing) and upper limits one, the answer is 'no'. It can be shown that, within this framework, a *family* of RMs follows from either of the following assumptions: (a) sufficiency of the raw score for the person parameter; (b) existence of a nontrivial likelihood function that can serve as a specifically objective comparator for the items or the testees; (c) existence of a nontrivial sufficient statistic for the testee parameter that is independent of the item parameters; and (d) stochastically consistent ordering of the items. (On formal definitions and proofs, see Chapter 2 of [7].) The term 'family of RMs' refers to all models of the form (1) where, however, the parameters θ_v and β_i are replaced by $a\theta_v + c$ and $a\beta_i + c$, respectively. Therein, the constant c , which immediately cancels from (1), is the normalization constant mentioned above, and $a > 0$

is an unspecified *discrimination* parameter, namely, the maximal slope of the (parallel) IRFs. One may set $\alpha = 1$, of course, which immediately yields the RM, but it has to be kept in mind that this specification is arbitrary.

These results have important consequences regarding an age-old question in behavioral science: What are the measurement properties of psychological or educational tests? If, for a given set of data X , the RM is found to fit and the parameters are adequately estimated, the scales of the parameters θ and β are unique only up to linear transformations $a\theta + c$ and $a\beta + c$, with arbitrary $a > 0$ and arbitrary normalization constant c . From this it is concluded that the scales are *interval scales* (see **Scales of Measurement**) with a common measurement unit.

There is one more *caveat*, though: The proofs leading to these conclusions rest on the assumption that there are a testee population and a *dense* universe of items such that both the person and the item parameters can vary continually. While the first assumption appears to be acceptable – for instance, growth of ability is usually conceived as continuous – the second assumption may be questioned. In the special case, however, that the item parameters are assumed to be k fixed discrete *rational* numbers, the conclusion about the measurement properties of the RM becomes somewhat weaker: abilities are measurable only on an *ordered metric scale* which has interval scale properties at certain lattice points (which are typically spaced narrowly), but elsewhere allows only an ordering of abilities. For practical purposes, however, the scale can still be considered an interval scale.

Parameter Estimation and Testing of Fit

Various approximate methods for parameter estimation have been proposed by Rasch and some of his students, but under the perspective that estimators should be *unique*, *statistically consistent*, and should have known *asymptotic properties*, only two approaches seem to prevail. The probably most attractive method is the *conditional maximum likelihood* (CML) method which maximizes the conditional probability of X , given the raw scores r_v , in terms of the item parameters. As mentioned above, this likelihood function is independent of the person parameters and thus is a comparator function – in the

spirit of Rasch – for establishing specifically objective comparisons of the items. In finite samples, the normalized CML estimates are unique if a certain directed graph C associated with matrix X is *strongly connected*, see [7]; checking this weak connectivity condition is easy in practice, using standard tools of graph theory. If a weak necessary and sufficient condition given by Pfanzagl [16] is met by the distribution of the person parameters in the respective testee population, the estimators are consistent and asymptotically normally distributed around the true parameters β_i for fixed test length k and $n \rightarrow \infty$. The standard errors of the estimates $\hat{\beta}_i$ can be determined from the **information matrix**. The estimator is not *efficient*, though, but the loss of statistical information entailed by conditioning on the raw scores is very slight, see [5]. Programs for CML estimation are, for example, LPCM-Win [8] and OPLM [21].

Another approach to the estimation of the β_i is the *marginal maximum likelihood* (MML) method. In the so-called *parametric* MML, a latent population distribution of the θ -parameters is specified, for instance, a normal distribution with unknown mean μ and standard deviation σ . To calculate the likelihood of the data under such an enhanced RM requires integration over θ , leading to the elimination of the person parameters from the likelihood function. The latter is then maximized with respect to both the item and distribution parameters. If the assumption about the latent population distribution happens to be true, then the parametric MML method is consistent and asymptotically efficient. If, on the other hand, the distributional assumption is not true, then the MML method can be strongly biased and even loses the property of consistency. A popular program for parametric MML estimation in the RM (and also for more general item response models like the two- and three-parameter logistic models) is BILOG [14].

Yet another method is *nonparametric* MML where the latent distribution is replaced by a step function with at most $(k+2)/2$ (if k is even) or $(k+1)/2$ (if k is odd) nodes on the θ -axis. The respective areas under the step function are then estimated along with the item parameters. De Leeuw and Verhelst [4] have shown that with the RM this method is asymptotically equivalent to CML. Pfanzagl [16] has proved that the RM is the only model where, under the assumption of an unknown latent distribution of θ , the item parameters are identifiable via nonparametric MML. (This is one more argument for the RM.)

From a practical point of view, the two most promising candidates among the estimation methods seem to be CML and parametric MML. (Nonparametric MML is asymptotically equivalent to CML and thus needs not be considered separately.) The choice should depend on the researcher's faith in a particular latent population distribution. Specifying the latent distribution, however, is always problematic: it has to be kept in mind that assuming, for instance, a normal distribution for a given population makes it practically impossible that the same holds for subpopulations like the male and female testees as well. Choosing CML implies a slight loss of precision of the estimates but, on the other hand, circumvents distributional assumptions that may entail a serious bias of the estimators.

Closely related to the question of item parameter estimation is the problem of testing of fit: only if the statistical consistency and asymptotic distribution (i.e., normality) of the estimators have been verified, it becomes possible to establish powerful test methods for assessing fit and/or testing other hypotheses on the parameters. Rasch [18] had mainly used heuristic graphical methods for the assessment of fit by checking whether the most characteristic property of the RM – independence of the item parameter estimates of the testees' abilities – could be empirically verified. Andersen [2] has shown that this null hypothesis can be tested by means of a *conditional likelihood ratio* (CLR) test: the sample is split in two or more subsamples which differ significantly with respect to their raw score distributions (e.g., in testees with above versus below average raw scores), and the item parameter estimates in these subgroups are compared with those of the total group by means of the respective conditional likelihoods. The same can be done by splitting the sample on the basis of an external criterion like age, gender, education, etc. Some authors, however, observed that these tests sometimes fail to reject the null hypothesis when there actually is lack of fit. Glas and Verhelst, see Chapter 5 of [7], therefore developed several powerful tests of fit, both global and item-wise. Ponocny [17] proposed nonparametric tests for hypotheses than can be chosen arbitrarily, based on a Monto-Carlo procedure. Klauer [11] and Fischer [6] developed exact single-case tests and a CML estimation method for the amount of change of θ_v between two time points, for instance, for the assessment of development or of a treatment effect in an individual.

Once the item parameters have been estimated and the fit of the RM has been established satisfactorily, the person parameters can be easily estimated by straightforward maximum likelihood. Many researchers see a disadvantage of ML estimates of θ in the fact that for raw score $r_v = 0$ the person parameter estimate diverges to $-\infty$, and for $r_v = k$, to ∞ . Warm [22] has therefore suggested a weighted ML estimator that yields finite values even for these extreme scores. (On further details about person parameter estimation, see [10]).

Remarks on Application of the RM, and an Example

A very wide range of applications of the RM both to achievement tests and questionnaires with dichotomous item format is seen in the literature. These abundant applications are often motivated as follows. If a scale consisting of dichotomous items has been constructed with the intention to measure a single latent trait via the simple raw score, then the RM should hold for that scale. Therefore, the RM is applied to check these assumptions. Two necessary conditions, however, are often overlooked: first, unidimensionality in the sense of the RM (see **Item Response Theory (IRT) Models for Dichotomous Data**) is a very strict requirement that can hold, for example, in a subscale of an intelligence test for items with homogeneous content, such like those of the Standard Progressive Matrices test or Gittler's [9] cubes test of spatial ability; but unidimensionality is extremely unlikely to occur in omnibus intelligence scales or, even more so, in questionnaires. Second, the IRFs are assumed (a) to tend to the lower limit zero and (b) the upper limit one. In order to satisfy these requirements, the response format must exclude guessing (or, at least, the guessing probabilities must be very low); this implies that a correct response occurs only if the testee has *solved* the item, so that this event is a reliable indicator of the respective ability. Technically, in questionnaires with response format 'yes' versus 'no' (or '+' versus '-') there is an extremely high guessing probability – namely, 0.50 – and hence there can be no certainty that the response '+' is an indicator of the trait of interest; the testee could as well have checked the response categories arbitrarily. Therefore, most applications of the dichotomous RM to questionnaire scales are of

little scientific value. (On generalized Rasch models for polytomous ordered response items, see below.)

To illustrate the procedure of an application of the RM, the analysis of a data sample from $n = 1160$ testees who took Gittler's '3DW' test [9] of spatial ability is now sketched. In each of the $k = 17$ items, a cube X is presented to the testee which is known to display different patterns on each of its sides, however, only three of them can be seen. Simultaneously with X, six other cubes are also presented, one of which may be equal to X but is shown in a rotated position. The testee is asked to point out which of the latter cubes is the same as X. The testee is moreover instructed to choose one of the alternative response categories 'None of the cubes is equal to X' or 'Don't know' if she/he feels uncertain whether any of the cubes equals X.

First the item parameters β_i are estimated by means of the CML method for the total sample. Then the data are split in two subsamples with raw scores below versus above the average raw score (denoted 'Group L' and 'Group H' for short). By virtue of the property of specific objectivity, the item parameter estimates in these two subgroups, denoted $\hat{\beta}_i^L$ and $\hat{\beta}_i^H$, should be the same except for sampling errors. Therefore, the points with coordinates $(\hat{\beta}_i^L, \hat{\beta}_i^H)$ in Figure 1 should fall near the straight line through the origin with slope 1. As can be seen, this is the case; the ellipses around the $k = 17$ points can be interpreted as confidence regions for $\alpha = 0.05$ with semiaxes $1.96\sigma_i^L$ in the horizontal and $1.96\sigma_i^H$ in the vertical direction. None of the points falls significantly apart from the line with slope 1.

This graphical control of the model is complemented by Andersen's [2] asymptotic CLR test: the log-likelihood is $\ln L_T = -7500.95$ for the total sample, $\ln L_L = -3573.21$ for Group L, and $\ln L_H = -3916.91$ for Group H. Therefore, Andersen's test statistic is $\chi^2 = -2[\ln L_T - (\ln L_L + \ln L_H)] = -2[-7500.95 - (-3573.21 - 3916.91)] = 21.65$ with $df = 16$, which is nonsignificant for $\alpha = 0.05$ (the critical value being $\chi_{\alpha}^2 = 26.29$). The H_0 that the RM fits the data is therefore retained under this test. (Similarly, when the sample is split by either of the external criteria gender, age, or education level, the respective test statistics are also nonsignificant.) This supports the hypothesis that the RM fits the data sufficiently well.

Another graphical tool for the assessment of fit is shown in Figure 2 (for three arbitrarily selected items

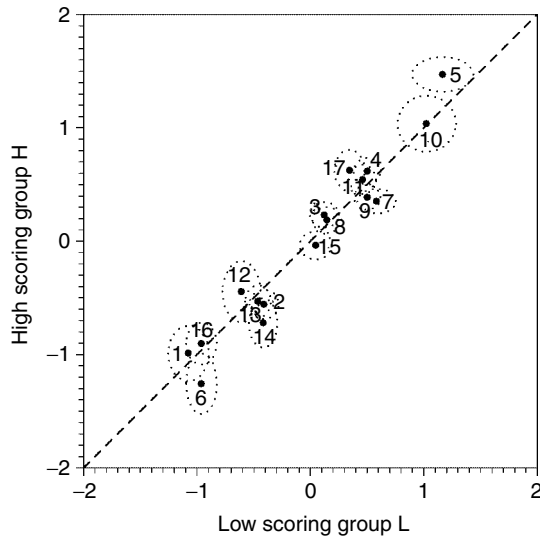


Figure 1 Graphic control of the RM for the 3DW items

of different difficulty levels, I_6 with $\hat{\beta}_6 = -1.06$, I_{10} with $\hat{\beta}_{10} = 1.02$, and I_{15} with $\hat{\beta}_{15} = -0.01$. It shows the IRFs (1) based on the parameter estimates as a function of θ , and ‘empirical IRFs’ based on the relative frequencies of correct responses in the different raw score subgroups. These relative frequencies were slightly smoothed using a so-called *normal kernel smoother* with a bandwidth of 3 adjacent frequencies at the margins and bandwidth 5 elsewhere. The confidence intervals for the estimates of the ordinates for $\alpha = 0.05$ are shown as dotted lines. (The graphs for the remaining items are similar.) This

nicely supports the hypothesis that the RM fits the items.

If the researcher has satisfied him(her)self that the RM fits, further advantages can be taken of the particular properties of the model. For instance, under the assumption that the item parameters have been estimated sufficiently well, the estimates can be considered as known constants and *exact* conditional tests can be made of the H_0 that the person parameters of two individuals are equal, or that the person parameters of the same person tested at two time points or under two different testing conditions are equal. ‘Exact’ means that the tests are based on the exact conditional distribution of the two scores, given their sum, rather than on asymptotic theory. To make such tests, it is not required that the two persons (or the same person at two time points) take exactly the same items: it suffices to present two item samples from the (unidimensional) item pool for which the RM has been found to hold. These item samples may be identical, or overlapping, or disjoint. In the present case, given that the item pool comprises only $k = 17$ items, there would be little point in choosing different (i.e., smaller) subsamples of items. Suppose therefore that the complete test is given to two testees (or the same testee twice), and the H_0 is to be tested that the two person parameters are equal (i.e., the H_0 of ‘no change’). The empirical researcher needs only to apply Table 1 which gives the significance levels for all possible score combinations r_1 and r_2 of the two testees (or of one testee at two time points) under the specified H_0 , based on the so-called ‘*Mid-P-Method*’ (cf. [6]). For example, in the following score

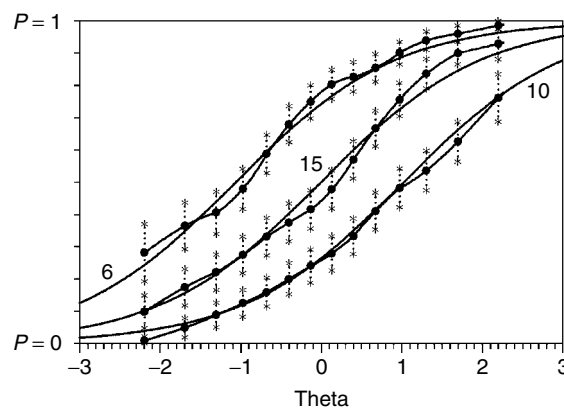


Figure 2 Theoretical and empirical IRFs for items 6, 10, and 15

Table 1 Significances of score combinations in the 3DW test (two-sided exact tests with $\alpha = 0.05$)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
0					s	s	S	S	T	T	T	T	T	T	T	T	T	T	0
1						.	s	s	S	S	T	T	T	T	T	T	T	T	1
2								.	s	S	S	S	T	T	T	T	T	T	2
3									.	s	s	S	S	T	T	T	T	T	3
4	s									.	s	s	S	S	T	T	T	T	4
5	s	.									.	s	s	S	S	T	T	T	5
6	S	s										.	s	s	S	S	T	T	6
7	S	s	.										.	s	s	S	T	T	7
8	T	S	s	.										.	s	S	S	T	8
9	T	S	S	s	.										.	s	S	T	9
10	T	T	S	s	s	.										.	s	S	10
11	T	T	S	S	s	s	.										s	S	11
12	T	T	T	S	S	s	s	.									.	s	12
13	T	T	T	T	S	S	s	s	.									s	13
14	T	T	T	T	T	S	S	s	s	.									14
15	T	T	T	T	T	T	S	S	S	s	.								15
16	T	T	T	T	T	T	T	T	S	S	s	s	.						16
17	T	T	T	T	T	T	T	T	T	T	S	S	s	s					17
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	

Note: Rows correspond to r_1 , columns to r_2 . Entries ‘.’ denote significance level 0.10, ‘s’ level 0.05, ‘S’ level 0.01, and ‘T’ level 0.001.

combinations (r_1, r_2) , score r_2 would be the lowest score that is significantly higher than score r_1 , under a two-sided alternative hypothesis with $\alpha = 0.05$: (0,4), (1,6), (2,8), (3,9), (4,10), (5,11), (6,12), (7,13), (8,14), (9,15), (10,16), (11,16), (12,17), (13,17).

The simplicity of the application of this table illustrates the advantages of having a test to which the RM can be fitted. Therefore, it is recommended to employ the RM whenever possible as a guideline for test development rather than just as a means for *post hoc* analysis of test data.

Some Extensions of the RM

Many extensions of the RM and of Rasch’s approach to measurement have been proposed. One group of them comprises models for dichotomous item responses, the other polytomous response models.

A. Dichotomous extensions. The *linear logistic test model* (LLTM) assumes that the item difficulty parameters of an RM can be explained as weighted sums of certain *basic parameters* assigned, for instance, to rules or cognitive operations involved in the solution process. The CML estimation method, uniqueness conditions, asymptotic properties, and

conditional likelihood ratio tests generalize directly from the RM to the LLTM. Studies trying to explain item parameters as weighted sums of parameters of cognitive operations or rules have often failed to be successful, however, because of lack of fit. The primary importance of the LLTM therefore seems to lie in studies with experimental or longitudinal designs (*see Clinical Trials and Intervention Studies and Longitudinal Data Analysis*) where the basic parameters are, for instance, effects of treatments given prior to – or experimental conditions prevailing at – the testing occasion. The LLTM can moreover be reformulated as a *multidimensional* longitudinal model, assigning different latent dimensions to different items without, however, requiring assumptions about the true dimensionality of the latent space. This *linear logistic model with relaxed assumptions* (LLRA) can serve to measure effects of treatments in educational, applied, or clinical psychology. (On these models, see Chapters 8 and 9 of [7].) Analogous to the LLTM is Linacre’s [12] FACETS model. Another kind of dichotomous generalization of the RM is the *mixed* RM by Rost [20], which assumes the existence of $c > 1$ latent classes between which the item parameters of the RM are allowed to differ. Yet another direction of generalization aims at

relaxing the assumption of equal discrimination of all items: the *one parameter logistic model* (OPLM) by Verhelst et al. [21] presumes a small set of discrete rational discrimination parameters α_i and assigns one of them, by means of a heuristic procedure, to each item. The difference between the OPLM and the *two-parameter logistic (or Birnbaum) model* lies in the nature of the discrimination parameters: in the latter, they are free parameters, while in the OPLM they are constants chosen by hypothesis. Sophisticated test methods are needed, of course, to verify or reject such hypotheses. Statistics to test the fit of the OPLM are given by Verhelst and Glas in Chapter 12 of [7].

B. Polytomous extensions. Andrich [3] and Masters [13] have introduced unidimensional rating scale models for items with ordered response categories (like ‘strongly agree’, ‘rather agree’, ‘rather disagree’, ‘strongly disagree’), namely, the *rating scale model* (RSM) and the *partial credit model* (PCM). The former assumes equal response categories for all items, whereas the latter allows for a different number and different definition of the response categories per item. The RM is a special case of the RSM, and the RSM a special case of the PCM. Fischer and Ponocny (see Chapter 19 of [7]) have embedded a linear structure in the parameters of these models analogously to the LLTM mentioned above. CML estimation, conditional hypothesis tests, and multidimensional reparameterizations are generalizable to this framework, as in the LLTM. These models are particularly suited for longitudinal and treatment effect studies. The individual-centered exact conditional tests of change are also applicable in these polytomous models (cf. [6]). A still more general class of multidimensional IRT models with linear structures embedded in the parameters has been developed by Adams et al. [1, 23]; these authors rely on parametric MML methods. Müller [15], moreover, has proposed an extension of the RM allowing for continuous responses.

References

- [1] Adams, R.J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model, *Applied Psychological Measurement* **21**, 1–23.
- [2] Andersen, E.B. (1973). A goodness of fit test for the Rasch model, *Psychometrika* **38**, 123–140.
- [3] Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**, 561–573.
- [4] De Leeuw, J. & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models, *Journal of Educational Statistics* **11**, 183–196.
- [5] Eggen, T.J.H.M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters, *Psychometrika* **65**, 337–362.
- [6] Fischer, G.H. (2003). The precision of gain scores under an item response theory perspective: a comparison of asymptotic and exact conditional inference about change, *Applied Psychological Measurement* **27**, 3–26.
- [7] Fischer, G.H. & Molenaar, I.W. eds (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York.
- [8] Fischer, G.H. & Ponocny-Seliger, E. (1998). *Structural Rasch Modeling. Handbook of the Usage of LPCM-WIN 1.0*, ProGAMMA, Groningen.
- [9] Gittler, G. (1991). *Dreidimensionaler Würfeltest (3DW)*, [The three-dimensional cubes test (3DW).] Beltz Test, Weinheim.
- [10] Hoijtink, H. & Boomsma, A. (1996). Statistical inference based on latent ability estimates, *Psychometrika* **61**, 313–330.
- [11] Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model, *Psychometrika* **56**, 213–228.
- [12] Linacre, J.M. (1996). *A User's Guide to FACETS*, MESA Press, Chicago.
- [13] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.
- [14] Mislevy, R.J. & Bock, R.D. (1986). *PC-Bilog: Item Analysis and Test Scoring with Binary Logistic Models*, Scientific software, Mooresville.
- [15] Müller, H. (1987). A Rasch model for continuous ratings, *Psychometrika* **52**, 165–181.
- [16] Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models, in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, G.H. Fischer & D. Laming, eds, Springer-Verlag, New York, pp. 249–263.
- [17] Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model, *Psychometrika* **66**, 437–460.
- [18] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Pædagogiske Institut, Copenhagen. (Expanded edition 1980. The University of Chicago Press, Chicago.).
- [19] Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements, in *The Danish Yearbook of Philosophy*, M. Blegvad, ed., Munksgaard, Copenhagen, pp. 58–94.
- [20] Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis, *Applied Psychological Measurement* **3**, 397–409.

8 Rasch Modeling

- [21] Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1994). *OPLM: Computer Program and Manual*, CITO, Arnhem.
- [22] Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models, *Psychometrika* **54**, 427–450.
- [23] Wu, M.L., Adams, R.J. & Wilson, M.R. (1998). *ACER ConQuest*, Australian Council for Educational Research, Melbourne.

GERHARD H. FISCHER

Rasch Models for Ordered Response Categories

DAVID ANDRICH

Volume 4, pp. 1698–1707

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rasch Models for Ordered Response Categories

Introduction

This entry explains the latent response structure and process compatible with the **Rasch model** (RM) for ordered response categories in standard formats (see **Rasch Modeling**). It considers the rationale for the RM but is not concerned with details of parameter estimation and tests of fit. There are a number of software packages that implement the RM at an advanced level. Detailed studies of the theory and applications of the RM can be found in [10], [11], [16], and [17].

Standard formats involve one response in one of the categories deemed *a priori* to reflect levels of the latent trait common in quantifying attitude, performance, and status in the social sciences. They are used by analogy to measurement in the natural sciences. Table 1 shows typical formats for four ordered categories.

The Model and its Motivation

The RM was derived from the following requirement of invariant comparisons:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison.

Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison [14, p. 332].

Rasch was not the first to require such invariance, but he was the first to formalize it in the form of a probabilistic mathematical model. Following a sequence of derivations in [14], [1], and [3], the

model was expressed in the form

$$P\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=0}^x \tau_k + x(\beta_n - \delta_i)\right), \quad (1)$$

where (a) $X_{ni} = x$ is an integer random variable characterizing $m + 1$ successive categories, (b) β_n and δ_i are respectively locations on the same latent continuum of person n and item i , (c) $\tau_k, k = 1, 2, 3, \dots, m$ are m thresholds which divide the continuum into to $m + 1$ ordered categories and which, without loss of generality, have the constraint $\sum_{k=0}^m \tau_k = 0$, and (d) $\gamma_{ni} = \sum_{x=0}^m \exp(-\sum_{k=0}^x \tau_k + x(\beta_n - \delta_i))$ is a normalizing factor that ensures that the probabilities in (1) sum to 1. For convenience of expression, though it is not present, $\tau_0 \equiv 0$. The thresholds are points at which the probabilities of responses in one of the two adjacent categories are equal.

Figure 1 shows the probabilities of responses in each category, known as category characteristic curves (CCCs) for an item with three thresholds and four categories, together with the location of the thresholds on the latent trait.

In (1) the model implies equal thresholds across items, and such a hypothesis might be relevant in example 3 of Table 1. Though the notation used in (1) focuses on the item–person response process and does not subscript the thresholds with item i , the derivation of the model is valid for the case that different items have different thresholds, giving [20]

$$P\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^x \tau_{ki} + x(\beta_n - \delta_i)\right), \quad (2)$$

in which the thresholds $\tau_{ki}, k = 1, 2, 3, \dots, m_i, \sum_{k=0}^{m_i} \tau_{ki} = 0$, are subscripted by i as well as k . $\tau_{0i} \equiv 0$. Such differences in thresholds among items might be required in examples of 1 and 2 of Table 1.

Models of (1) and (2) have become known as the *rating scale* model and *partial credit* models

Table 1 Standard response formats for the Rasch model

Example	Category 1		Category 2		Category 3		Category 4
1	Fail	<	Pass	<	Credit	<	Distinction
2	Never	<	Sometimes	<	Often	<	Always
3	Strongly disagree	<	Disagree	<	Agree	<	Strongly agree

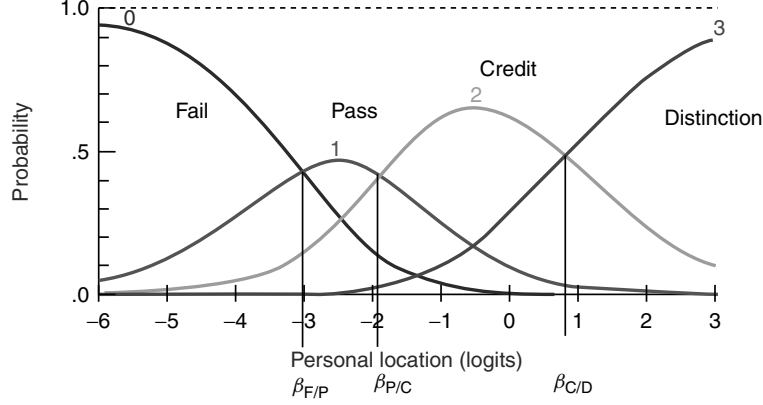


Figure 1 Category characteristic curves showing the probabilities of responses in each of four ordered categories

respectively. Nevertheless, they have an identical response structure and process for a single person responding to a single item.

Let $\delta_{ki} = \delta_i + \tau_{ki}$, $\delta_{0i} \equiv 0$. Then (2) simplifies to

$$\begin{aligned} P\{X_{ni} = x\} &= \frac{1}{\gamma_{ni}} \exp \sum_{k=1}^x (\beta_n - \delta_{ki}) \\ &= \frac{1}{\gamma_{ni}} \exp \left(x\beta_n - \sum_{k=1}^x \delta_{ki} \right). \end{aligned} \quad (3)$$

In this form, the thresholds δ_{ki} are immediately comparable across items; in the form of (2), the τ_{ki} are referenced to the location δ_i of item i , which is the mean of the thresholds δ_{ki} for each item, that is, $\delta_i = \bar{\delta}_{ki}$. It is convenient to use (3) in some illustrations and applications.

Elimination of Person Parameters

The formalization of invariance of comparisons rests on the existence of sufficient statistics, which implies that conditional on that statistic, the resultant distribution is independent of the relevant parameter. The sufficient statistic for the person parameter β_n is simply the total score $r = \sum_{i=1}^I x_{ni}$. Then for a pair of items i and j , and using directly the responses x_{ni}, x_{nj} the conditional equation

$$\begin{aligned} P\{x_{ni}, x_{nj} | r_n = x_{ni} + x_{nj}\} \\ = \frac{1}{\Psi_{nij}} \exp \sum_{k=1}^{x_{ni}} (-\delta_{ki}) \exp \sum_{k=1}^{x_{nj}} (-\delta_{kj}), \end{aligned} \quad (4)$$

where $\Psi_{nij} = \sum_{(x_{ni}, x_{nj} | r_n)} \exp \sum_{k=1}^{x_{ni}} (-\delta_{ki}) \exp \sum_{k=1}^{x_{nj}} (-\delta_{kj})$ is the summation over all possible pairs of responses given a total score of r_n . Equation (4) is clearly independent of the person parameters β_n , $n = 1, 2, \dots, N$. It can be used to estimate the item threshold parameters independently of the person parameters. In the implementation of the estimation, (4) may be generalized by considering all possible pairs of items or conditioning on the total score across more than two items. Different software implements the estimation in different ways. The person parameters are generally estimated following the estimation of the item parameters using the item parameter estimates as known. Because there are generally many less items than persons, the procedure for estimating the person parameters by conditioning out the item parameters is not feasible. **Direct maximum likelihood estimates** of the person parameters are biased and methods for reducing the bias have been devised.

Controversy and the Rasch Model

The construction of a model on the basis of *a priori* requirements rather than on the basis of characterizing data involves a different paradigm from the traditional in the data model relationship. In the traditional paradigm, if the model does not fit the data, then consideration is routinely given to finding a different model which accounts for the data better. In the Rasch paradigm, the emphasis is on whether the data fit the chosen model, and if not, then consideration is routinely given to understanding what aspect of the data is failing to fit the model.

The RM has the required structure of fundamental measurement or additive conjoint measurement in a probabilistic framework (see **Measurement: Overview**) [9, 19]. This is the main reason that those who adhere to the RM and try to construct data that fit it, give for adhering to it and for trying to collect data that fit it. In addition, they argue that the constructing measuring instruments, is not simply a matter of characterizing data, but a deliberate attempt to construct measures which satisfy important properties, and that the RM provides an operational criterion for obtaining fundamental measurement. Thus, the data from a measuring instrument are seen to be deliberately constructed to be empirical valid and at the same time satisfy the requirements of measurements [5]. Some controversy, which has been discussed in [7], has arisen from this distinctive use of the RM.

Short biographies of Rasch can be found in [2], [6], and [18].

The Latent Structure of the Model

The RM for dichotomous responses, which specializes from (3) to

$$P\{X_{ni} = x\} = \frac{\exp x(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}; \quad x \in \{0, 1\}, \quad (5)$$

is also the basis of the RM for more than two ordered categories. In this case there is only the one threshold, the location of the item δ_i .

The Derivation of the Latent Structure Assuming a Guttman Pattern

To derive the latent structure of the model, assume an instantaneous latent dichotomous response process $\{Y_{nki} = y\}$, $y \in \{0, 1\}$ at each threshold [3]. Let this latent response take the form

$$P\{Y_{nki} = y\} = \frac{1}{\eta_{ni}} \exp y(\beta_n - \delta_{ki}), \quad (6)$$

where η_{ki} is the normalizing factor $\eta_{ni} = 1 + \exp(\beta_n - \delta_{ki})$.

Although instantaneously assumed to be independent, there is only *one* response in *one* of $m + 1$ categories. Therefore, the responses must be *latent* (see **Latent Variable**). Furthermore, if the responses were independent, there would be 2^m possible response patterns. Therefore, the responses must

also be *dependent* and a *constraint* must be placed on any process in which the latent responses at the thresholds are instantaneously considered independent. The Guttman structure provides this constraint in exactly the required way. Table 2 shows the responses according to the Guttman structure for three items.

The rationale for the Guttman patterns in Table 2 [12] is that for unidimensional responses across items, if a person succeeds on an item, then the person should succeed on all items that are easier than that item and that if a person failed on an item, then the person should fail on all items more difficult than that item. A key characteristic of the Guttman pattern is that the total score across items recovers the response pattern perfectly. Of course, with experimentally independent items, that is, where each item is responded to independently of every other item, it is possible that the Guttman structure will not be observed in data and that given a particular total score the Guttman pattern will not be observed.

The rationale for the Guttman structure, as with the ordering of items in terms of their difficulty, is that the thresholds with an item are required to be ordered, that is

$$\tau_1 < \tau_2 < \tau_3 \cdots < \tau_{m-1} < \tau_m. \quad (7)$$

This requirement of ordered thresholds is independent of the RM – it is required by the Guttman structure and ordering of the categories. However, it is completely *compatible* with the structure of the responses in the RM and implies that a person at the threshold of two higher categories is at a higher location than a person at the boundary of two lower categories. For example, it requires that a person who is at the threshold between a Credit and Distinction in the first example in Table 1, and reflected in Figure 1, has a higher ability than a person who is at the threshold between a Fail and a Pass. That is, if the categories are to be ordered, then the thresholds

Table 2 The Guttman structure with dichotomous items in difficulty order

Items	1	2	3	Total score
	0	0	0	0
	1	0	0	1
	1	1	0	2
	1	1	1	3

4 Rasch Models for Ordered Response Categories

that define them should also be ordered. The derivation of the model with this requirement is shown in detail in [3].

To briefly outline this derivation, consider the case of only four ordered categories as in Table 1 and therefore three thresholds. Then if the responses at the thresholds were independent, the probability of any set of responses across the thresholds is given by

$$P\{y_{n1i}, y_{n2i}, y_{n3i}\} = \prod_{k=1}^3 \frac{\exp y_{nki}(\beta_n - \delta_{ki})}{1 + \exp(\beta_n - \delta_{ki})}, \quad (8)$$

where the sum of probabilities of *all* patterns $\sum_{All} \prod_{k=1}^3 \frac{\exp y_{nki}(\beta_n - \delta_{ki})}{1 + \exp(\beta_n - \delta_{ki})} = 1$.

The subset of Guttman G patterns has a probability of occurring

$$\begin{aligned} \Gamma &= \sum_G \prod_{k=1}^3 \frac{\exp y_{nki}(\beta_n - \delta_{ki})}{1 + \exp(\beta_n - \delta_{ki})} \\ &= \frac{\sum_G \exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}{\prod_{k=1}^3 (1 + \exp(\beta_n - \delta_{ki}))} < 1. \end{aligned} \quad (9)$$

Then the probability of a particular Guttman response pattern, conditional on the response being one of the Guttman patterns, is given by

$$\begin{aligned} P\{y_{n1i}, y_{n2i}, y_{n3i}|G\} &= \frac{\exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}{\prod_{i=1}^3 (1 + \exp(\beta_n - \delta_{ki}))} \bigg/ \Gamma \\ &= \frac{\exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}{\prod_{k=1}^3 (1 + \exp(\beta_n - \delta_{ki}))} \end{aligned}$$

$$\begin{aligned} &\bigg/ \frac{\sum_G \exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}{\prod_{k=1}^3 (1 + \exp(\beta_n - \delta_{ki}))} \\ &= \frac{\exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}{\sum_G \exp \sum_{k=1}^3 y_{nki}(\beta_n - \delta_{ki})}. \end{aligned} \quad (10)$$

For example, suppose the response pattern is $\{1, 1, 0\}$, in which case the total score is 2. Then

$$\begin{aligned} P\{1, 1, 0|G\} &= \frac{\exp[1(\beta_n - \delta_{1i}) + 1(\beta_n - \delta_{2i}) + 0(\beta_n - \delta_{3i})]}{\sum_G \exp \sum_{i=1}^3 y_{nki}(\beta_n - \delta_{ki})} \\ &= \frac{\exp(2\beta_n - \delta_{1i} - \delta_{2i})}{\sum_G \exp \sum_{i=1}^3 y_{nki}(\beta_n - \delta_{ki})}. \end{aligned} \quad (11)$$

Notice that the coefficient of the person location β_n in the numerator of (11) is 2, the total score of the number of successes at the thresholds. This scoring can be generalized and the total score can be used to define the Guttman response pattern. Thus, define the integer random variable $X_{ni} = x \in \{0, 1, 2, 3\}$ as the total score for each of the Guttman patterns: $0 \equiv (0, 0, 0)$, $1 \equiv (1, 0, 0)$, $2 \equiv (1, 1, 0)$, $3 \equiv (1, 1, 1)$. Then (11) simplifies to

$$P\{X_{ni} = x\} = \frac{\exp(x\beta_n - \delta_{1i} - \delta_{2i} \dots - \delta_{xi})}{\sum_{x=0}^3 \exp \sum_{k=1}^x (\beta_n - \delta_{ki})}, \quad (12)$$

which is the special case of (3) in the case of just three thresholds and four categories.

Effectively, the successive categories are scored with successive integers as in elementary analyses of ordered categories. However, no assumption of equal distances between thresholds is made; the thresholds are estimated from the data and may be unequally spaced. The successive integer scoring rests on the discrimination of the latent dichotomous responses

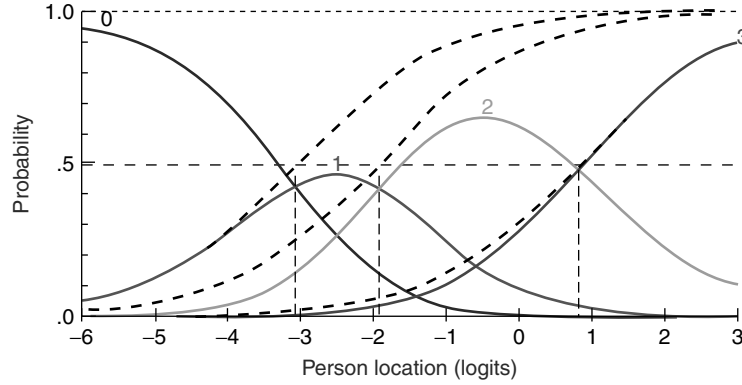


Figure 2 Probabilities of responses in each of four ordered categories showing the probabilities of the latent dichotomous responses at the thresholds

at the thresholds within an item being the same. Although the successive categories are scored with successive integers, it is essential to recognize that the response in any category implies a success at thresholds up to and including the lower threshold defining a category and failure on subsequent thresholds including the higher threshold defining a category. That is, the latent response structure of the model is a Guttman pattern of *successes* and *failures* at all the thresholds. Figure 2 shows Figure 1 augmented by the probabilities of the latent dichotomous responses at the thresholds according to (6).

The Derivation of the Latent Structure Resulting in the Guttman Pattern

This response structure is confirmed by considering the ratio of the probability of a response in any category, conditional on the response being in one of two adjacent categories. The probability of the response being in the higher of the two categories is readily shown to be

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\} + P\{X_{ni} = x\}} = \frac{\exp(\beta_n - \delta_{xi})}{1 + \exp(\beta_n - \delta_{xi})}, \quad (13)$$

which is just the dichotomous model at the threshold defined in (6).

This conditional latent response between two adjacent categories is dichotomous and again latent. It is

an implication of the latent structure of the model because there is no sequence of observed conditional responses at the thresholds: there is just one response in one of the m categories.

In the above derivation, a Guttman pattern based on the ordering of the thresholds was imposed on an initially independent set of responses. Suppose now that the model is defined by the latent responses at the thresholds according to (13). Let

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\} + P\{X_{ni} = x\}} = P_x, \quad (14)$$

and its complement

$$\frac{P\{X_{ni} = x - 1\}}{P\{X_{ni} = x - 1\} + P\{X_{ni} = x\}} = Q_x = 1 - P_x. \quad (15)$$

Then it can be shown readily that

$$P\{X_{ni} = x\} = P_1 P_2 P_3 \dots P_x Q_{x+1} Q_{x+2} \dots Q_m / D, \quad (16)$$

where $D = Q_1 Q_2 Q_3 \dots Q_m + P_1 Q_2 Q_3 \dots Q_m + P_1 P_2 Q_3 \dots Q_m + \dots P_1 P_2 P_3 \dots P_m$.

Clearly, the particular response $X_{ni} = x$ implies once again successes at the first x thresholds and failures at all the remaining thresholds. That is, the response structure results in successes at the thresholds consistent with the Guttman pattern. This in turn implies an ordering of the thresholds. Thus, both derivations lead to the same structure at the thresholds.

6 Rasch Models for Ordered Response Categories

The Log Odds Form and Potential for Misinterpretation

Consider (12) again.

Taking the ratio of the response in two adjacent categories gives the odds of success at the threshold:

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \exp(\beta_n - \delta_{xi}). \quad (17)$$

Taking the logarithm gives

$$\ln \frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \beta_n - \delta_{xi}. \quad (18)$$

This log odds form of the model, while simple, eschews its richness and invites making up a response process that has nothing to do with the model. It does this because it can give the impression that there is an independent response at each threshold, an interpretation which incorrectly ignores that there is only one response among the categories and that the dichotomous responses at the thresholds are latent, implied, and never observed. This form loses, for example, the fact that the probability of a response in any category is a function of all thresholds. This can be seen from the normalizing constant denominator in (3), which contains all thresholds. Thus, the probability of a response in the first category is affected by the location of the last threshold.

Possibility of Reversed Thresholds in Data

Although the ordering of the thresholds is required in the data and the RM is compatible with such ordering, it is possible to have data in which the thresholds, when estimated, are not in the correct order. This can occur because there is only one response in one category, and there is no restriction on the distribution of those responses.

The relative distribution of responses for a single person across triplets of successive categories can be derived simply from (16) for pairs of successive categories:

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \exp(\beta_n - \delta_{xi}) \quad (19)$$

and

$$\frac{P\{X_{ni} = x + 1\}}{P\{X_{ni} = x\}} = \exp(\beta_n - \delta_{x+1,i}). \quad (20)$$

Therefore,

$$\begin{aligned} & \frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} \frac{P\{X_{ni} = x\}}{P\{X_{ni} = x + 1\}} \\ &= \exp(\delta_{x+1,i} - \delta_{xi}). \end{aligned} \quad (21)$$

If $\delta_{x+1,i} > \delta_{xi}$, then $\delta_{x+1,i} - \delta_{xi} > 0$, $\exp(\delta_{x+1,i} - \delta_{xi}) > 1$ then from (21),

$$[P\{X_{ni} = x\}]^2 > P\{X_{ni} = x - 1\}P\{X_{ni} = x + 1\}. \quad (22)$$

Because each person responds only in one category of an item, there is no constraint in the responses to conform to (22); this is an empirical matter. The violation of order can occur even if the data fit the model statistically. It is a powerful property of the RM that the estimates of the thresholds can be obtained independently of the distribution of the person parameters, indicating that the relationship among the thresholds can be inferred as an intrinsic structure of the of the operational characteristics of the item.

Figure 3 shows the CCCs of an item in which the last two thresholds are reversed. It is evident that the threshold between Pass and Credit has a greater location than the threshold between Credit and Distinction. It means that if this is accepted, then the person who has 50% chance of being given a Credit or a Distinction has less ability than a person who has 50% chance of getting a Pass or a Credit. This clearly violates any a-priori principle of ordering of the categories. It means that there is a problem with the empirical ordering of the categories and that the successive categories do not reflect increasing order on the trait.

Other symptoms of the problem of reversed thresholds is that there is no region in Figure 3 in which the grade of Credit is most likely and that the region in which Credit should be assigned is undefined – it implies some kind of negative distance. Compatible with the paradigm of the RM, reversed threshold estimates direct a closer study of the possible reasons that the ordering of the categories are not working as intended.

Although lack of data in a sample in any category can result in estimates of parameters with large standard errors, the key factor in the estimates is the relationship amongst the categories of the implied probabilities of (22). These cannot be inferred directly from raw frequencies in categories in a sample. Thus, in the

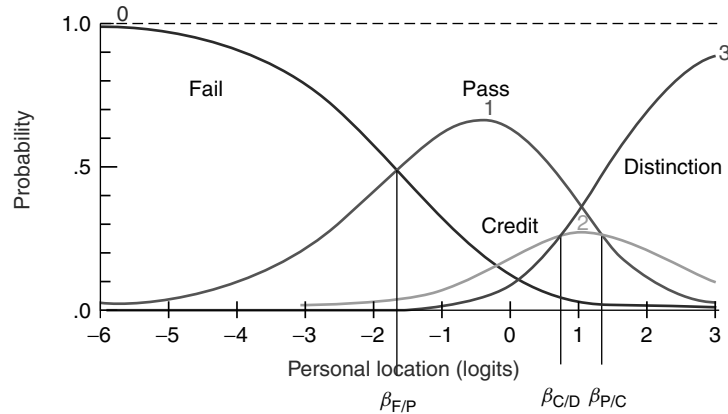


Figure 3 Category characteristic curves showing the probabilities of responses in each of four ordered categories when the thresholds are disordered

case of Figure 2, any *single* person whose ability estimate is between the thresholds identified by $\beta_{C/D}$ and $\beta_{P/C}$ will, simultaneously, have a higher probability of a getting a Distinction and a Pass than getting a Credit. This is not only incompatible with ordering of the categories, but it is *not a matter of the distribution of the persons in the sample of data analyzed*.

To consolidate this point, Table 3 shows the frequencies of responses of 1000 persons for two items each with 12 categories. These are simulated data, which fit the model to have correctly ordered thresholds. It shows that in the middle categories, the frequencies are very small compared to the extremes, and in particular, the score of 5 has a 0 frequency for item 1. Nevertheless, the threshold estimates shown in Table 4 have the required order. The method of estimation, which exploits the structure of responses

among categories to span and adjust for the category with 0 frequency and conditions out the person parameters, is described in [8]. The reason that the frequencies in the middle categories are low or even 0 is that they arise from a bimodal distribution of person locations. It is analogous to having heights of a sample of adult males and females. This too would be bimodal and therefore heights somewhere in between the two modes would have, by definition, a low frequency. However, it would be untenable if the low frequencies in the middle heights would reverse the lines (thresholds) which define the units on the ruler. Figure 4 shows the frequency distribution of the estimated person parameters and confirms that it is bimodal. Clearly, given the distribution, there would be few cases in the middle categories with scores of 5 and 6.

Table 3 Frequencies of responses in two items with 12 categories

Item	0	1	2	3	4	5	6	7	8	9	10	11
I0001	81	175	123	53	15	0	8	11	51	120	165	86
I0002	96	155	119	57	17	5	2	26	48	115	161	87

Table 4 Estimates of thresholds for two items with low frequencies in the middle categories

Threshold estimates												
Item	$\hat{\delta}_i$	1	2	3	4	5	6	7	8	9	10	11
1	0.002	-3.96	-2.89	-2.01	-1.27	-0.62	-0.02	0.59	1.25	2.00	2.91	4.01
2	-0.002	-3.78	-2.92	-2.15	-1.42	-0.73	-0.05	0.64	1.36	2.14	2.99	3.94

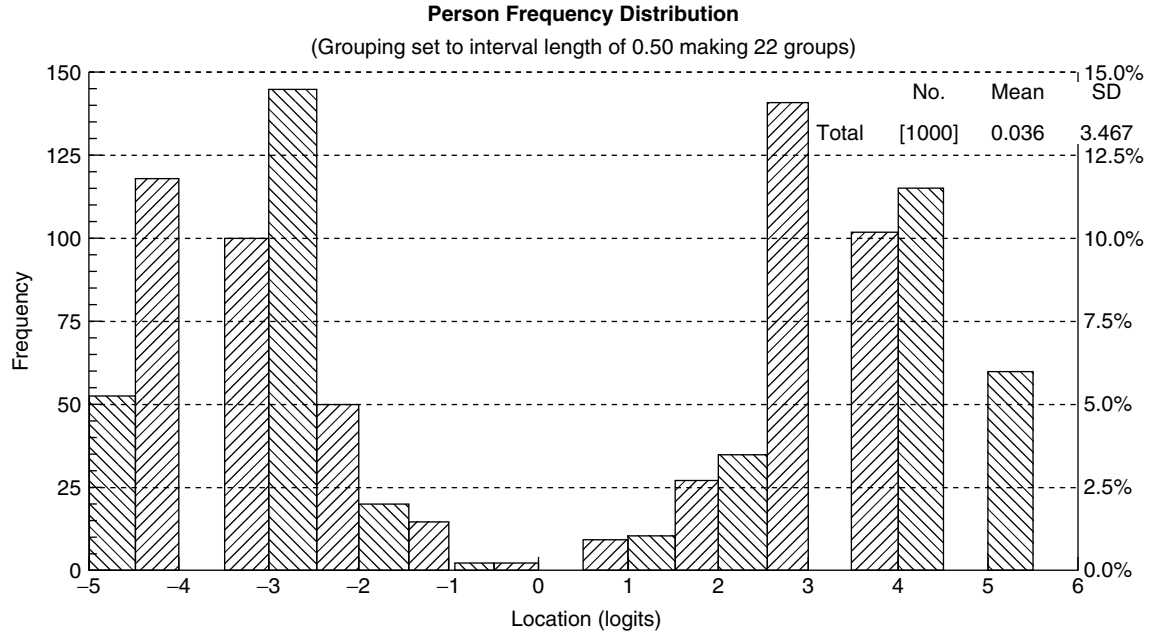


Figure 4 A bimodal distribution of locations estimates

The Collapsing of Adjacent Categories

A distinctive feature of the RM is that summing the probabilities of adjacent categories, produces a model which is no longer a RM. In particular, dichotomizing in that way is not consistent with the model. That is, taking (3), and forming

$$\begin{aligned}
 &P\{X_{ni} = x\} + P\{X_{ni} = x + 1\} \\
 &= \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^x \tau_{ki} + x(\beta_n - \delta_i)\right) \\
 &\quad + \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^{x+1} \tau_{ki} + x(\beta_n - \delta_i)\right) \quad (23)
 \end{aligned}$$

gives (23) which cannot be reduced to the form of (3). This result has been discussed in [4], [13] and was noted by Rasch [15]. It implies that collapsing categories is not arbitrary but an integral property of the data revealed through the model. This nonarbitrariness in combining categories contributes to the model providing information on the empirical ordering of the categories.

The Process Compatible with the Rasch Model

From the above outline of the structure of the RM, it is possible to summarize the response *process* that is compatible with it. The response process is one of *simultaneous ordered classification*. That is, the process is one of considering the property of an object, which might be a property of oneself or of some performance, relative to an item with two or more than two ordered categories, and deciding the category of the response.

The examples in Table 1 show that each successive category implies the previous category in the order *and in addition*, reflects more of the assessed trait. This is compatible with the Guttman structure. Thus, a response in a category implies that the latent response was a success at the lower of the two thresholds, and a failure at the greater of the two thresholds. *And this response determines the implied latent responses at all of the other thresholds.* This makes the response process a *simultaneous classification* process across the thresholds. The further implication is that when the manifest responses are used to estimate the thresholds, the threshold

locations are themselves empirically defined *simultaneously* - that is, the estimates arise from data in which all the thresholds of an item were involved simultaneously in every response. This contributes further to the distinctive feature of the RM that it can be used to assess whether or not the categories are working in the intended ordering, or whether on this feature, the empirical ordering breaks down.

References

- [1] Andersen, E.B. (1977). Sufficient statistics and latent trait models, *Psychometrika* **42**, 69–81.
- [2] Andersen, E.B. & Olsen, L.W. (2000). The life of Georg Rasch as a mathematician and as a statistician, in *Essays in Item Response Theory*, A. Boomsma, M.A.J. van Duijn & T.A.B. Snijders, eds, Springer, New York, pp. 3–24.
- [3] Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**, 357–374.
- [4] Andrich, D. (1995). Models for measurement, precision and the non-dichotomization of graded responses, *Psychometrika* **60**, 7–26.
- [5] Andrich, D. (1996). Measurement criteria for choosing among models for graded responses, in *Analysis of Categorical Variables in Developmental Research*, Chap. 1, A. von Eye & C.C. Clogg, eds, Academic Press, Orlando, pp. 3–35.
- [6] Andrich, D. (2004a). *Georg Rasch: Mathematician and Statistician*. *Encyclopaedia of Social Measurement*, Academic Press.
- [7] Andrich, D. (2004b). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care* **42**, 7–16.
- [8] Andrich, D. & Luo, G. (2003). Conditional estimation in the Rasch model for ordered response categories using principal components, *Journal of Applied Measurement* **4**, 205–221.
- [9] Brogden, H.E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement, *Psychometrika* **42**, 631–634.
- [10] Engelhard Jr, G. & Wilson, M., eds (1996). *Objective Measurement: Theory into Practice*, Vol. 3, Norwood, Ablex Publishing.
- [11] Fischer, G.H. & Molenaar, I.W., eds, (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, Springer, New York.
- [12] Guttman, L. (1950). The basis for scalogram analysis, in *Measurement and Prediction*, S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star & J.A. Clausen, eds, Wiley, New York, pp. 60–90.
- [13] Jansen, P.G.W. & Roskam, E.E. (1986). Latent trait models and dichotomization of graded responses, *Psychometrika* **51**(1), 69–91.
- [14] Rasch, G. (1961). On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. IV, J. Neyman, ed., University of California Press, Berkeley, pp. 321–334.
- [15] Rasch, G. (1966). An individualistic approach to item analysis, in *Readings in Mathematical Social Science*, P.F. Lazarsfeld & N.W. Henry, eds, Science Research Associates, Chicago, pp. 89–108.
- [16] Smith Jr, E.V. & Smith, R.M., eds, (2004). *Introduction to Rasch Measurement*, JAM Press, Maple Grove.
- [17] Van der Linden, W. & Hambleton, R., eds (1997). *Handbook of Modern Item Response Theory*, Springer, New York, pp. 139–152.
- [18] Wright, B.D. (1980). Foreword to *Probabilistic Models for Some Intelligence and Attainment Tests*, G., Rasch, ed., 1960 Reprinted University of Chicago Press.
- [19] Wright, B.D. (1997). A history of social science measurement, *Educational Measurement: Issues and Practice* **16**(4), 33–45.
- [20] Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*, MESA Press, Chicago.

(See also **Rasch Modeling**)

DAVID ANDRICH

Rater Agreement

CHRISTOF SCHUSTER AND DAVID SMITH

Volume 4, pp. 1707–1712

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rater Agreement

Introduction

Human judgment is prone to error that researchers routinely seek to quantify, understand, and minimize. To quantify the quality of nominal judgments, agreement among multiple raters of the same target is examined using either global summary indices, such as Cohen's kappa (*see* **Rater Agreement – Kappa**), or by modeling important properties of the judgment process using **latent class analysis**. In this entry, we give an overview of the latent class analysis approach to nominal scale rater agreement data.

Latent class models of rater agreement assume there is an underlying or 'true' latent category to which each target belongs, and that clues to the nature of this *latent class*, and to the nature of the judgment process itself, can be found in the observed classifications of multiple raters. With targets belonging to only one of the latent classes, manifest disagreement among observed judgments requires that at least one judgment be erroneous. However, without knowledge of the true latent class, or a 'gold standard' indicator of the true latent class, correct and erroneous judgments cannot be distinguished. Even though individual target classifications cannot be directly established at the latent level, given certain assumptions are fulfilled, latent class analysis can nevertheless estimate overall misclassification probabilities.

Data Example

To illustrate the rater agreement models we discuss in this article, suppose that 212 patients were diagnosed according to four raters as either 'schizophrenic' or 'not schizophrenic.' For two categories and four raters, there are 16 possible rating profiles, which are given in the first four columns of Table 1. The frequency with which each profile occurred is given in the fifth column.¹ As the frequencies in this table show, all four judges agree on 111 of the 212 targets. In other words, in approximately 48% of the targets there is at least some disagreement among the judges.

The Latent Class Model

The event $A = i$, $i = 1, \dots, I$, indicates a target assigned to the i th category by rater A . Similarly,

Table 1 Rating profile frequencies, n , of psychiatric diagnoses according to four raters, A_1, A_2, A_3, A_4 , where '1' = schizophrenic and '2' = not schizophrenic

A_1	A_2	A_3	A_4	n
1	1	1	1	12
1	1	1	2	10
1	1	2	1	4
1	1	2	2	8
1	2	1	1	10
1	2	1	2	6
1	2	2	1	7
1	2	2	2	8
2	1	1	1	0
2	1	1	2	1
2	1	2	1	1
2	1	2	2	8
2	2	1	1	8
2	2	1	2	15
2	2	2	1	15
2	2	2	2	99

$X = t$, $t = 1, \dots, T$, indicates a target truly belongs to the t th latent class. The probabilities of these events are denoted as $P(A = i) = \pi_A(i)$ and $P(X = t) = \pi_X(t)$, respectively. The conditional probability (*see* **Probability: An Introduction**) that a rater will assign a target of the t th latent class to the i th category is denoted as $P(A = i|X = t) = \pi_{A|X}(i|t)$. To avoid unnecessarily complex notation, the illustrative latent class model example we present in this chapter (*viz.* Table 1) assumes four raters, denoted as A_1, A_2, A_3 , and A_4 .

Latent class analysis is based on the assumption of local (conditional) independence (*see* **Conditional Independence**). According to this assumption, multiple judgments of the same target are independent. Therefore, the raters and latent status joint probability factors as

$$\begin{aligned} \pi_{A_1 A_2 A_3 A_4 X}(i, j, k, l, t) \\ = \pi_{A_1|X}(i|t) \pi_{A_2|X}(j|t) \pi_{A_3|X}(k|t) \pi_{A_4|X}(l|t) \pi_X(t). \end{aligned} \quad (1)$$

This raters and latent status joint probability can be related to the rating profile probability by

$$\pi_{A_1 A_2 A_3 A_4}(i, j, k, l) = \sum_{t=1}^T \pi_{A_1 A_2 A_3 A_4 X}(i, j, k, l, t). \quad (2)$$

2 Rater Agreement

This formula shows that the probabilities of the rating profiles, the left-side of (2), are obtained by collapsing the joint probability $\pi_{A_1 A_2 A_3 A_4 X}(i, j, k, l, t)$ over the levels of the latent class variable. Combining (1) and (2) yields

$$\begin{aligned} & \pi_{A_1 A_2 A_3 A_4}(i, j, k, l) \\ &= \sum_{t=1}^T \pi_{A_1|X}(i|t) \pi_{A_2|X}(j|t) \pi_{A_3|X}(k|t) \pi_{A_4|X}(l|t) \pi_X(t) \end{aligned} \quad (3)$$

This equation relates the probabilities of the observed rating profiles to conditional response probabilities and the latent class probabilities. For this model to be identified, it is necessary, though not sufficient, for the degrees of freedom, calculated as $df = I^R - (RI - R + 1)T$, where R denotes the number of raters, to be nonnegative.

Although the present review of latent class models only involves one categorical latent variable, the model is applicable also to more than one categorical latent variable. This is because the cells in a cross-classification can be represented equivalently as categories of a single nominal variable. Thus, if models are presented in terms of multiple latent variables – as is often the case when modeling rater agreement data – this is done only for conceptual clarity, not out of statistical necessity. We return to this issue in the next section.

Furthermore, it is often desirable to restrict the probabilities on the right-hand side of (3), either for substantive reasons or to ensure model identification. Probabilities can be fixed to known values, set equal to one another, or any combination of the two. The wide variety of rater agreement latent class models found in the literature emerges as a product of the wide variety of restrictions that can be imposed on the model probabilities.

Distinguishing Target Types

In the rater agreement literature, it has become increasingly popular to introduce a second latent class variable that reflects the common recognition that targets differ not only with respect to their status on the substantive latent variable of interest but also with respect to the target's prototypicality with respect to the latent class (e.g., [6]). These 'target type' models consider the ease with which targets can be classified,

although 'ease' in this context is only metaphorical and should not necessarily be equated with any perceived experience on the part of the raters.

We will illustrate the operation of the target type latent class variable by considering three target types; obvious, suggestive, and ambiguous. Where Y is the target type latent variable, we arbitrarily define $Y = 1$ for obvious targets, $Y = 2$ for suggestive targets, and $Y = 3$ for ambiguous targets. The latent variables X and Y are assumed to be fully crossed. In other words, all possible combinations of levels of X and Y have a positive probability of occurrence. However, we do not assume populations that necessarily involve all three target types.

Obvious Targets. Targets belonging to the obvious latent class can be readily identified. These targets are often referred to as prototypes or 'textbook cases'. To formalize this idea, one restricts the conditional probabilities $\pi_{A|XY}(i|t, 1) = \delta_{it}$, where $\delta_{it} = 1$ if $i = t$ and $\delta_{it} = 0$ else. Because this restriction is motivated by characteristics of the targets, it applies to all raters in the same manner. In connection with the local independence assumption, and continuing our four judge example, one obtains

$$\pi_{A_1 A_2 A_3 A_4|Y}(i, j, k, l|1) = \delta_{ijkl} \pi_{X|Y}(i|1), \quad (4)$$

where $\delta_{ijkl} = 1$ if $i = j = k = l$ and $\delta_{ijkl} = 0$ else.

Suggestive Targets. While targets belonging to the suggestive latent class are not obvious, their rate of correct assignment is better than chance. Suggestive targets possess features tending toward, but not decisively establishing, a particular latent status. In this sense, the suggestive targets are the ones for which the latent class model has been developed. Formally speaking, the suggestive class model is

$$\begin{aligned} & \pi_{A_1 A_2 A_3 A_4|Y}(i, j, k, l|2) \\ &= \sum_{t=1}^T \pi_{A_1|XY}(i|t, 2) \pi_{A_2|XY}(j|t, 2) \pi_{A_3|XY}(k|t, 2) \\ & \quad \times \pi_{A_4|XY}(l|t, 2) \pi_{X|Y}(t|2). \end{aligned} \quad (5)$$

Note that (5) is equivalent to (3).

Ambiguous Targets. Judgments of ambiguous targets are no better than random. In conditional probability terms, this means that ambiguous target judgments do not depend on the target's true latent class.

In other words, $\pi_{A|XY}(i|t, 3) = \pi_{A|Y}(i|3)$. Together with the local independence assumption, this yields

$$\begin{aligned} \pi_{A_1 A_2 A_3 A_4 | Y}(i, j, k, l|3) \\ = \pi_{A_1 | Y}(i|3) \pi_{A_2 | Y}(j|3) \pi_{A_3 | Y}(k|3) \pi_{A_4 | Y}(l|3). \end{aligned} \quad (6)$$

Models Having a Single Target Type

Having delineated three different target types, it is possible to consider models that focus on a single target type only. Of the three possible model classes, – models for suggestive, obvious, or ambiguous targets – only models for suggestive targets are typically of interest to substantive researchers. Nevertheless, greater appreciation of the latent class approach can be gained by briefly considering the three situations in which only a single target type is present.

Suggestive Targets Only. Because latent class models for rater agreement are based on the premise that for each observed category there exists a corresponding latent class, it is natural to consider the general latent class model that has as many latent classes as there are observed categories. In other words, one can consider the general latent class model for which $T = I$. This model has been considered by, for example, Dillon and Mulani [2].²

Fitting this model to the data given in Table 1 produces an acceptable model fit. Specifically, one obtains $X^2 = 9.501$ and $G^2 = 10.021$ based on $df = 7$, allowing for one boundary value. Boundary values are probabilities that are estimated to be either zero or one. For the hypothetical population from which this sample has been taken the prevalence of schizophrenia is estimated to be $\pi_X(1) = 0.2788$. Individual rater sensitivities, $\pi_{A_r | X}(1|1)$, for $r = 1, 2, 3, 4$ are estimated as 1.0000, .5262, .5956, and .5381. Similarly, individual rater specificities, $\pi_{A_r | X}(2|2)$, are estimated as .9762, .9346, .8412, and .8401.

Obvious Targets Only. If only obvious targets are judged, there would be perfect agreement. Every rater would be able to identify the latent class to which each target correctly belongs. Of course, in this situation, statistical models would be unnecessary.

Ambiguous Targets Only. If a population consists only of ambiguous targets, the quality of the judgments would be no better than chance. Consequently,

the judgments would be meaningless from a substantive perspective. Nevertheless, it is possible to distinguish at least two different modes either of which could underlie random judgments. First, raters could produce random judgments in accordance with category specific base-rates, that is, base-rates that could differ across raters. Alternatively, judgments could be random in the sense of their being evenly distributed among the categories.

Models Having Two Types of Targets

Allowing for two different types of targets produces three different model classes of the form

$$\begin{aligned} \pi_{A_1 A_2 A_3 A_4}(i, j, k, l) = \pi_{A_1 A_2 A_3 A_4 | Y}(i, j, k, l|u) \pi_Y(u) \\ + \pi_{A_1 A_2 A_3 A_4 | Y}(i, j, k, l|v) \pi_Y(v), \end{aligned} \quad (7)$$

where $u, v = 1, 2, 3$. For models that include only obvious and ambiguous targets, $u = 1$ and $v = 3$; for obvious and suggestive targets only, $u = 1$ and $v = 2$; and for suggestive and ambiguous targets only, $u = 2$ and $v = 3$. Of course, in each of the three cases the two conditional probabilities on the right-hand side of (7) will be replaced with the corresponding expressions given in (4), (5), and (6). Because the resulting model equations are evident, they are not presented.

Obvious and Ambiguous Targets. Treating all targets as either obvious or ambiguous has been suggested by Clogg [1] and by Schuster and Smith [8]. Fitting this model to the data example yields a poor fit. Specifically, one obtains $X^2 = 22.6833$ and $G^2 = 28.1816$ based on $df = 9$. The proportion of obvious targets, $\pi_Y(1)$, is estimated as 44.80%. Among these, only 8.43% are estimated to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .0843$. The rater specific probabilities of a schizophrenia diagnosis for the ambiguous targets, $\pi_{A_r | Y}(1|3)$, $r = 1, 2, 3, 4$, are .4676, .2828, .4399, and .4122 respectively. Clearly, while raters one, three, and four seem to behave similarly when judging ambiguous targets, the second rater's behavior is different from the other three. Note that this model does not produce a prevalence estimate of schizophrenia for ambiguous targets.

Obvious and Suggestive Targets. Treating all targets as either obvious or suggestive has been considered by Espeland and Handelman [3]. Fitting this

4 Rater Agreement

model to the data produces an acceptable model fit. Specifically, one obtains $X^2 = 6.5557$ and $G^2 = 7.1543$ based on $df = 5$, allowing for one boundary value. The proportion of obvious targets is estimated as 23.43%. Among these, only 5.22% are estimated to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .0522$. Similarly, among the 76.57% suggestive targets one estimates 35.32% to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .3532$. The individual rater sensitivities for the suggestive targets, $\pi_{A_r|XY}(1|1, 2)$, $r = 1, 2, 3, 4$, are 1.0000, .5383, .6001, .5070, and the corresponding specificities, $\pi_{A_r|XY}(2|2, 2)$, are .9515, .8994, .7617, and .7585. Overall, the values for the sensitivities and specificities are very similar to the model that involves only suggestive targets.

Suggestive and Ambiguous Targets. Treating all targets as either suggestive or ambiguous has been considered by Espeland and Handelman [3] and by Schuster and Smith [8]. Fitting this model to the data yields an acceptable model fit of $X^2 = 1.8695$ and $G^2 = 1.7589$ based on $df = 4$, allowing for three boundary values. The proportion of suggestive targets, $\pi_Y(2)$, is estimated to be 85.81% of which 28.30% are schizophrenic, that is, $\pi_{X|Y}(1|2) = .2830$. The individual rater sensitivities for the suggestive targets, $\pi_{A_r|XY}(1|1, 2)$, $r = 1, 2, 3, 4$, are 1.0000, .5999, .6373, .5082 for raters A_1 to A_4 , and the corresponding specificities, $\pi_{A_r|XY}(2|2, 2)$, are .9619, .9217, .8711, and 1.0000. The rater specific probabilities of a schizophrenia diagnosis for the ambiguous targets are .2090, .0000, .3281, and .9999 for raters A_1 to A_4 . It is remarkable how much the likelihood of a schizophrenia diagnosis for ambiguous targets differs for raters two and four. If the relatively large number of boundary values is not indicative of an inappropriate model, then it is as if rater 2 will not diagnose schizophrenia unless there is enough specific evidence of it, while rater 4 views ambiguity as diagnostic of schizophrenia. Note that this model cannot produce a prevalence estimate of schizophrenia for ambiguous targets.

Three Types of Targets

When the population of targets contains all three types, one obtains

$$\begin{aligned} \pi_{A_1 A_2 A_3 A_4}(i, j, k, l) &= \pi_{A_1 A_2 A_3 A_4|Y}(i, j, k, l|1)\pi_Y(1) \\ &+ \pi_{A_1 A_2 A_3 A_4|Y}(i, j, k, l|2)\pi_Y(2) \end{aligned}$$

$$+ \pi_{A_1 A_2 A_3 A_4|Y}(i, j, k, l|3)\pi_Y(3), \quad (8)$$

where, as before, each of the three conditional probabilities on the right-hand side are replaced using (4–6). Although this model follows naturally from consideration of different targets, we are not aware of an application of this model in the literature. For the present data example the number of parameters of this model exceeds the number of rating profiles, and, therefore, the model is not identified.

Espeland and Handelman [3] have discussed this model assuming equally probable categories. In this case, the model can be fitted to the data in Table 1. However, for the present data set, the probability of ambiguous targets is estimated to be zero. Therefore, the model fit is essentially equivalent to the model involving only obvious and suggestive targets.

Discussion

Table 2 summarizes the goodness-of-fit statistics for the models fitted to the data in Table 1. When comparing the goodness-of-fit statistics of different models one has to keep in mind that differences between likelihood-ratio statistics follow a central chi-square distribution only if the models are nested. Thus, the only legitimate comparisons are between Model 1 and Model 3 ($\Delta G^2 = 2.8663$, $\Delta df = 3$, $p = .413$) and between Model 1 and Model 4 ($\Delta G^2 = 8.2617$, $\Delta df = 3$, $p = .041$). Clearly, insofar as these data are concerned allowing for obvious targets in addition to suggestive targets does not improve model fit considerably. However, allowing for ambiguous targets in addition to suggestive targets improves the model fit considerably.

Depending on the target type, additional restrictions can be imposed on the latent class model. In particular, for suggestive targets one can constrain the hit-rates, $\pi_{A|X}(i|i)$, to be equal across raters,

Table 2 Goodness-of-fit statistics for models fitted to the data in Table 1. The models are specified in terms of the target types involved

Model	Target-types	df	X^2	G^2
1	$Y = 2$	7	9.5010	10.0206
2	$Y = 1, Y = 3$	9	22.6833	28.1816
3	$Y = 1, Y = 2$	5	6.5557	7.1543
4	$Y = 2, Y = 3$	4	1.8695	1.7589

across categories, or across raters and categories simultaneously, see [2] or [8] for details. Alternatively, one can constrain error-rates across raters or targets.

For ambiguous targets the response probabilities can be constrained to be equal across raters, that is, $\pi_{A_r|Y}(i|3) = \pi_{A_q|Y}(i|3)$ for $r \neq q$. It is also possible to define random assignment more restrictively, such as requiring each category be equally probable. In this case, each of the four probabilities on the right-hand side of (6) would be replaced with $1/I$, where I s denotes the number of categories.

Of course, the restrictions for suggestive and ambiguous targets can be combined. In particular, if the rater panel has been randomly selected, one should employ restrictions that imply homogeneous rater margins, that is, $\pi_{A_r}(i) = \pi_{A_q}(i)$ for $r \neq q$. For random rater panels one could also consider symmetry models, that is, models that imply $\pi_{A_1 A_2 A_3 A_4}(i, j, k, l)$ is constant for all permutations of index vector (i, j, k, l) .

Finally, an alternative way in which a second type of latent variable can be introduced is to assume two different rater modes or states in which the raters operate [5], which is an idea closely related to the models proposed by Schutz [9] and Perrault and Leigh [7]. Ratets in reliable mode judge targets correctly with a probability of 1.0, while raters in unreliable mode will 'guess' the category. This model is different from the target type models inasmuch as the representation of the rater mode will require a separate latent variable for each rater. In addition, the rater modes are assumed independent. Therefore, the model involves restrictions among the latent variables, which is not the case for traditional latent class models.

Notes

1. A similar data set has been presented by Young, Tanner, and Meltzer [10].

2. However, note that for a given number of raters, the number of latent classes that are identifiable may depend on the number of observed categories. For dichotomous ratings, model identification requires at least three raters. In cases of either three or four categories, four raters are required to ensure model identification (see [4]).

References

- [1] Clogg, C.C. (1979). Some latent structure models for the analysis of likert-type data, *Social Science Research* **8**, 287–301.
- [2] Dillon, W.R. & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability, *Multivariate Behavioral Research* **19**, 438–458.
- [3] Espeland, M.A. & Handelman, S.L. (1989). Using latent class models to characterize and assess relative error in discrete measurements, *Biometrics* **45**, 587–599.
- [4] Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**(2), 215–231.
- [5] Klauer, K.C. & Batchelder, W.H. (1996). Structural analysis of subjective categorical data, *Psychometrika* **61**(2), 199–240.
- [6] Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation, *British Journal of Psychiatry* **130**, 79–83.
- [7] Perrault, W.D. & Leigh, L.E. (1989). Reliability of nominal data based on qualitative judgments, *Journal of Marketing Research* **26**, 135–148.
- [8] Schuster, C. & Smith, D.A. (2002). Indexing systematic rater agreement with a latent class model, *Psychological Methods* **7**(3), 384–395.
- [9] Schutz, W.C. (1952). Reliability, ambiguity and content analysis, *Psychological Review* **59**, 119–129.
- [10] Young, M.A., Tanner, M.A. & Meltzer, H.Y. (1982). Operational definitions of schizophrenia: what do they identify? *Journal of Nervous and Mental Disease* **170**(8), 443–447.

CHRISTOF SCHUSTER AND DAVID SMITH

Rater Agreement – Kappa

EUN YOUNG MUN

Volume 4, pp. 1712–1714

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rater Agreement – Kappa

Whether there is agreement or consensus among raters or observers on their evaluation of the objects of interest is one of the key questions in the behavioral sciences. Social workers assess whether parents provide appropriate rearing environments for their children. School psychologists evaluate whether a child needs to be placed in a special class. If agreement among raters is low or absent, it begs many questions, including the validity of the guidelines or criteria, and the reliability of the raters involved in their judgment. Thus, it is important to establish rater agreement in the behavioral sciences. One way to measure rater agreement is to have two raters make independent observations on the same group of objects, to classify them into a limited number of categories, and then to see to what extent the raters' evaluations overlap. There are several measures to assess rater agreement for this type of situation. Of all measures, Cohen's kappa (κ , [1]) is one of the best known and most frequently used measures to assess the extent of agreement by raters in a summary statement for entire observations [3].

Cohen's kappa measures the strength of rater agreement against the expectation of independence of ratings. Independence of ratings refers to a situation where the judgment made by one rater or observer is independent of, or unaffected by, the judgment made by the other rater. Table 1 illustrates a typical cross-classification table of two raters (*see Contingency Tables*). Three classification categories by both raters are completely crossed, resulting in a square table of nine cells. Of the nine cells, the three diagonal cells from the left to the right, shaded, are called the *agreement cells*. The remaining six other cells

are referred to as the *disagreement cells*. The numbers in the cells are normally frequencies, and are indexed by m_{ij} . Subscripts i and j index I rows and J columns. m_{11} , for example, indicates the number of observations for which both raters used Category 1. The last row and column in Table 1 represent the row and column totals. m_{1i} indicates the row total for Category 1 by Rater A, whereas m_{i1} indicates the column total for Category 1 by Rater B. N is for the total number of objects that are evaluated by the two raters.

Cohen's kappa is calculated by

$$\hat{\kappa} = \frac{\sum p_{ii} - \sum p_{i.}p_{.i}}{1 - \sum p_{i.}p_{.i}},$$

where $\sum p_{ii}$ and $\sum p_{i.}p_{.i}$ indicate the *observed* and the *expected* sample proportions of agreement based on independence of ratings, respectively. Based on Table 1, the observed proportion of agreement can be calculated by adding frequencies of all agreement cells and dividing the total frequency of agreement by the number of total observations, $\sum p_{ii} = \sum m_{ii}/N = (m_{11} + m_{22} + m_{33})/N$. The expected sample proportion of agreement can be calculated by summing the products of the row and the column sums for each category, and dividing the sum by the squared total number of observations, $\sum p_{i.}p_{.i} = \sum m_{i.}m_{.i}/N^2 = (m_{1.}m_{.1} + m_{2.}m_{.2} + m_{3.}m_{.3})/N^2$.

The numerator of Cohen's kappa formula suggests that when the observed proportion of cases in which the two independent raters agree is greater than the expected proportion, then kappa is positive. When the observed proportion of agreement is less than the expected proportion, Cohen's kappa is negative. In addition, the difference in magnitude between the observed and the expected proportions is factored into Cohen's kappa calculation. The greater the

Table 1 A typical cross-classification of rater agreement

		Rater B			Row sum
		Category 1	Category 2	Category 3	
Rater A	Category 1	m_{11}	m_{12}	m_{13}	m_{1i}
	Category 2	m_{21}	m_{22}	m_{23}	m_{2i}
	Category 3	m_{31}	m_{32}	m_{33}	m_{3i}
	Column Sum	m_{i1}	m_{i2}	m_{i3}	N

2 Rater Agreement – Kappa

Table 2 Cross-classification of rater agreement when Cohen's kappa = 0.674

		Psychologist B			Row sum
		Securely attached	Resistant or ambivalent	Avoidant	
Psychologist A	Securely Attached	64	4	2	70
	Resistant or Ambivalent	5	4	1	10
	Avoidant	1	2	17	20
	Column Sum	70	10	20	100

difference between the two proportions, the higher the absolute value of Cohen's kappa. The difference is scaled by the proportion possible for improvement. Namely, the denominator of the formula indicates the difference between the perfect agreement (i.e., Cohen's kappa = 1) and the expected proportion of agreement. In sum, Cohen's kappa is a measure of the proportion of agreement relative to that expected based on independence. κ can also be characterized as a measure of *proportionate reduction in error* (PRE; [2]). A κ value multiplied by 100 indicates the percentage by which two raters' agreement exceeds the expected agreement from chance.

Perfect agreement is indicated by Cohen's kappa = 1. Cohen's kappa = 0 if ratings are completely independent of each another. The higher Cohen's kappa, the better the agreement. Although it is rare, it is possible to have a negative Cohen's kappa estimate. A negative kappa estimate indicates that observed agreement is worse than expectation based on chance. Cohen's kappa can readily be calculated using a general purpose statistical software package and a significance test is routinely carried out to see whether κ is different from zero. In many instances, however, researchers are more interested in the magnitude of Cohen's kappa than in its significance. While there are more than one suggested guidelines for acceptable or good agreement, in general, a $\hat{\kappa}$ value less than 0.4 is considered as fair or slight agreement. A $\hat{\kappa}$ value greater than 0.4 is considered as good agreement (see [3] for more information).

Cohen's kappa can be extended to assess agreement by more than two raters. It can also be used to assess ratings or measurements carried out for multiple occasions across time. When more than two raters are used in assessing rater agreement, the expected proportion of agreement can be calculated by utilizing a standard main effect loglinear analysis.

Data Example

Suppose two psychologists observed infants' reactions to a separation and reunion situation with their mothers. Based on independent observations, psychologists A and B determined the attachment styles of 100 infants. Typically 70%, 10%, and 20% of infants are classified as *Securely Attached*, *Resistant or Ambivalent*, and *Avoidant*, respectively. Table 2 shows an artificial data example. The proportion of *observed* agreement is calculated at 0.85 (i.e., $(64 + 4 + 17)/100 = 0.85$), and the *expected* agreement is calculated at 0.54 (i.e., $(70 * 70 + 10 * 10 + 20 * 20)/10000 = 0.54$). The difference of 0.31 in proportion between the observed and expected agreement is compared against the maximum proportion that can be explained by rater agreement. $\hat{\kappa} = (0.85 - 0.54)/(1 - 0.54) = 0.674$, with standard error = 0.073, $z = 8.678$, $p < 0.01$. Thus, we conclude that two psychologists agree to 67.4% more than expected by chance, and that the agreement between two psychologists is significantly better than chance.

References

- [1] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.
- [2] Fleiss, J.L. (1975). Measuring agreement between two judges in the presence or absence of a trait, *Biometrics* **31**, 651–659.
- [3] von Eye, A. & Mun, E.Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum, Mahwah.

EUN YOUNG MUN

Rater Agreement – Weighted Kappa

EUN YOUNG MUN

Volume 4, pp. 1714–1715

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rater Agreement – Weighted Kappa

One of the characteristics of Cohen’s kappa (κ , [1]) (see **Rater Agreement – Kappa**) is that any discrepancy between raters is equally weighted as zero. On the other hand, any agreement means *absolute agreement* between raters, and is equally weighted as one. Thus, the distinction between agreement and disagreement is categorical. For example, Table 1 from the article on Cohen’s kappa (see **Rater Agreement – Kappa**) shows three agreement cells and six disagreement cells. In deriving Cohen’s kappa, all six disagreement cells are treated equally. However, there could be situations in which the discrepancy between adjacent categories such as Categories 1 and 2 can be considered as partial agreement as opposed to complete disagreement. In those situations, Cohen’s weighted kappa [2] can be used to reflect the magnitude of agreement. Thus, Cohen’s weighted kappa is used when categories are measured by ordinal (or interval) scales. What we mean by ordinal is that categories reflect orderly magnitude. For example, categories of *Good*, *Average*, and *Bad* reflect a certain order in desirability. The disagreement by raters between *Good* and *Average*, or *Average* and *Bad* can be considered as less of a problem than the discrepancy between *Good* and *Bad*, and furthermore, the disagreement by one category or scale can sometimes be considered as partial agreement.

Cohen’s weighted kappa is calculated by

$$\hat{\kappa}_w = \frac{\sum \omega_{ij} p_{ij} - \sum \omega_{ij} p_{i.} p_{.j}}{1 - \sum \omega_{ij} p_{i.} p_{.j}}, \quad (1)$$

where the ω_{ij} are the weights, and $\sum \omega_{ij} p_{ij}$ and $\sum \omega_{ij} p_{i.} p_{.j}$ are the *weighted observed* and the *weighted expected* sample proportions of agreement,

based on the assumption of independence of ratings, respectively (see **Rater Agreement – Kappa**). Subscripts i and j index I rows and J columns. The weights take a value between zero and one, and they should be ratios. For example, the weight of one can be assigned to the agreement cells (i.e., $\omega_{ij} = 1$, if $i = j$), and 0.5 can be given to the partial agreement cells of adjacent categories (i.e., $\omega_{ij} = 0.5$, if $|i - j| = 1$). And zero for the disagreement cells that are off by more than one category (i.e., $\omega_{ij} = 0$, if $|i - j| > 1$). Thus, all weights range from zero to one, and the weights of 1, 0.5, and 0 are ratios of one another (see Table 1). Cohen’s weighted kappa tends to be higher than Cohen’s unweighted kappa because weighted kappa takes into account partial agreement between raters (see [3] for more information).

Data Example

Suppose two teachers taught a course together and independently graded 100 students in their class. Table 2 shows an artificial data example. Students were graded on an ordinal scale of A, B, C, D, or F grade. The five diagonal cells, shaded, indicate *absolute agreement* between two teachers. These cells get a weight of one. The adjacent cells that differ just by one category get a weight of 0.75 and the adjacent cells that differ by two and three categories get weights of 0.50 and 0.25, respectively. The proportions of the *weighted observed* agreement and the *weighted expected* agreement are calculated at 0.92 and 0.71, respectively. Cohen’s weighted kappa is calculated at 0.72 (e.g., $\hat{\kappa}_w = (0.92 - 0.71)/(1 - 0.71)$) with standard error = 0.050, $z = 14.614$, $p < 0.01$. This value is higher than the value for Cohen’s unweighted kappa = 0.62 with standard error = 0.061, $z = 10.243$, $p < 0.01$. The 10% increase in agreement reflects the partial weights

Table 1 An example of weights

		Rater B		
		Category 1	Category 2	Category 3
Rater A	Category 1	1(1)	0.5(0)	0(0)
	Category 2	0.5(0)	1(1)	0.5(0)
	Category 3	0(0)	0.5(0)	1(1)

Note: The numbers in parenthesis indicate weights of Cohen’s unweighted kappa.

2 Rater Agreement – Weighted Kappa

Table 2 Two teachers' ratings of students' grades (weights in parenthesis)

		Teacher B					Row sum
		A	B	C	D	F	
Teacher A	A	14 (1)	2 (0.75)	1 (0.50)	0 (0.25)	0 (0)	17
	B	4 (0.75)	25 (1)	6 (0.75)	1 (0.50)	0 (0.25)	36
	C	1 (0.50)	3 (0.75)	18 (1)	5 (0.75)	0 (0.50)	27
	D	0 (0.25)	1 (0.50)	3 (0.75)	14 (1)	1 (0.75)	19
	F	0 (0)	0 (0.25)	0 (0.50)	0 (0.75)	1 (1)	1
	Column sum	19	31	28	20	2	N = 100

given to the adjacent cells by one, two, and three categories.

References

- [1] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.
- [2] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.
- [3] von Eye, A. & Mun, E.Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum, Mahwah.

EUN YOUNG MUN

Rater Bias Models

KIMBERLY J. SAUDINO

Volume 4, pp. 1716–1717

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Rater Bias Models

Rater bias occurs when behavioral ratings reflect characteristics of the rater in addition to those of the target. As such, rater bias is a type of method variance (i.e., systematically contributes variance to behavioral ratings that are due to sources other than the target). Rater personality, response styles (e.g., the tendency to rate leniently or severely), normative standards, implicit theories (e.g., stereotypes, halo effects) and even mood can influence ratings of a target's behaviors.

In phenotypic research, rater biases are of concern because they result in disagreement between raters while at the same time inflating correlations between variables that are rated by the same rater [2]. In twin studies (*see Twin Designs*), the effects of rater biases are more complex. When both members of a twin pair are assessed by the *same* rater, rater biases act to inflate estimates of **shared environmental** variance. That is, there is covariance between the biases that affect the ratings of each twin such that the rater tends to consistently overestimate or underestimate the behavior of *both* cotwins. This consistency across cotwins would act to inflate the similarity of both monozygotic (MZ) and dizygotic (DZ) twins. Thus, the correlation between cotwins is due in part to bias covariance. Since bias covariance would not be expected to differ according to twin type, its net effect will be to result in overestimates of shared environmental variance. However, if each member of a twin pair is assessed by a *different* informant, rater biases will result in overestimates of **nonshared environmental** variance. Here, there is no bias covariance between the ratings of each twin – different raters will have different biases that influence their behavioral ratings. Thus, rater bias will be specific to each twin and will contribute to differences between cotwins and, consequently, will be included in estimates of nonshared environment in quantitative genetic analyses.

Rater bias can be incorporated into quantitative genetic models [e.g., [1, 3]]. According to the Rater Bias model, observed scores are a function of the individual's latent phenotype, rater bias, and unreliability. The basic model requires at least two raters, each of whom has assessed both cotwins. Under this model, it is assumed that raters agree because they are assessing the same latent phenotype (i.e., what

is common between raters is reliable trait variance). This latent phenotype is then decomposed into its genetic and environmental components. Disagreement between raters is assumed to be due to rater bias and unreliability. The model includes latent bias factors for each rater that account for bias covariance between the rater's ratings of each twin (i.e., what is common *within* a rater *across* twins is bias). Unreliability is modeled as rater-specific, twin-specific variances. Thus, this model decomposes the observed variance in ratings into reliable trait variance, rater bias, and unreliability; and allows for estimates of genetic and environmental influences on the reliable trait variance independent of bias and error [1].

As indicated above, the Rater Bias model assumes that rater bias and unreliability are the only reasons why raters disagree, but this may not be the case. Different raters might have different perspectives or knowledge about the target's behaviors. Therefore, it is important to evaluate the relative fit of the Rater Bias model against a model that allows raters to disagree because each rater provides different but valid information regarding the target's behavior. Typically, the Rater Bias model is compared to the Psychometric model (also known as the **Common Pathway model**). Like the Rater Bias model, this model suggests that correlations between raters arise because they are assessing a common phenotype that is influenced by genetic and/or environmental influences; however, this model also allows genetic and environmental effects specific to each rater. Thus, behavioral ratings include a common phenotype (accounting for agreement between raters) and specific phenotypes unique to each rater (accounting for disagreement between raters). As with the Rater Bias model, the common phenotype represents reliable trait variance. Because neither rater bias nor unreliability can result in the systematic effects necessary to estimate genetic influences, the specific genetic effects represent real effects that are unique to each rater [4]. Specific shared environmental influences may, however, be confounded by rater biases. When the Psychometric model provides a relatively better fit to the data than the Rater Bias model and rater-specific genetic variances estimated, it suggests that rater differences are not simply due to rater bias (i.e., that the raters to some extent assess different aspects of the target's behaviors).

2 Rater Bias Models

References

- [1] Hewitt, J.K., Silberg, J.L., Neale, M.C., Eaves, L.J. & Erikson, M. (1992). The analysis of parental ratings of children's behavior using LISREL, *Behavior Genetics* **22**, 292–317.
- [2] Hoyt, W.T. (2000). Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods* **5**, 64–86.
- [3] Neale, M.C. & Stevenson, J. (1989). Rater bias in the EASI temperament survey: a twin study, *Journal of Personality and Social Psychology* **56**, 446–455.
- [4] van der Valk, J.C., van den Oord, E.J.C.G., Verhulst, F.C. & Boomsma, D.I. (2001). Using parent ratings to study the etiology of 3-year-old twins' problem behaviors: different views or rater bias? *Journal of Child Psychology and Psychiatry* **42**, 921–931.

KIMBERLY J. SAUDINO

Reactivity

PATRICK ONGHENA

Volume 4, pp. 1717–1718

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Reactivity

Reactivity is a threat to the construct validity of observational and intervention studies that refers to the unintended reaction of participants on being observed, or, more in general, on being in a study. It can entail a threat to the construct validity of outcomes or a threat to the construct validity of treatments [4, 5, 12].

Reactivity as a threat to the *outcome* construct validity may be involved if some sort of reactive or obtrusive assessment procedure is used, that is, an assessment procedure for which it is obvious to the participants that some aspect of their functioning is being observed or measured [13, 14]. Reactivity can occur both with respect to pretests and with respect to posttests. For example, in self-report pretests, participants may present themselves in a way that makes them eligible for a certain treatment [1]. Posttests can become a learning experience if certain ideas presented during the treatment ‘fall into place’ while answering a posttreatment questionnaire [3]. Pretest as well as posttest reactivity yields observations or measurements that tap different or more complex constructs (including participant perceptions and expectations) than the constructs intended by the researcher.

Reactivity as a threat to the *treatment* construct validity may be involved if participants are very much aware of being part of a study and interpret the research or treatment setting in a way that makes the actual treatment different from the treatment as planned by the researcher [9, 11]. Such confounding treatment reactivity can be found in the research literature under different guises: hypothesis guessing within experimental conditions [10], **demand characteristics** [6], **placebo effects** [2], and evaluation apprehension [7]. A closely related risk for treatment construct invalidity is formed by (*see Expectancy Effect by Experimenters*) [8], but here the prime source for bias are the interpretations of the researcher himself, while in reactivity, the interpretations of the participants are directly at issue.

Finally, it should be remarked that reactivity is a validity threat to both **quasi-experiments** and ‘true’ experiments. Random assignment procedures clearly provide no solution to reactivity problems. In fact,

the random assignment itself might be responsible for changes in the measurement structure or meaning of the constructs involved, even before the intended treatment is brought into action [1]. In-depth discussion and presentation of methods to obtain unobtrusive measures and guidelines to conduct nonreactive research can be found in [9], [13], and [14].

References

- [1] Aiken, L.S. & West, S.G. (1990). Invalidity of true experiments: self-report pretest biases, *Evaluation Review* **14**, 374–390.
- [2] Beecher, H.K. (1955). The powerful placebo, *Journal of the American Medical Association* **159**, 1602–1606.
- [3] Bracht, G.H. & Glass, G.V. (1968). The external validity of experiments, *American Educational Research Journal* **5**, 437–474.
- [4] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.
- [5] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand-McNally, Chicago.
- [6] Orne, M.T. (1962). On the social psychology of the psychological experiment, *American Psychologist* **17**, 776–783.
- [7] Rosenberg, M.J. (1965). When dissonance fails: on elimination of evaluation apprehension from attitude measurement, *Journal of Personality and Social Psychology* **1**, 28–42.
- [8] Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York.
- [9] Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*, McGraw-Hill, New York.
- [10] Rosenzweig, S. (1933). The experimental situation as a psychological problem, *Psychological Review* **40**, 337–354.
- [11] Rosnow, R.L. & Rosenthal, R. (1997). *People Studying People: Artifacts and Ethics in Behavioral Research*, Freeman, New York.
- [12] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.
- [13] Webb, E.J., Campbell, D.T., Schwartz, R.D. & Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences in the Social Sciences*, Rand McNally Chicago.
- [14] Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L. & Grove, J.B. (1981). *Nonreactive Measures in the Social Sciences*, 2nd Edition, Houghton Mifflin, Boston.

PATRICK ONGHENA

Receiver Operating Characteristics Curves

DANIEL B. WRIGHT

Volume 4, pp. 1718–1721

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Receiver Operating Characteristics Curves

The receiver operating characteristics curve, better known as ROC and sometimes called *receiver operating curve* or *relative operating characteristics*, is the principal graphical device for **signal detection theory** (SDT). While ROCs were originally developed in engineering and psychology [3], they have become very common in medical diagnostic research (see [6] for a recent textbook, and also journals such as *Medical Decision Making*, *Radiology*, *Investigative Radiology*).

Consider an example: I receive about 100 emails a day, most of which are unsolicited junk mail. When each email arrives, it receives a ‘spamicity’ value from an email filter. Scores close to 1 mean that the filter predicts the email is ‘spam’; scores close to 0 predict the email is nonspam (‘ham’). I have to decide above what level of spamicity, I should automatically delete the emails. For illustration, suppose the data for 1000 emails and two different filters are those shown in Table 1.

The two most common approaches for analyzing such data are **logistic regression** and SDT. While in their standard forms, these are both types of the **generalized linear model**, they have different origins and different graphical methods associated with them. However, the basic question is the same: what is the relationship between spamicity and whether the email is spam or ham? ROCs are preferred when the decision criterion is to be determined and when the decision process itself is of interest, but when discrimination is important researchers should choose the logistic regression route. Because both

these approaches are based on similar methods [1], it is sometimes advisable to try several graphical methods and use the ones which best communicate the findings.

There are two basic types of curves: empirical and fitted. Empirical curves show the actual data (sometimes with slight modifications). Fitted curves are based on some model. The fitted curves vary on how much they are influenced by the data versus the model. In general, the more parameters estimated in the model, the more influence the data will have. Because of this, some ‘fitted’ models are really just empirical curves that have been smoothed (*see Kernel Smoothing*) (see [5] and [6] for details of the statistics underlying these graphs).

The language of SDT is explicitly about accuracy and focuses on two types: sensitivity and specificity. Sensitivity means being able to detect spam when the email is spam; and specificity means just saying an email is spam if it is spam. In psychology, these are usually referred to as hits (or true positives) and correct rejections. The opposite of specificity is the false alarm rate: the proportion of time that the filter predicts that real email. Suppose the data in Table 1 were treated as predicting ham if spamicity is 0.5 or less and predicting spam if it is above. Table 2 shows the breakdown of hits and false alarms by whether an email is or is not spam, and whether the filter decrees it as spam or ham. Included also are the calculations for hit and false alarm rates. The filters only provide a spamicity score. I have to decide above what level of spamicity the email should be deleted. The choice of criterion is important. This is dealt with later and is the main advantage of ROCs over other techniques.

It is because all the values on one side of a criterion are classified as either spam or ham that ROCs are cumulative graphs. Many statistical packages

Table 1 The example data used throughout this article. Each filter had 1000 emails, about half of which were spam. The filter gave each email a spamicity rating

	Spamicity ratings										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Filter one											
Email is ham	47	113	30	122	104	41	21	20	4	1	0
Email is spam	0	2	3	10	78	66	27	114	3	28	166
Filter two											
Email is ham	199	34	0	44	11	40	17	8	21	66	63
Email is spam	51	37	3	20	21	45	21	11	17	74	197

2 Receiver Operating Characteristics Curves

Table 2 The decision table if the criterion to delete emails was that the spamicity scores be 0.6 or above. While the two filters have similar hit rates, the first filter has a much lower false alarm rate (i.e., better specificity)

	Filter 1			Filter 2			
	≤ 0.5	> 0.6	Rate	≤ 0.5	> 0.6	Rate	Total
Is ham	457 correct rejection	46 false alarm	9%	328 correct rejection	175 false alarm	35%	503
Is spam	159 miss	338 hit	68%	177 miss	320 hit	64%	497
Total	616	384		505	495		1000

have cumulative data transformation functions (for example, in SPSS the CSUM function), so this can be done easily in mainstream packages. In addition, good freeware is available (for example, ROCKIT, RscorePlus), macros have been written for some general packages (for example, SAS and S-Plus), and other general packages contain ROC procedures (for example, SYSTAT). For each criterion, the number of hits and false alarms is divided by the number of spam and ham emails, respectively.

Figures 1(a) and 1(b) are the standard empirical ROCs and the fitted binormal curves (using RscorePlus [2]). The program assumes ham and spam vary on some dimension of spamicity (which is related

to, but not the same as, the variable spamicity) and that these distributions of ham and spam are normally distributed on this dimension. The normal distribution assumption is there for historical reasons though nowadays researchers often use the logistic distribution, which yields nearly identical results and is simpler mathematically; the typical package offers both these alternatives plus others. In psychology, usually normality is assumed largely because Swets [3] showed that much psychological data fits with this assumption (and therefore also with the logistic distribution). In other fields like medicine, this assumption is less often made [6]. ROCs usually show the concave pattern of Figures 1(a) and

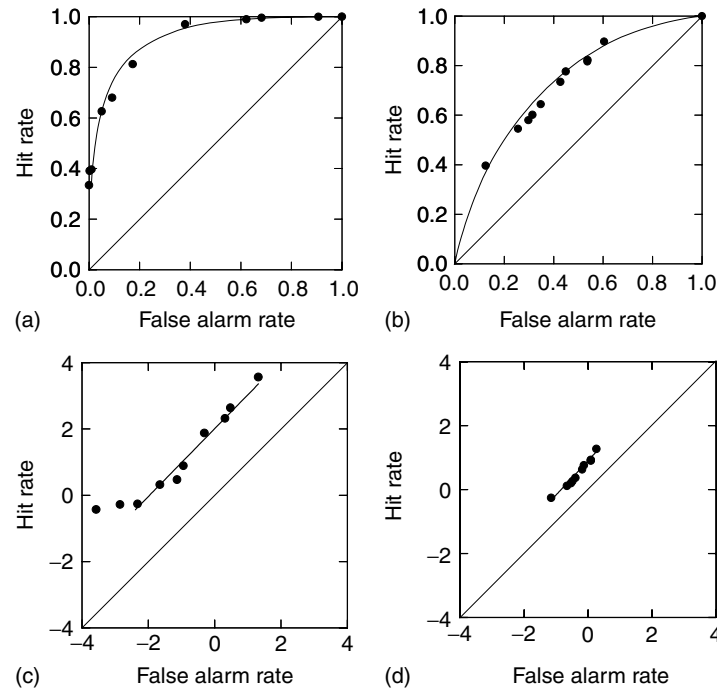


Figure 1 (a) and (b) plot the empirical hit rate and false alarm rates with fitted binormal model. Figures (c) and (d) show these graphs after they have been normalized so that the fitted line is straight

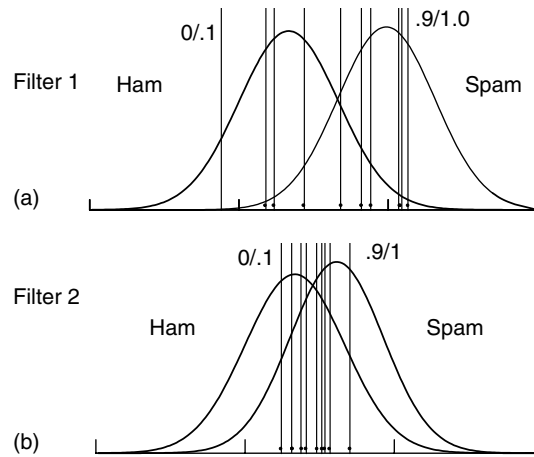


Figure 2 The models for filters (a) and (b) based on the fitted ROCs. The vertical lines show the 10 different response criteria

1(b). The diagonal lines stand for chance responding, and the further away from the diagonal the more diagnostic the variable spamicity is. Thus, comparing these two ROCs shows that the first filter is better.

The data are more easily seen if they are transformed so that the predicted lines are straight. These are sometimes called *normalized*, *standardized*, or *z-ROCs*. They can be calculated with the inverse normal function in most statistics packages and are available in most SDT software. These are shown in Figures 1(c) and 1(d). The straight lines in these graphs provide two main pieces of information. First, if the slopes are near 1, which they are for both these graphs, then they suggest the distributions for ham and spam have the same variance. The distance from the line to the diagonal is a measure of diagnosticity. If the slope of the line is 1, then this distance is the same at all points on the line and it corresponds to the SDT statistic d' . If the slope is not 1, then the distance between the diagonal and the line depends on the particular decision criterion. Several statistics are available for this (see [5] and [6]).

The fitted models in Figures 1(a–d) can be shown as normal distributions. Figures 2(a) and 2(b) indicate how well the ham and the spam can be separated by the filters. For the first filter, the spam distribution is about two standard deviations away from the ham distribution, while it is only about one standard deviation adrift for the second filter. The decision criteria are also included on these graphs.

For example, the far left criterion is at the cut-off between 0.0 and 0.1 on the spamicity variable. As can be seen, for the second filter the criteria are all close together, which means the users would have less choice about where along the dimension they could choose a criterion. These graphs are useful ways of communicating the findings.

An obvious question is whether the normal distribution assumption is valid. There are statistical tests that look at this, but it can also be examined graphically. Note, however, that as these graphs are cumulative, the data are not independent. Consequently, conducting standard regressions for the data in Figures 1(c) and 1(d) is not valid, and in these cases, the analysis should be done on the noncumulative data or using specialized packages like those cited earlier.

More advanced and less restrictive techniques are discussed in [6], but are relatively rare. The more common procedure is simply to draw straight lines between each of the observed points. This is called the *trapezoid method* because the area below the line between each pair of points is a trapezoid. Summing these trapezoids gives a measure called A . This tends to underestimate the area under real ROC curves. The values of A can range between 0 and 1 with chance discrimination as 0.5. The first filter has $A = 0.92$ and the second has $A = 0.73$.

An important characteristic of SDT is how the relative values of sensitivity and specificity are calculated and used to determine a criterion. In the

4 Receiver Operating Characteristics Curves

spam example, having low specificity would mean many real emails (ham) are labelled as spam and deleted. Arguably, this is more problematic than having to delete a few unsolicited spam. Therefore, if my filter automatically deletes messages, I would want the criterion to be high because specificity is more important than sensitivity.

To decide the relative value of deleting ham and not deleting spam, an equation from expected utility theory needs to be applied (slightly adapting the notation from [3], p.127):

$$S_{\text{opt}} = \frac{P(\text{ham})}{P(\text{spam})} \times \frac{V_{\text{CR}} - V_{\text{FA}}}{V_{\text{Hit}} - V_{\text{miss}}} \quad (1)$$

where S_{opt} is the optimal slope, $P(\text{ham})$ is the probability that an item will be ham, $P(\text{spam})$ is $1 - P(\text{ham})$, V_{CR} is the value of correctly not identifying ham as spam (this will be positive), V_{FA} is the value of incorrectly identifying ham as spam (negative), V_{Hit} is the value of correctly identifying spam (positive), and V_{miss} is the value of not identifying spam (negative). (It is worth noting here that a separate study is needed to estimate these utilities unless the minimum sensitivity or specificity is set externally; in such cases, simply go to this value on the ROC.) Thus, the odds value (see **Odds and Odds Ratios**) of ham is directly proportional to the slope. It is important to realize how important this baseline is for deciding the decision criterion. Often, people do not consider the baseline information when making decisions (see [4]).

Once S_{opt} is found, if one of the standard fitted ROC curves is used, then the optimal decision point is where the curve has this slope. For more complex

fitted curves and empirical curves, start in the upper left-hand corner of the ROC with a line of slope S_{opt} and move towards the opposite corner. The point where the line first intersects the ROC shows where the optimal decision criterion should be. Because there are usually only a limited number of possible decision criteria, the precision of this method is usually adequate to identify the optimal criterion.

This discussion only touches the surface of an exciting area of contemporary statistics. This general procedure has been expanded to many different experimental designs (see [2] and [5]), and has been generalized for **meta-analyses**, correlated and biased data, robust methods, and so on [6].

References

- [1] DeCarlo, L.T. (1998). Signal detection theory and generalized linear models, *Psychological Methods* **3**, 186–205.
- [2] Harvey Jr, L.O. (2003). *Parameter Estimation of Signal Detection Models: RscorePlus User's Manual*. Version 5.4.0, (<http://psych.colorado.edu/~lharvey/>, as at 17.08.04).
- [3] Swets, J.A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum.
- [4] Tversky, A. & Kahneman, D. (1980). Causal schemes in judgements under uncertainty, in *Progress in Social Psychology*, M. Fishbein, ed., Lawrence Erlbaum, Hillsdale.
- [5] Wickens, T.D. (2002). *Elementary Signal Detection Theory*, Oxford University Press, New York.
- [6] Zhou, X.-H., Obuchowski, N.A. & McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York.

DANIEL B. WRIGHT

Recursive Models

LILIA M. CORTINA

Volume 4, pp. 1722–1723

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Recursive Models

In behavioral research, statistical models are often *recursive*, meaning that causality flows only in one direction. In other words, these models only include unidirectional effects (e.g., variable A influences variable B, which in turn influences variable C; see Figure 1).

These cross-sectional recursive models do not include circular effects or *reciprocal causation* (e.g., variable A influences variable B, which in turn influences variable A; see Figure 2(a)), nor do they permit feedback loops (variable A influences variable B, which influences variable C, which loops back to influence variable A; see Figure 2(b)).

In addition, recursive models do not allow variables that are linked in a causal chain to have correlated disturbance terms (also known as ‘error terms’ or ‘residuals’). If all of these criteria for being recursive are met, then the model is *identified* – that is, a unique value can be obtained for each free parameter from the observed data, yielding a single solution for all equations underlying the model (see **Identification**) [1, 4].

Distinctions are sometimes made between models that are ‘recursive’ versus ‘fully recursive’. In a *fully recursive* model (or a ‘fully saturated recursive model’), each variable directly influences all other variables that follow it in the causal chain; Figure 3(a) displays an example. Fully recursive models are *exactly identified* (or ‘just identified’).

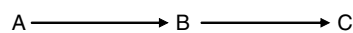


Figure 1 Graphical representation of a basic recursive model

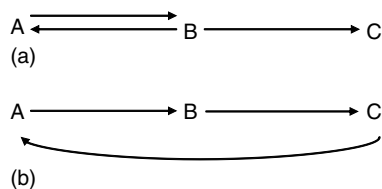


Figure 2 Examples of nonrecursive models; (a) nonrecursive model depicting reciprocal causation between variables A and B and (b) nonrecursive model depicting a feedback loop between variables A and C

It is impossible to disconfirm these models, because they will fit any set of observations perfectly. Moreover, fully recursive models often lack parsimony, being as complex as the observed relationships [3]. For these reasons, fully recursive models are most useful in the context of exploratory data analysis, but they are suboptimal for the testing of theoretically derived predictions.

Models that are recursive (but not fully so), on the other hand, omit one or more direct paths in the causal chain. In other words, some variables only influence other variables indirectly. In Figure 3(b), for instance, variable A influences variable D only by way of variables B and C (i.e., variables B and C *mediate* the effects of A on D). This amounts to a more restrictive model (i.e., constraining the direct relationship from variable A to variable D to zero) than in the fully recursive case, so these not-fully-recursive models are more impressive when they fit well [2, 3]. Recursive models in the behavioral sciences typically fall into this ‘not fully saturated’ category.

In contrast to recursive models, *nonrecursive* models include bidirectional or feedback effects (displayed in Figures 2(a) and 2(b), respectively), and/or they contain correlated disturbances for variables that are part of the same causal chain. Nonrecursive models are intuitively appealing in the behavioral sciences, given that many phenomena in the real world would seem to have mutual influences or feedback loops. However, a major drawback of nonrecursive models is that estimation can be difficult, especially with cross-sectional data. These models are often *underidentified*, meaning that there is more than one

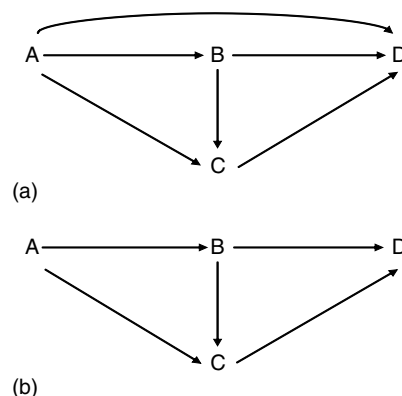


Figure 3 Variants of recursive models; (a) fully recursive model and (b) (not fully) recursive model

2 Recursive Models

possible value for one or more parameters (*see Identification*); that is, multiple estimates fit the data equally well, making it impossible to arrive at a single unique solution [2]. For this reason, nonrecursive models are less common than recursive models in behavioral research.

References

- [1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, John Wiley & Sons, New York.
- [2] Klem, L. (1995). Path analysis, in *Reading and Understanding Multivariate Statistics*, L.G. Grimm & P.R. Yarnold, eds, American Psychological Association, Washington, pp. 65–98.
- [3] MacCallum, R.C. (1995). Model specification: procedures, strategies, and related issues, in *Structural Equation Modeling: Concepts, Issues, and Applications*, R.H. Hoyle, ed., Sage Publications, Thousand Oaks, pp. 16–36.
- [4] Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd Edition, Wadsworth Publishers.

(See also **Structural Equation Modeling: Overview**)

LILIA M. CORTINA

Regression Artifacts

DAVID A. KENNY

Volume 4, pp. 1723–1725

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Regression Artifacts

Regression toward the mean has a long and somewhat confusing history. The term was invented by **Galton** [3] in 1886 when he noted that tall parents tended to have somewhat shorter children. Its presence is readily apparent in everyday life. Movie sequels tend to be worse than the original movies, married couples tend to experience a decline in marital satisfaction over time, and therapy usually results in improvement in symptoms.

The formal definition of regression toward the mean is as follows: Given a set of scores measured on one variable, X, it is a mathematical necessity that the expected value of a score on another variable, Y, will be closer to the mean, when both X and Y are measured in standard deviation units. As an example, if a person were two standard deviation units above the mean on intelligence, the expectation would be that the person would be less than two standard deviation units above the mean on any other variable. Typically, regression toward the mean is presented in terms of the same variable measured at two times, but it applies to any two variables measured on the same units or persons. As discussed by Campbell and Kenny [1], regression toward the mean does not depend on the assumption of linearity, the level of measurement of the variables (i.e., the variables can be dichotomous), or measurement error. Given a less than perfect correlation between X and Y, regression toward the mean is a mathematical necessity. It is not something that is inherently biological or psychological, although it has important implications for both biology and psychology.

Regression toward the mean applies in both directions, from Y to X as well as from X to Y. For instance, Galton [3] noticed that tall children tended to have shorter parents. Regression toward the mean does not imply increasing homogeneity over time because it refers to expected or predicted scores, not actual scores. On average, scores regress toward the mean, but some scores may regress away from the mean and some may not change at all.

Campbell and Kenny [1] discuss several ways to illustrate graphically regression toward the mean. The simplest approach is a **scatterplot**. An alternative is a *pair-link diagram*. In such a diagram, there are two vertical bars, one for X and one for Y. An individual unit is represented by a line that goes from one

bar to the other. A particularly useful method for displaying regression toward the mean is a *Galton squeeze diagram* [1], which is shown in Figure 1. The left axis represents the scores on one measure, the pretest, the right axis represents the means on the second measure, the posttest. The line connecting the two axes represents the score on one measure and the mean score of those on the second measure. Regression toward the mean is readily apparent.

Because regression toward the mean refers to standard scores and an entire distribution of scores, its implications are less certain for raw scores. Consider, for instance, a measure of educational ability on which children are measured. All children receive some sort of intervention, and all are remeasured on that same variable. If the children improve at the second testing (i.e., the mean of the second testing is greater than the mean on the first testing), can we attribute that improvement to regression toward the mean or to the intervention? Without further information, we would be unable to answer the question definitively. If the children in the sample were below the mean in that population at time one and the mean and variance of the scores in the population was the same at both times without any intervention, then it is likely that the change is due to regression toward the mean. Of course, an investigator is not likely to know whether the children in the study are below the population mean and whether the population mean and standard deviation are unchanging.

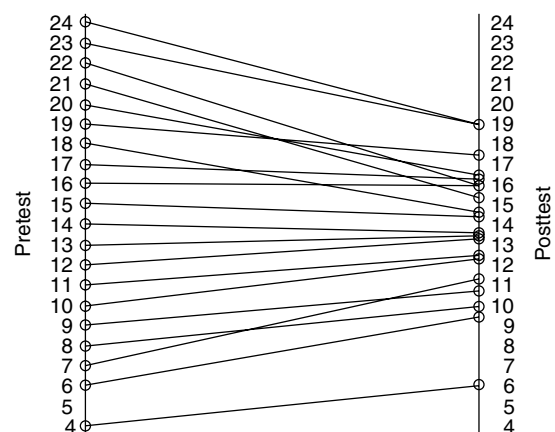


Figure 1 A Galton squeeze diagram illustrating regression toward the mean

2 Regression Artifacts

Because raw scores do not necessarily regress toward the mean, some (e.g., [5]) have argued that regression toward the mean is not a problem in interpreting change. However, it should be realized that regression toward the mean is omnipresent with standard scores and therefore it has implications, uncertain though they are, for the interpretation of raw score change. Certainly, a key signal that regression toward the mean is likely to be a problem is when there is selection of extreme scores from a distribution.

Further complications also arise when two populations are studied over time. Consider two populations whose means and standard deviations differ from each other but their means and standard deviations do not change over time. If at time one, a person from each of the two populations is selected and these two persons have the very same score, at time two the two persons would not be expected to have the same score because each is regressing toward a different mean. It might be that one would be improving and the other worsening. Furby [2] has a detailed discussion of this issue.

Gilovich [4] and others have discussed how it is that lay people fail to take into account regression toward the mean in everyday life. For example, some parents think that punishment is more effective than reward. However, punishment usually follows bad behavior and, given regression toward the mean, the expectation is for improvement. Because reward usually follows good behavior, the expectation is a decline in good behavior and an apparent ineffectiveness of reward. Also, Harrison and Bazerman [6]

discuss the often unnoticed effects of regression toward the mean in organizational contexts.

Regression toward the mean has important implications in prediction. In situations in which one has little information to make a judgment, often the best advice is to use the mean value as the prediction. In essence, the prediction is regressed to the mean.

In summary, regression toward the mean is a universal phenomenon. Nonetheless, it can be difficult to predict its effects in particular applications.

References

- [1] Campbell, D.T. & Kenny, D.A. (1999). *A Primer of Regression Artifacts*, Guilford, New York.
- [2] Furby, L. (1973). Interpreting regression toward the mean in developmental research, *Developmental Psychology* **8**, 172–179.
- [3] Galton, F. (1886). Regression toward mediocrity in hereditary stature, *Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263.
- [4] Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*, Free Press, New York.
- [5] Gottman, J.M., ed. (1995). *The Analysis of Change*, Lawrence Erlbaum Associates, Mahwah.
- [6] Harrison, J.R. & Bazerman, M.H. (1995). Regression toward the mean, expectation inflation, and the winner's curse in organizational contexts, in *Negotiation as a Social Process: New Trends in Theory and Research*, R.M. Kramer & D.M. Messick, eds, Sage Publications, Thousand Oaks, pp. 69–94.

DAVID A. KENNY

Regression Discontinuity Design

WILLIAM R. SHADISH AND JASON K. LUELLEN

Volume 4, pp. 1725–1727

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Regression Discontinuity Design

Thistlewaite and Campbell [9] published the first example of a regression discontinuity design, the only **quasi-experimental design** that can yield unbiased estimates of the treatment effects [3, 4, 6, 7]. More extensive treatments of this design appear in [2], [8], [10], [11], and [12].

With the regression discontinuity design, the experimenter assigns participants to treatment conditions using a cutoff score on an assignment variable that is measured prior to treatment and is at least an ordinal measurement scale. The assignment variable is often a measure of need or merit, meaning that those who need a treatment or merit a reward the most are also the most likely to receive it. This feature of the design provides an ethical alternative when objections occur to randomly assigning needy or meritorious participants to a treatment, or to randomly depriving others of those same treatments.

The simplest design places units scoring on one side of the cutoff into the treatment condition, and those on the other side into the control condition. This design is diagrammed as shown in Table 1:

Table 1 A diagram of the regression discontinuity design

O_A	C	X	O
O_A	C		O

where O_A is the assignment variable, C indicates that participants are assigned to conditions using a cutoff score, X denotes treatment, O is a posttest observation, and the position of these letters from left to right indicates the time sequence in which they occur. If j is a cutoff score on O_A , then any participant scoring greater than or equal to j is in one group, and anything less than j is in the other. For example, suppose an education researcher implements a treatment program to improve math skills of third graders, but resource restrictions prohibited providing the treatment to all third-graders. Instead, the researcher could administer all third-graders a test of math achievement, and then assign those below a cutoff score on that test to the treatment, with

those above being in the control group. If we draw a **scatterplot** of the relationship between scores on the assignment variable (horizontal axis) and scores on a math outcome measure (vertical axis), and if the training program improved math skills, the points in the scatterplot on the treatment side of the cutoff would be displaced upwards to reflect higher posttest math scores. And a regression line (*see Regression Models*) would show a discontinuity at the cutoff, that is, the size of the treatment effect. If the treatment had no effect, a regression line would not have this discontinuity. Shadish et al. [8] present many real-world studies that used this design. Many variants of the basic regression discontinuity design exist that allow comparing more than two conditions, that combine regression discontinuity with random assignment or with a quasi-experiment, or that improve statistical **power** [8, 10].

In most other quasi-experiments, it is difficult to rule out plausible alternative explanations for observed treatment effects – what Campbell and Stanley [1] call threats to **internal validity**. Threats to internal validity are less problematic with the regression discontinuity design, especially when the change in the regression line occurs at the cutoff and is large. Under such circumstances, it is difficult to conceive of reasons other than treatment that such a change in the outcome measure would occur for those immediately to one side of the cutoff, but not for those immediately to the other side.

Statistical regression (*see Regression to the Mean*) may seem to be a plausible threat to internal validity with the regression discontinuity design. That is, because groups were formed from the extremes of the distribution of the assignment variable, a participant scoring high on the assignment variable will likely not score as high on the outcome measure, and a participant scoring low on the assignment variable will likely not score as low on the outcome measure. But this regression does not cause a discontinuity in the regression line at the cutoff; it simply causes the regression line to turn more horizontal.

Selection is not a threat even though groups were selected to be different. Selection can be controlled statistically because the selection mechanism is fully known and measured. Put intuitively, the small difference in scores on the assignment variable for participants just to each side of the cutoff is not likely to account for a large difference in their scores on the

2 Regression Discontinuity Design

outcome measure at the cutoff. *History* is a plausible threat if other interventions use the same cutoff for the same participants, which is usually unlikely. Because both groups are administered the same test measures, *testing*, per se, would not likely result in differences between the two groups. For the threat of *instrumentation* to apply, changes to the measurement instrument would need to occur exactly at the cutoff value of the assignment variable. *Attrition*, when correlated with treatment assignment, is probably the most likely threat to internal validity with this design.

In the regression discontinuity design, like a randomized experiment, the selection process is completely known and perfectly measured – the condition that must be met in order to successfully adjust for selection bias and obtain unbiased estimates of treatment effects. The selection process is completely known, assignment to conditions is based solely on whether a participant's score on the assignment variable is above or below the cutoff. No other variables, hidden or observed, determine assignment. Selection is perfectly measured because the pretest is strictly used to measure the selection mechanism. For example, if IQ is the assignment variable, although IQ scores imperfectly measure global intelligence, they have no error as a measure of how participants got into conditions.

The design has not been widely utilized because of a number of practical problems in implementing it. Two of those problems are unique to the regression discontinuity design. Treatment effect estimates are unbiased only if the functional form of the relationship between the assignment variable and the outcome measure is correctly specified, including nonlinearities and interactions. And statistical power of the regression discontinuity design is always lower than a comparably sized randomized experiment. Other problems are shared in common with a randomized experiment. Treatment assignment must adhere to the cutoff, just as assignment must adhere to a randomized selection process. In both cases, treatment professionals are not allowed to use judgment in assigning treatment. Cutoff-based assignments may be difficult to implement when the rate of participants entering the study is too slow or too fast [5]. Treatment crossovers may occur when participants assigned to treatment do not take it, or participants

assigned to control end up being treated. Despite these difficulties, the regression discontinuity design holds a special place among cause-probing methods and deserves more thoughtful consideration when researchers are considering design options.

References

- [1] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.
- [2] Cappelleri, J.C. (1991). *Cutoff-Based Designs in Comparison and Combination with Randomized Clinical Trials*, Unpublished doctoral dissertation, Cornell University, Ithaca, New York.
- [3] Goldberger, A.S. (1972a). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*, (Discussion paper No. 123), University of Wisconsin, Institute for Research on Poverty, Madison.
- [4] Goldberger, A.S. (1972b). *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*, (Discussion paper), University of Wisconsin, Institute for Research on Poverty, Madison.
- [5] Havassy, B. (1988). *Efficacy of Cocaine Treatments: A Collaborative Study*, (NIDA Grant Number DA05582), University of California, San Francisco.
- [6] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [7] Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics* **2**, 1–26.
- [8] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.
- [9] Thistlewaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment, *Journal of Educational Psychology* **51**, 309–317.
- [10] Trochim, W.M.K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, California, Sage Publications.
- [11] Trochim, W.M.K. (1990). The regression discontinuity design, in *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, L. Sechrest, E. Perrin & J. Bunker, eds, Public Health Service, Agency for Health Care Policy and Research, Rockville, pp. 199–140.
- [12] Trochim, W.M.K. & Cappelleri, J.C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials, *Controlled Clinical Trials* **13**, 190–212.

WILLIAM R. SHADISH AND JASON K. LUELLEN

Regression Model Coding for the Analysis of Variance

PATRICIA COHEN

Volume 4, pp. 1727–1729

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Regression Model Coding for the Analysis of Variance

Using a **multiple linear regression** (MR) program to carry out **analysis of variance** (ANOVA) has certain advantages over use of an ANOVA program. They use the same statistical model (ordinary least squares, OLS), and thus make the same assumptions (normal distribution of population residuals from the model). Ordinarily, one's conclusions from an ANOVA will be the same as those from MR for the same data. However, with MR, you can code directly for tests of study hypotheses, whereas ANOVA can require a two-step process, beginning with tests of the significance of model 'factors', followed by specific tests relevant to your hypotheses. Second, MR permits using as many covariates as needed for substantive questions, including potential interactions among covariates. Third, MR permits tests of interactions of covariates with research factors, whereas **analysis of covariance** assumes such interactions to be absent in the population. Fourth, MR allows elimination of improbable and uninterpretable higher-order interactions in complex designs, improving the statistical power of the other tests. Finally, MR allows alternative treatments of unequal group *ns*, whereas standard ANOVA programs select a particular one (*see Type I, Type II and Type III Sums of Squares*). However, OLS MR programs handle **repeated measures analysis of variance** awkwardly at best.

How do you do it? An ANOVA study design consists of one or more research 'factors'. Each factor has two or more groups (categories) with participants in each group. Thus, one research factor (A) can consist of the three levels of some experimentally induced independent variable (IV). Another factor (B) can consist of four different ages, ethnicities, or litters of origin, and a third factor (C) can consist of two different testing times during the day. Thus, this study would consist of $3(A) \times 4(B) \times 2(C) = 24$ different combinations or conditions; each combination includes some fraction of the total *N* participants on whom we have measured a dependent variable (DV). In an ANOVA, these three factors would be tested for significant differences among group means on the DV by the following *F* tests:

A with $g_A - 1 = 2$ *df* (degrees of freedom)
B with $g_B - 1 = 3$ *df*,
C with $g_C - 1 = 1$ *df*,
A $\times$ B interaction with 2 *df* $\times$ 3 *df* = 6 *df*,
A $\times$ C interaction with 2 *df* $\times$ 1 *df* = 2 *df*,
B $\times$ C interaction with 3 *df* $\times$ 1 *df* = 3 *df*,
A $\times$ B $\times$ C interaction with 2 *df* $\times$ 3 *df* $\times$ 1 *df* = 6 *df*, with a total of $2 + 3 + 1 + 6 + 2 + 3 + 6 = 23$ *df*.

In a MR analysis of the same data, each of these 23 *df* is represented by a unique independent variable. These variables are usually created by coding each research factor and their interactions using one or more of the following coding systems.

Dummy variable coding: To dummy variable code each factor with *g* groups, each of the *g* - 1 groups is coded 1 on one and only one of the coded variables representing the factor and 0 on all the others. The 'left out' group is coded 0 on all *g* - 1 variables. When these *g* - 1 variables are entered into the regression analysis, (counterintuitively) the regression coefficient for each of the variables reflects the mean difference between the group coded 1 and the 'left out' group consistently coded 0. Thus, a statistical test of the coefficient provides a test for that difference. For this reason, dummy variable coding is most appropriate when there is a group that is logical to compare with all other groups in the research factor. The *F*-value for the contribution to *R*² of the *g* - 1 variables is precisely the same as it would have been for the ANOVA. However, the statistical **power** for the comparison of certain groups with the control group can be greater when the hypothesis of a difference from the control group is weak for some other groups. This occurs because the tests of group-control comparisons will not necessarily require an overall significant *F* test prior to the individual comparisons.

Effects coding of the *g* - 1 variables results in a contrast of each group's mean with the mean of the set of group means. This coding method is similar to dummy variable coding with an important difference and a very different interpretation. Instead of one group being selected as the reference group and coded 0 on all *g* - 1 variables, one group is coded -1 rather than 0 on every variable. This group should be selected to be the group of *least* interest because now the *g* - 1 regression coefficients contrast each of the other group means with the mean of the *g* group

2 Regression Model Coding for the Analysis of Variance

Table 1 Alternative codes for three variables representing a four-group factor

Research factor	Dummy variable	Effects	Contrasts A	Contrasts B
Group L	1, 0, 0	1, 0, 0	+0.5, -0.5, 0	+0.5, +0.5, +0.5
Group M	0, 1, 0	-1, -1, -1 (least important)	+0.5, +0.5, 0	+0.5, -0.5, -0.5
Group N	0, 0, 1	0, 1, 0	-0.5, 0, -0.5	-0.5, +0.5, -0.5
Group P	0, 0, 0 (reference)	0, 0, 1	-0.5, 0, +0.5	-0.5, -0.5, +0.5

means. No explicit comparison of the ‘-1’ group is represented in the regression coefficients, although it makes the usual contribution to the overall test of the variance in the group means, which is exactly equivalent to any of the other coding methods. (See Table 1 for dummy and effects codes for a four-group factor).

Contrast coding is an alternative method of coding the $g - 1$ variables representing the research factor. Contrast coding is actually not a single method, but a set of methods designed to match at least some of the study hypotheses. Thus, again, these tests are carried out directly in MR, whereas an ANOVA will typically provide an omnibus F test of a research factor that will need to be followed up with planned comparisons. Let us compare two alternative contrast codes representing a four-group research factor, comprised of groups L, M, N, and P, for which we will need three variables.

Suppose our primary interest is in whether groups L and M are different from groups N and P. In order for the regression coefficient to reflect this contrast, we will make the difference between the variable codes for these groups = 1. Thus, we code L and M participants = 1.0 and N and P participants = 0 on the first IV_1 . At this point, we invoke the first rule of contrast coding: let the sum of the codes for each variable = 0, and recode L = M = +0.5 and N = P = -0.5. Our second major interest is in whether there are differences between L and M. Therefore, for IV_2 : L = -0.5 and M = +0.5, and, since we want to leave them out of consideration, N = P = 0.

Our third major interest is in whether there are differences between N and P. Therefore, for IV_3 : N = -0.5, P = +0.5, and L = M = 0. At this point, we invoke the second rule of contrast coding: let the sum of the products of each pair of variables = 0. The product of the codes for IV_1 and IV_2 is $(0.5 \times -0.5) = -0.25$ for group L, $(0.5 \times 0.5) = +0.25$ for M, and 0 for N and P. The code product of IV_1 and IV_3 is $(\pm 0.5 \times 0) = 0$ for L and M,

$(-0.5 \times -0.5) = 0.25$ for N and $(-0.5 \times 0.5) = -0.25$ for P which also sum to 0. The $IV_2 - IV_3$ products similarly sum to 0. Finally, although the tests of statistical significance do not require it, a third rule also is useful: for each variable, let the difference between the negative weights and the positive weights = 1. Under these conditions, with these three variables in the regression equation predicting the dependent variable, the first coefficient equals the difference between combined groups L and M as compared to groups N and P, with its appropriate standard error (SE) and significance (t or F) test. The second regression coefficient equals the difference between groups L and M, with its appropriate SE and test, and the third coefficient equals the difference between groups N and P.

Suppose our study had a different content that made a different set of contrasts appropriate to our hypotheses. Perhaps the first contrast was the same L and M versus N and P as above, but the second was more appropriately a contrast of L and N (= +0.5) with M and P (= -0.5). The third variable that will satisfy the code rules would combine L = P = +0.5 and M = N = -0.5 (see Table 1).

A given study can use whatever combination of dummy variable, effects, and contrast codes that fits the overall hypotheses for different research factors. The coded variables representing the factors are then simply multiplied to create the variables that will represent the interactions among the factors (*see Interaction Effects*). Thus, if a given study participant had 1, 0, 0 on the three dummy variables representing the first factor and -0.5, -0.5, 0.25 on three contrast variables representing the second research factor, there would be nine variables representing the interaction between these factors. This participant would be scored -0.5, -0.5, 0.25, 0, 0, 0, 0, 0, 0. Because the first factor is dummy variable coded, all interactions are assessed as differences on the second factor contrasts between a particular group coded 1 and the reference group. The first three variables, therefore, reflect these differences for the group including this

participant and the reference group on the three contrasts represented by the codes for the second research factor. The next six variables provide the same information for the other two groups coded 1 on the first factor variables 2 and 3, respectively.

Suppose that sample sizes in study cells representing the different combinations of research factors are not all equal. The analysis used in computer ANOVA programs in this case is equivalent to MR using effects codes, with regression coefficients contrasting each group's mean with the equally weighted mean of the g group means. Alternatively, rarely the investigator wishes the statistical tests to take the different cell sizes into account because their numbers appropriately represent the population of interest. Such a 'weighted means' set of comparisons involves using ratios of group sample sizes as coefficients (see Cohen, Cohen, West, & Aiken, 2003, p. 329 for details). (See **Type I, Type II and Type III Sums of Squares**.)

Repeated measure ANOVAs have multiple error terms, each of which would need to be identified in a separate MR using different functions of the original DV. Thus, MR is not generally employed. If the advantages of a regression model are desired, repeated DVs are better handled by multilevel regression analysis, in which, however, the statistical model is Maximum Likelihood rather than OLS. One advantage of this method is that it does not require that every trial is available for every individual. Codes of 'fixed effects' in such cases may employ the same set of options as described here. 'Nested' ANOVA designs in which groups corresponding to one research factor (B) are different for different levels of another research factor (A) are also currently usually managed in multilevel MR.

PATRICIA COHEN

Regression Models

MICHAEL S. LEWIS-BECK

Volume 4, pp. 1729–1731

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Regression Models

Regression is the most widely used procedure for analysis of observational, or nonexperimental, data in the behavioral sciences. A regression model offers an explanation of behavior in the form of an equation, tested against systematically gathered empirical cases. The simplest regression model may be written as follows:

$$Y = a + bX + e \quad (1)$$

where Y = the dependent variable (the phenomenon to be explained) and X = the independent variable (a factor that helps explain, or at least predict, the dependent variable). The regression coefficient, b , represents the link between X and Y . The constant value, a , is always added to the prediction of Y from X , in order to reduce the level of error. The last term, e , represents the error that still remains after predicting Y .

The regression model of Eq. 1 is bivariate, with one independent variable and one dependent variable. A regression model may be multivariate, with two or more independent variables and one dependent variable. Here is a **multiple regression model**, with two independent variables:

$$Y = a + bX + cZ + e \quad (2)$$

where Y = the dependent variable, X and Z = independent variables, a = the constant, b and c = regression coefficients, and e = error. This multiple regression model has advantages over the bivariate regression model, first, because it promises a more complete account of the phenomenon under study, and second, because it more accurately estimates the link between X and Y .

The precise calculations of the link between the independent and dependent variables come from application of different estimation techniques, most commonly that of ordinary least squares (OLS) (*see Least Squares Estimation*). When researchers 'run regressions', the assumption is that the method of calculation is OLS unless otherwise stated. The least squares principle minimizes the sum of squared errors in the prediction of Y , from a line or plane. Since the principle is derived from the calculus, the sum of these squared errors is guaranteed 'least' of all possibilities, hence the phrase 'least squares'. If

a behavioral scientist says 'Here is my regression model and here are the estimates', most likely the reference is to results from a multiple regression equation estimated with OLS.

Let us consider a hypothetical, but not implausible, example. Suppose a scholar of education policy, call her Dr. Jane Brown, wants to know why some American states spend more money on secondary public education than others. She proposes the following regression model, which explains differential support for public secondary education as a function of median income, elderly population, and private schools in the state:

$$Y = a + bX + cZ + dQ + e, \quad (3)$$

where Y = the per pupil dollar expenditures (in thousands) for secondary education by the state government, X = median dollar income of those in the state workforce; Z = people over 65 years of age (in percent); Q = private school enrollment in high schools of the state (scored 1 if greater than 15%, 0 otherwise).

Her hypotheses are that more income means more expenditure, that is, $b > 0$; more elderly means less expenditure, that is, $c < 0$; and a substantial private school enrollment means less expenditure, that is, $d < 0$. To test these hypotheses, she gathers the variable scores on the 48 mainland states from a 2003 statistical yearbook, enters these data into the computer and, using a popular statistical software package, estimates the equation with OLS. Here are the results she gets for the coefficient estimates, along with some common supporting statistics:

$$Y = 1.31 + .50*X - 0.29Z - 4.66*Q + e$$

$$(0.44) \quad (3.18) \quad (1.04) \quad (5.53)$$

$$R^2 = 0.55 \quad N = 48 \quad (4)$$

where Y , X , Z , Q and e are defined and measured as with Eq. 3; the estimates of coefficients a , b , c , and d are presented; e is the remaining error; the numbers in parentheses below these coefficients are absolute t -ratios; the * indicates statistical significance at the 0.05 level; the R^2 indicates the coefficient of multiple determination; N is the size of the sample of 48 American states, excluding Alaska and Hawaii.

These findings suggest that, as hypothesized, income positively affects public high school spending, while the substantial presence of private schools

negatively affects it. This assertion is based on the signs of the coefficients b and d , respectively, $+$ and $-$, and on their statistical significance at 0.05. This level of statistical significance implies that, if Dr. Brown claims income matters for education, the claim is probably right, with 95% certainty. Put another way, if she did the study 100 times, only 5 of those times would she not be able to conclude that income related to education. (Note that the t -ratios of both coefficients exceed 2.0, a common rule-of-thumb for statistical significance at the 0.05 level.) Further, contrary to hypothesis, a greater proportion of elderly in the state does not impact public high school spending. This assertion is based on the lack of statistical significance of coefficient c , which says we cannot confidently reject the possibility that the percentage elderly in a state is not at all related to these public expenditures. Perhaps, for instance, with somewhat different measures, or a sample from another year, we might get a very different result from the negative sign we got in this sample.

Having established that there is a link between income and spending, what is it exactly? The coefficient $b = 0.50$ suggests that, on average, a one-unit increase in X (i.e., \$1000 more in income) leads to a 0.50 unit increase in Y (i.e., a \$500 increase in education expenditure). So we see that an income dollar translates strongly, albeit not perfectly, into an education dollar, as Dr. Brown expected. Further, we see that states that have many private schools (over 15%) are much less likely to fund public high school education for, on average, it is $4.66 \times 1000 = \$4660$ per pupil less in those states.

How well does this regression model, overall, explain state differences in public education spending? An answer to this question comes from the R^2 , a statistic that reports how well the model fits the data. Accordingly, the R^2 here indicates that 55% of the variation in public high school expenditures across the states can be accounted for. Thus, the model tells us a lot about why states are not the same on this variable. However, almost half the variation ($1 - 0.55 = 0.45$) is left unexplained, which means an important piece of the story is left untold. Dr. Brown should reconsider her explanation, perhaps including more variables in the model to improve its theoretical specification.

With a classical regression model, the OLS coefficients estimate real-world connections between variables, assuming certain assumptions are met. The assumptions include no specification error, no measurement error, no perfect multicollinearity, and a well-behaved error term. When these assumptions are met, the least squares estimator is BLUE, standing for Best Linear Unbiased Estimator. Among other things, this means that, on average, the estimated coefficients are true. Once assumptions are violated, they may be restored by variable transformation, or they may not. For example, if, in a regression model, the dependent variable is dichotomous (say, values are 1 if some property exists and 0 otherwise), then no transformation will render the least squares estimator BLUE. In this case, an alternative estimator, such as **maximum likelihood (MLE)**, is preferred.

The need to use MLE, usually in order to achieve more efficient estimation, leads to another class of regression models, which includes logistic regression (when the dependent variable is dichotomous), polytomous **logistic regression** (when the dependent variable is ordered), or poisson regression (*see Generalized Linear Models (GLM)*) (when the dependent variable is an event count). Other kinds of regression models, which may use a least squares solution, are constructed to deal with other potential assumption violations, such as weighted least squares (to handle heteroskedasticity), local regression (to inductively fit a curve), censored regression (to deal with truncated observations on the dependent variable), seemingly unrelated regression (for two equations related through correlated errors but with different independent variables), spatial regression (for the problem of geographic autocorrelation), spline regression when there are smooth turning points in a line (*see Scatterplot Smoothers*), or stepwise regression (for selecting a subset of independent variables that misleadingly appears to maximize explanation of the dependent variable). All these variations are regression models of some sort, united by the notion that a dependent variable, Y , can be accounted for some independent variable(s), as expressed in an equation where Y is a function of some $X(s)$.

Further Reading

Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition, Lawrence Erlbaum, Mahwah.

Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots*, Sage Publications, Thousand Oaks.
Kennedy, P. (2003). *A Guide to Econometrics*, 5th Edition, MIT Press, Cambridge.
Lewis-Beck, M.S. (1980). *Applied Regression: An Introduction*, Sage Publications, Beverly Hills.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks.

MICHAEL S. LEWIS-BECK

Regression to the Mean

RANDOLPH A. SMITH

Volume 4, pp. 1731–1732

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Regression to the Mean

Francis Galton discovered regression to the mean in the 1800s; he referred to it as a 'tendency toward mediocrity' [1, p. 705]. In 1877, his work with sweet peas revealed that large parent seeds tended to produce smaller seeds and that small parent seeds tended to produce larger seeds [3]. Later, Galton confirmed that the same principle operated with human height: Tall parents tended to produce children who were shorter and short parents tended to have children who were taller. In all these examples, subsequent generations moved closer to the mean (regressed toward the mean) than the previous generation. These observations led to the definition of regression to the mean: extreme scores tend to become less extreme over time. Regression is always in the direction of a population mean [2].

Regression to the mean may lead people to draw incorrect causal conclusions. For example, a parent who uses punishment may conclude that it is effective because a child's bad behavior becomes better after a spanking. However, regression to the mean would predict that the child's behavior would be better shortly after bad behavior.

In the context of measurement, regression to the mean is probably due to measurement error. For example, if we assume that any measurement is made up of a true score + error, a high degree of error (either positive or negative) on one measurement should be followed by a lower degree of error on a subsequent measurement, which would result in a second score that is closer to the mean than the first score. For example, if a student guesses particularly well on a multiple-choice test, the resulting score will be high. On a subsequent test, the same student will likely guess correctly at a lower rate, thus resulting in a lower score. However, the lower score is due to error in measurement rather than the student's knowledge base.

Regression to the mean can be problematic in experimental situations; Cook and Campbell [2] listed it ('statistical regression', p. 52) as a threat to internal validity. In a repeated measures or prepost design, the experimenter measures the participants more than once (*see Repeated Measures Analysis of Variance*). Regression to the mean would predict that low scorers would tend to score higher on the second measurement and that high scorers would tend to score lower on the second measurement. Thus, a change in scores could occur as a result of a statistical artifact rather than because of an independent variable in an experiment. This potential problem becomes even greater if we conduct an experiment in which we use pretest scores to select our participants. If we choose the high scorers in an attempt to decrease their scores (e.g., depression or other psychopathology) or if we choose low scorers in an attempt to increase their scores (e.g., sociability, problem-solving behavior), regression to the mean may account for at least some of the improvement we observe from pre- to posttest [4].

Although regression to the mean is a purely statistical phenomenon, it can lead people to draw incorrect conclusions in real life and in experimental situations.

References

- [1] Bartoszyński, R. & Niewiadomska-Bugaj, M. (1996). *Probability and Statistical Inference*, Wiley, New York.
- [2] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Houghton Mifflin, Boston.
- [3] Heyde, C.C. & Seneta, E. (2001). *Statisticians of the Centuries*, Springer, New York.
- [4] Hsu, L.M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy, *Journal of Consulting and Clinical Psychology* **63**, 141–144.

RANDOLPH A. SMITH

Relative Risk

BRIAN S. EVERITT

Volume 4, pp. 1732–1733

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Relative Risk

Quantifying risk and assessing risk involve the calculation and comparison of probabilities, although most expressions of risk are compound measures that describe both the probability of harm and its severity (*see* **Risk Perception**). The way risk assessments are presented can have an influence on how the associated risks are perceived. You might, for example, be worried if you heard that occupational exposure at your place of work doubled your risk of serious disease compared to the risk working at some other occupation entailed. You might be less worried, however, if you heard that your risk had increased from one in a million to two in a million. In the first case, it is relative risk that is presented, and in the second, it is an absolute risk.

Relative risk is generally used in medical studies investigating possible links between a risk factor and a disease. Formally relative risk is defined as

$$\text{Relative risk} = \frac{\text{incidence rate among exposed}}{\text{incidence rate among nonexposed}}. \quad (1)$$

Thus, a relative risk of five, for example, means that an exposed person is five times as likely to have the disease than a person who is not exposed.

Relative risk is an extremely important index of the strength of the association (*see* **Measures of Association**) between a risk factor and a disease (or other outcome of interest), but it has no bearing on the probability that an individual will contract the disease. This may explain why airline pilots who presumably have relative risks of being killed in airline crashes that are of the order of a thousandfold greater than the rest of us occasional flyers can still sleep easy in their beds. They know that the absolute risk of their being the victim of a crash remains extremely small.

BRIAN S. EVERITT

Reliability: Definitions and Estimation

ALAN D. MEAD

Volume 4, pp. 1733–1735

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Reliability: Definitions and Estimation

Reliability refers to the degree to which test scores are free from error. Perfectly reliable scores would contain no error and completely unreliable scores would be composed entirely of error. In **classical test theory**, reliability is defined precisely as the ratio of the true score variance to the observed score variance and, equivalently, as one minus the ratio of error score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}, \quad (1)$$

where $\rho_{XX'}$ is the reliability, σ_X^2 is the observed score variance, σ_E^2 is the error score variance, and σ_T^2 is the true score variance (see [2, 4]). If $\sigma_E^2 = 0$, then $\rho_{XX'} = 1$, while if $\sigma_E^2 = \sigma_X^2$ (i.e., error is at its maximum possible value), then $\rho_{XX'} = 0$. The symbol for reliability is the (hypothetical) correlation of a test score, X , with an independent administration of itself, X' .

Estimates of Reliability

True and error scores are not observable, so reliability must be estimated. Several different kinds of estimates have been proposed. Each estimate may suffer from various sources of error and, thus, no single estimate is best for all purposes. **Generalizability theory** may be used to conduct a generalizability study that can be helpful in understanding various sources of error, including interacting influences [1].

Test-retest reliability estimates require administration of a test form on two separate occasions separated by a period of time (the 'retest period'). The correlation of the test scores across the two administrations is the reliability estimate. The retest period may be quite short (e.g., minutes) or quite long (months or years). Sometimes, different retest periods are chosen, resulting in multiple test-retest estimates of reliability. For most tests, reliabilities decrease with longer retest periods, but the *rate of decline* decreases as the retest period becomes longer.

Test-retest reliability estimates may be biased because the same form is used on two occasions

and test-takers may recall the items and answers from the initial testing ('memory effects'). Also, during the retest period, it is possible that the test-taker's true standing on the test could change (due to learning, maturation, the dynamic nature of the assessed trait, etc.). These 'maturation effects' might be a cause for the reliability estimate to underestimate the reliability at a single point in time. As a practical matter, test-retest reliability estimates entail the costs and logistical problems of two independent administrations.

Parallel forms estimates of reliability require the preparation of parallel forms of the test. By definition, perfectly parallel forms have equal reliability but require different questions. Both forms are administered to each test-taker and the correlation of the two scores is the estimate of reliability.

Parallel forms reliability estimates eliminate the 'memory effects' concerns that plague test-retest estimates but there still may be 'practice effects' and if the two forms are not administered on the same occasion, 'maturation effects' may degrade the estimate. As a partial answer to these concerns, administration of forms is generally counterbalanced. However, perhaps the most serious problem with this estimate can occur when the forms are substantially nonparallel. Reliability will be misestimated to the degree that the two forms are not parallel.

Split-half reliability estimates are computed from a single administration by creating two (approximately) parallel halves of the test and correlating them. This represents the reliability of half the test, and the *Spearman-Brown formula* is used to 'step up' the obtained correlation to estimate the reliability of the entire form.

Split-half reliability estimates retain many of the same strengths and pitfalls as parallel forms estimates while avoiding a second administration. To the extent that the method of dividing the items into halves does not create parallel forms, the reliability will be underestimated – because the lack of parallelism suppresses their correlation and also because the 'stepping up' method assumes essentially parallel forms. Also, different splitting methods will produce differing results. Common splitting methods include random assignment; odd versus even items; first-half versus second-half; and methods based on content or statistical considerations.

The *coefficient alpha* and *Kuder-Richardson* methods – commonly referred to generically as

2 Reliability: Definitions and Estimation

internal consistency methods – are very common estimates computed from single administrations. These methods are based on the assumption that the test measures a single trait and each item is essentially parallel to all other items. These methods compare the common covariance among items to the overall variability. Greater covariance among the items leads to higher estimates of reliability. Coefficient alpha is also commonly interpreted as the theoretical mean of all possible split-half estimates (based on all possible ways that the halves could be split).

Internal consistency methods make relatively stronger assumptions about the underlying test items and suffer from inaccuracies when these assumptions are not met. When the test items are markedly nonparallel, these reliability estimates are commonly taken to be lower-bound estimates of the true reliability. However, because these methods do not involve repeated administrations, they may be quite different from test-retest or parallel forms estimates and are unsuited to estimating the stability of test scores over time.

Other Reliability Topics

Standard Error of Measurement. The *standard error of measurement* (SEM) is the standard deviation of the error score component. The SEM is very important because it can be used to characterize the degree of error in individual or group scores. The SEM is

$$SEM = \sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}}. \quad (2)$$

For users of test information, the SEM may be far more important and useful than the reliability. By assuming that the error component of a test score is approximately normally distributed, the SEM can be used to construct a true score **confidence interval** around an observed score. For example, if an individual's test score is 10 and the test has an SEM of 1.0, then the individual's true score is about 95% likely to be between 8 and 12 (or more correctly, 95% of the candidate confidence bands formed in this way will contain the unknown true scores of candidates).

The interpretation of the SEM should incorporate the kind of reliability estimate used. For example, a common question raised by test-takers is how they might score if they retest. Confidence intervals using

a test-retest reliability are probably best suited to answering such questions.

Reliability is *not* Invariant Across Subpopulations.

Although reliability is commonly discussed as an attribute of the test, it is influenced by the variability of observed scores (σ_X^2) and the variability of true scores (σ_T^2). The reliability of a test when used with some subpopulations will be diminished if the subpopulation has a lower variability than the general population. For example, an intelligence test might have a high reliability when calculated from samples drawn from the general population but when administered to samples from narrow subpopulations (e.g., geniuses), the reduced score variability causes the reliability to be attenuated; that is, the intelligence test is less precise in making fine distinctions between geniuses as compared to ranking members representative of the breadth of the general population.

As a consequence, test developers should carefully describe the sample used to compute reliability estimates and test users should consider the comparability of the reliability sample to their intended population of test-takers.

The Spearman-Brown Formula. The Spearman-Brown formula provides a means to estimate the effect of lengthening or shortening a test:

$$\rho_{XX'}^* = \frac{n\rho_{XX'}}{1 + (n-1)\rho_{XX'}}, \quad (3)$$

where $\rho_{XX'}^*$ is the new (estimated) reliability, n is the fraction representing the change in test length ($n = 0.5$ implies halving test length, $n = 2$ implies doubling) and $\rho_{XX'}$ is the current reliability of the test. A critical assumption is that the final test is essentially parallel to the current test (technically, the old and new forms must be 'tau-equivalent'). For example, if a 10-item exam were increased to 30 items, then the Spearman-Brown reliability estimate for the 30-item exam would only be accurate to the extent that the additional 20 items have the same characteristics as the existing 10.

True Score Correlation. The correlation between the observed score and the true score is equal to the square root of the reliability: $\rho_{XT} = \sqrt{(\rho_{XX'})}$. If the classical test theory assumptions regarding the

independence of error score components are true, then the observed test score, X , cannot correlate with any variable more highly than it does with its true score, T . Thus, the square root of the reliability is considered an upper bound on correlations between test scores and other variables. This is primarily a concern when the test score reliability is low. A correlation of 0.60 between a test score and a criterion may be quite high if the reliability of the test score is only about 0.40. In this case, the observed correlation of 0.60 is about the maximum possible ($\sqrt{0.40}$).

These considerations give rise to the standard formula for estimating the correlation between two variables as if they were both perfectly reliable:

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'} \rho_{YY'}}}, \quad (4)$$

where $\rho_{T_X T_Y}$ is the estimated correlation between the true scores of variables X and Y , ρ_{XY} is the observed correlation between variables X and Y , $\rho_{XX'}$ is the reliability of score X , and $\rho_{YY'}$ is the reliability of score Y . This hypothetical relationship is primarily of interest when comparing different sets of variables without the obscuring effect of different reliabilities and when assessing the correlation of two constructs without the obscuring effect of the unreliability of the measures.

Reliability for Speeded Tests

Some estimates of reliability are not applicable to highly speeded tests. For example, split-half and internal consistency estimates of reliability are inappropriate for highly speeded tests. Reliability methods that involve retesting are probably best, although they

may be subject to practice effects. A generalizability study may be particularly helpful in identifying factors that influence the reliability of speeded tests.

Reliability and IRT

Reliability is a concept defined in terms of classical test theory. Modern **item response theory (IRT)** provides much stronger and more flexible results (see, for example, [3]). Such results reveal that reliability and SEM are simplifications of the actual characteristics of tests. In place of reliability and SEM, IRT provides *information functions* for tests and items. The inverse of the test information function provides the standard error of measurement *at a particular point on the latent trait* (see **Latent Variable**). IRT analysis generally reveals that scores near the middle of the score distribution are considerably more reliable (i.e., have lower SEM) than scores in the tails of the distribution. This often means that ordering test-takers in the upper and lower percentiles is unreliable, perhaps to the point of being arbitrary.

References

- [1] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.
- [2] Gulliksen, H. (1950). *Theory of Mental Test Scores*, Wiley, New York.
- [3] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.
- [4] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

ALAN D. MEAD

Repeated Measures Analysis of Variance

BRIAN S. EVERITT

Volume 4, pp. 1735–1741

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Repeated Measures Analysis of Variance

In many studies carried out in the behavioral sciences and related disciplines. The response variable is observed on each subject under a number of different conditions. For example, in an experiment reported in [6], the performance of field-independent and field-dependent subjects (twelve of each type) in a reverse Stroop task was investigated. The task required reading of words of two types, color and form names, under three cue conditions – normal, congruent, and incongruent. For instance, the word ‘yellow’ displayed in yellow would be congruent, whereas ‘yellow’ displayed in blue would be incongruent. The dependent measure was the time (msec) taken by a subject to read the words. The data are given in

Table 1. Note that each subject in the experiment has six time measurements recorded. The response variable is time in milliseconds.

The data in Table 1 contain repeated measurements of a response variable on each subject. Researchers, typically, adopt the repeated measures paradigm as a means of reducing error variability and/or as a natural way of assessing certain phenomena (e.g., developmental changes over time, and learning and memory tasks). In this type of design, effects of experimental factors giving rise to the repeated measurements (the so-called *within subject factors*; word type and cue condition in Table 1) are assessed relative to the average response made by a subject on all conditions or occasions. In essence, each subject serves as his or her own control, and, accordingly, variability caused by differences in average response time of the subjects is eliminated from the extraneous error variance. A consequence of this

Table 1 Field independence and a reverse Stroop task

Subject	Form names			Color names		
	Normal condition	Congruent condition	Incongruent condition	Normal condition	Congruent condition	Incongruent condition
<i>Field-independent</i>						
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
<i>Field-dependent</i>						
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

2 Repeated Measures Analysis of Variance

is that the **power** to detect the effects of within-subject experimental factors is increased compared to testing in a between-subject design. But this advantage of a repeated measures design comes at the cost of an increase in the complexity of the analysis and the need to make an extra assumption about the data than when only a single measure of the response variable is made on each subject (see later). This possible downside of the repeated measures approach arises because the repeated observations made on a subject are very unlikely to be independent of one another. In the data in Table 1, for example, an individual who reads faster than average under say one cue condition, is likely to read faster than average under the other two cue conditions. Consequently, the repeated measurements are likely to be correlated, possibly substantially correlated. Note that the data in Table 1 also contains a *between-subject factor*, cognitive style with two levels, field-independent and field-dependent.

Analysis of Variance for Repeated Measure Designs

The variation in the observations in Table 1 can be partitioned into a part due to between-subject variation, and a part due to within-subject variation. The former can then be split further to give a between-cognitive style mean square and an error term that can be used to calculate a mean square ratio and assessed against the appropriate F distribution to test the hypothesis that the mean reading time in the population of field-dependent and field-independent subjects is the same (see later for the assumptions under which the test is valid).

The within-subject variation can also be separated into parts corresponding to variation between the levels of the within-subject factors, their **interaction**, and their interaction with the between-subject factor along with a number of error terms. In detail, the partition leads to sums of squares and so on for each of the following:

- Cue condition, Cognitive style $\times$ Cue condition, error;
- Word type, Cognitive style $\times$ Word type, error;
- Cue condition $\times$ Word type, Cognitive style $\times$ Cue condition $\times$ Word type, error.

Again, mean square ratios can be formed using the appropriate error term and then tested against the relevant F distribution to provide tests of the following hypotheses:

- No difference in mean reading times for different cue conditions, and no cognitive style $\times$ cue condition interaction.
- No difference in mean reading time for the two types of word, and no interaction of word type with cognitive style.
- No interaction between cue condition and word type, and no second order interaction of cognitive style, cue condition, and word type.

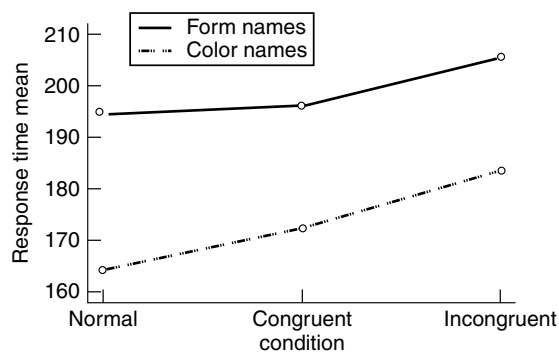
Full details of both the model behind the partition of within-subject variation and the formulae for the relevant sums of squares and so on are given in [2, 4]. The resulting **analysis of variance** table for the cognitive style data is given in Table 2.

Before interpreting the results in Table 2, we need to consider under what assumptions is such an analysis of variance approach to repeated measure designs valid? First, for the test of the between-subject factor, we need only normality of the response and homogeneity of variance, both familiar from the analysis of variance of data not involving repeated measures. But, for the tests involving the within-subject factors, there is an extra assumption needed to make the F tests valid, and it is this extra assumption that is particularly critical in the analysis of variance ANOVA of repeated measures data. The third assumption is known as **sphericity**, and requires that the variances of the differences between all pairs of repeated measurements are equal to each other and the same in all levels of the between-subject factors. The sphericity condition also implies a constraint on the covariance matrix (see **Correlation and Covariance Matrices**) of the repeated measurements, namely, that this has a *compound symmetry structure*, i.e., equal values on the main diagonal (the variances of the response under the different within-subject experimental conditions), and equal off-diagonal elements (the covariances of each pair of repeated measures). And this covariance matrix has to be equal in all levels of the between-group factors.

If, for the moment, we simply assume that the cognitive style data meet the three assumptions of normality, homogeneity, and sphericity, then the

Table 2 Repeated measures analysis of variance of reverse Stroop data

Source	Sum of squares	DF	Mean square	Mean square ratio	<i>P</i> value
Cognitive style	17578.34	1	17578.34	2.38	0.137
Error	162581.49	22	7390.07	2.38	
Word type	22876.56	1	22876.56	11.18	0.003
Word type $\times$ cognitive style	4301.17	1	4301.17	2.10	0.161
Error	45019.10	22	2046.32		
Condition	5515.39	2	2757.69	21.81	<0.001
Condition $\times$ Cognitive style	324.06	2	162.03	1.28	0.288
Error	5562.56	44	126.42		
Word type $\times$ Condition	450.17	2	225.08	3.14	0.053
Word type $\times$ Condition $\times$ Cognitive style	111.05	2	55.53	0.78	0.467
Error	3153.44	44	71.67		

**Figure 1** Interaction plot for cue condition $\times$ word type means

results in Table 2 lead to the conclusion that mean reading time differs in the three cue conditions and for the two types of word. There is also a suggestion of a cue condition $\times$ word type interaction. The mean reaction times for the three cue conditions are; normal –179.6 msec, congruent –184.3 msec, incongruent –194.5 msec. Under the incongruent condition, reaction times are considerably longer than under the other two conditions. The mean reaction times for the two word types are: form names –198.8 msec, color names –173.5 msec. Reaction times are quicker for the color names. To examine the cue condition $\times$ word type interaction, a graph of the six relevant mean reaction times is useful. This is shown in Figure 1 (*see Interaction Plot*). There is some slight evidence that the difference in reaction

time means for the two word conditions is greater in the normal condition than in the other two conditions.

What happens if the assumptions are not valid? Concentrating on the one specific to repeated measures data, namely, sphericity, if this is not valid then the *F* tests in the repeated measures analysis of variance are positively biased, leading to an increase in the probability of rejecting the null hypothesis when it is true, that is, increasing the size of the Type I error over the nominal value set by the researcher. This will lead to an investigator claiming a greater number of ‘significant’ results than are actually justified by the data.

How likely are repeated measures data in behavioral science experiments to depart from the sphericity assumption? This is a difficult question to answer, but if the within-subjects conditions are given in a random order to each subject in the study, then the assumption is probably easier to justify than when they are presented in the same order, particularly if there is a substantial time difference between the first condition and the last. The problem is that, in the latter case, observations closer together in time are likely to be more highly correlated than those taken further apart, thus violating the required compound symmetry structure for the repeated measures covariance matrix.

Where departures from sphericity *are* suspected or, perhaps, indicated by the formal test for the condition (*see Sphericity Test*), there are two approaches that might be used:

1. Correction factors,
2. Multivariate analysis of variance (MANOVA).

Correction Factors in the Analysis of Variance of Repeated Measures Data

The effects of departures from the sphericity assumption in a repeated measures analysis of variance have been considered in [3, 5], and it has been shown that the extent to which a set of repeated measures data deviates from the sphericity assumption can be summarized in terms of a quantity that is a function of the covariances and variances of the measures (see [2] for an explicit definition). Furthermore, an estimate of this quantity based upon sample variances and covariances can be used to decrease the degrees of freedom of the F tests for within-subject effects, to account for the departure from sphericity. (In fact, two different estimates of the correction factor have been suggested, one by Greenhouse and Geisser [3], and one by Huynh and Feldt [5].) The result is that larger mean square ratio values will be needed to claim statistical significance than when the correction is not used; consequently, the increased risk of falsely rejecting the null hypothesis is confronted. For the cognitive style data, the adjusted P values obtained using each estimate of the correction factor are shown in Table 3. Note that since word type has only two levels, the ‘corrected’ values are identical to those in Table 2, and for these data, the P values that do change do not differ greatly from the uncorrected values in Table 2. (This is, perhaps, not surprising here since the test for sphericity indicates that the assumption is valid.) More detailed examples of the use of this correction factor approach are given in [2, 4].

Table 3 Adjusted P values for reverse Stroop data

Factor	Greenhouse/ Geisser	Huynh/ Feldt
Word type	0.003	0.003
Word type × Cognitive style	0.161	0.161
Condition	<0.001	<0.001
Condition × Cognitive style	0.286	0.287
Word type × Condition	0.063	0.057
Word type × Condition × Cognitive style	0.447	0.460

Multivariate Analysis of Variance for Repeated Measure Designs

An alternative to the use of the correction factors in the analysis of repeated measures data when the sphericity assumption is judged to be inappropriate is to use a multivariate analysis of variance. Details are given in [2], but the essential feature of this approach is to transform the repeated measures so that they characterize different aspects of the within-subject factors, that is, main effects and interactions, and then use multivariate test criteria on the resulting sets of variables in the usual way. For example, in the cognitive style data, differences between the three cue conditions could be represented by differences (averaged over word types) between normal and congruent, and congruent and incongruent. The cue condition × word type interaction effect would involve differences between color and form names for these same differences in cue conditions. In the multivariate situation, there is no single test statistic that is always the most powerful for detecting all types of departures from the null hypothesis of the equality of mean vectors, and a number of different test statistics are in use. For details of these test statistics and the differences between them, see the multivariate analysis of variance entry. (Here, since there are only two groups involved, all the test criteria are equivalent.)

The results from the multivariate procedure are given in Table 4. (Note that since word type has only two levels, the multivariate result is equivalent to the univariate result given in Table 2.) The results again indicate highly significant condition and word type main effects, but, here, the condition × word type interaction is, unlike in the univariate analysis, also highly significant.

The main advantage of using MANOVA for the analysis of repeated measures designs is that no assumptions now have to be made about the pattern of covariances between the repeated measures. In particular, these covariances need not satisfy the compound symmetry condition. A disadvantage of using MANOVA for repeated measures is often stated to be the technique’s relatively low power when the assumption of compound symmetry is actually valid. However, the power of the univariate and multivariate analysis of variance approaches when compound symmetry holds is compared in [1] with the conclusion that the latter is nearly as powerful as

Table 4 Multivariate analysis of variance of reverse Stroop data

Effect	Value	<i>F</i>	Hypothesis <i>df</i>	Error <i>df</i>	<i>P</i> value
Word type:					
Pillai's trace	0.34	11.18	1	22	0.003
Wilk's lambda	0.66	11.18	1	22	0.003
Hotelling's trace	0.51	11.18	1	22	0.003
Roy's largest root	0.51	11.18	1	22	0.003
Word type $\times$ Cognitive style:					
Pillai's trace	0.09	2.10	1	22	0.161
Wilk's lambda	0.91	2.10	1	22	0.161
Hotelling's trace	0.10	2.10	1	22	0.161
Roy's largest root	0.10	2.10	1	22	0.161
Condition:					
Pillai's trace	0.66	20.73	2	21	<0.001
Wilk's lambda	0.34	20.73	2	21	<0.001
Hotelling's trace	1.97	20.73	2	21	<0.001
Roy's largest root	1.97	20.73	2	21	<0.001
Condition $\times$ Cognitive style:					
Pillai's trace	0.13	1.63	2	21	0.220
Wilk's lambda	0.87	1.63	2	21	0.220
Hotelling's trace	0.16	1.63	2	21	0.220
Roy's largest root	0.16	1.63	2	21	0.220
Word type $\times$ Condition:					
Pillai's trace	0.37	5.32	2	21	0.014
Wilk's lambda	0.66	5.32	2	21	0.014
Hotelling's trace	0.51	5.32	2	21	0.014
Roy's largest root	0.51	5.32	2	21	0.014
Condition $\times$ Word type $\times$ Cognitive style:					
Pillai's trace	0.48	0.53	2	21	0.595
Wilk's lambda	0.95	0.53	2	21	0.595
Hotelling's trace	0.51	0.53	2	21	0.595
Roy's largest root	0.51	0.53	2	21	0.595

the former when the number of observations exceeds the number of repeated measures by more than 20.

Summary

An analysis of variance can often be safely applied to repeated measures data arising from psychological experiments when the within-subject conditions are presented in a random order to each subject, since with such designs, the sphericity assumption is likely to be justified. There is, however, one type of repeated measures data where sphericity is very unlikely to hold, namely, **longitudinal data**. For such data, the only within-subject factor is "time", and so randomization is no longer an option. Consequently, it is very likely that observations taken closer together in

the study will be more similar than those separated by a longer time interval; consequently, assuming compound symmetry will not, in general, be sensible. For this reason, longitudinal data requires more sophisticated approaches, for example, **linear multi-level models**. Such models can also deal with the frequently occurring practical problem of **missing data** in longitudinal data sets, in particular when such observations are missing because subjects drop out of the study (*see Dropouts in Longitudinal Studies: Methods of Analysis*).

References

- [1] Davidson, M.L. (1972). Univariate versus multivariate tests in repeated measures experiments, *Psychological Bulletin* 77, 446–452.

6 Repeated Measures Analysis of Variance

- [2] Everitt, B.S. (2000). *Statistics for Psychologists*, Lawrence Erlbaum, Mahwah.
- [3] Greenhouse, S.W. & Geisser, S. (1959). On the methods in the analysis of profile data, *Psychometrika* **24**, 95–112.
- [4] Howell, D.C. (2002). *Statistical Methods for Psychology*, Duxbury, Pacific Grove.
- [5] Huynh, H. & Feldt, L.S. (1976). Estimated of the correction for degrees of freedom for sample data in randomized block and split-plot designs, *Journal of Educational Statistics* **1**, 69–82.
- [6] Pahwa, A. & Broota, K.D. (1981). Field-independence, field dependence as a determinant of colour-word interference, *Journal of Psychological Research* **30**, 55–61.

Further Reading

Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*, Chapman & Hall, London.

(See also **Generalized Estimating Equations (GEE); Generalized Linear Mixed Models; Marginal Models for Clustered Data**)

BRIAN S. EVERITT

Replicability of Results

BRUCE THOMPSON

Volume 4, pp. 1741–1743

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Replicability of Results

Quantitative researchers seek to identify relationships that recur under stated conditions. Scholars in the physical sciences have the luxury of the idiosyncrasies of human personality not confounding their results. A physicist observing a quark and a neutrino running away from each other may make an inference that these two atomic particles have like charges. The physicist need not qualify generalizations with statements about the nutrition of the quark during gestation, or the quality of schooling received by different quarks during their early years.

The business of behavioral science is much more difficult. Behavioral scientists attempt to formulate generalizations about people, which hold up reasonably well, but recognize that few statements apply equally well to all people. Behavioral scientists attempt to overcome the vagaries of individual differences in various ways, including conducting studies with large numbers of people, so that the idiosyncrasies may 'wash out' within the large sample.

Some behavioral scientists erroneously believe that statistical significance testing evaluates the replicability of results. In fact, statistical significance does not evaluate result replicability, as **Cohen** [1] and others [10] have shown. Because statistical tests do not evaluate result replicability, and replicability is very important, other methods must be used to test result replicability.

Thompson [9] suggested that there are two kinds of replicability evidence: external and internal. External replicability evidence requires the researchers to conduct the research study a second time, to determine whether the results are stable.

Another form of external replicability analysis involves 'meta-analytic thinking' (see **Meta-Analysis**) in which the researcher focuses on explicitly and directly comparing study effect sizes with the effect sizes in the related prior studies [2, 11, 13]. If all the effects across studies are similar, the researcher has more confidence that the results in a given study are not purely serendipitous.

The most persuasive evidence of result replicability is actual study replication. The problem is that most researchers do not have the time or the resources to replicate every study prior to

publishing their results or submitting their theses. In such cases, internal replicability analyses [9] are a partial substitute for true external replication.

The basic idea of internal replicability analyses is to mix up the participants in different ways, to determine whether the results remain stable across different combinations of people. The intent is to approximate modeling the variations in personalities that would occur if an actual new sample had been drawn. These internal replicability analyses are never as good as true replication, but are generally more informative as regards replicability compared to what many researchers do to establish replicability of their results - nothing!

There are three primary principles of logic for conducting internal replicability analyses: **cross-validation**, the **jackknife**, and the **bootstrap**. (Carl Huberty has suggested combining the latter two methods to create another alternative - the jackstrap. The more serious point is that the researcher can do whatever seems reasonable to evaluate result replicability, even if the approach has not been previously employed.)

Traditionally, cross-validation involves randomly splitting the sample into two nonoverlapping subsamples, replicating the analysis in both subsamples, and then conducting additional empirical analyses to determine whether the results are similar in the two subsamples. Thompson [6] provides details in a heuristic example for the multiple regression case. Factor analytic examples are provided by Thompson [12].

The jackknife was popularized by **John Tukey**. In the jackknife, the primary resampling mechanism is to redo the analysis by successively dropping out subsets of participants of a given size, k (e.g., $k = 1$, $k = 2$). For example, the researcher might drop out subsets of participants where each dropped subset is size 1. In a regression involving 300 participants, the analysis would be done with all 300 participants. And the regression would then be done 300 more times, each with a sample size equal to $n - 1$, where a different participant is dropped each time.

The bootstrap was popularized by Bradley Efron; [3] is a readable explanation. Lunneborg [4] provides a technical but comprehensive explanation. Here resamples are drawn each of a size n , but typically are drawn with replacement. For example, in a regression analysis involving 250 participants, the first resample

2 Replicability of Results

might be drawn such that Wendy's data on all the variables is drawn as a set five times, while Deborah's data is not drawn at all. However, in the second resample, Wendy's data might not be drawn at all, while Deborah's data might be drawn three times.

Usually when the bootstrap is invoked, thousands of resamples are drawn and analyzed. Indeed, both the jackknife and the bootstrap are called 'computationally intensive', because they usually cannot be done unless specialized software is used to execute the numerous analyses required. This software, though not part of SPSS, is widely available, especially as regards the bootstrap.

The cross-validation method is the least sophisticated of the three methods because it involves only one sample split. The problem is that for a given data set numerous splits are possible, and different splits might yield contradictory results as regards the same data.

Because both the jackknife and the bootstrap are computationally intensive, but the bootstrap has the appeal of mixing up the participants in very many ways to see if the results remain stable, researchers considering these two choices quite frequently opt for the bootstrap. Although the bootstrap sounds like a tremendous amount of work, the work is done by a modern microcomputer in seconds, and is thus quite painless.

A special challenge arises when using the bootstrap with multivariate statistics (see **Multivariate Analysis: Overview**). Multivariate statistics will usually yield several functions or factors in a given analysis. The orders of the factors may arbitrarily vary across resamples. This is only a logistical issue, because usually the order of a given construct is not that meaningful. But some way must be found to compare a given construct across the thousands of resamples such that apples are compared to apples [12]. Thompson [5] proposed invoking a special statistical rotation method, called Procrustean rotation, to solve this problem (see **Procrustes Analysis**). This solution has been generalized to bootstrap factor analysis [5], descriptive **discriminant analysis** [8], and **canonical correlation analysis** [7].

References

- [1] Cohen, J. (1994). The earth is round ($p < 0.05$), *American Psychologist* **49**, 997–1003.
- [2] Cumming, G. & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions, *Educational and Psychological Measurement* **61**, 532–575.
- [3] Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics, *Scientific American* **248**(5), 116–130.
- [4] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury, Pacific Grove.
- [5] Thompson, B. (1988). Program FACSTRAP: a program that computes bootstrap estimates of factor structure, *Educational and Psychological Measurement* **48**, 681–686.
- [6] Thompson, B. (1989). Statistical significance, result importance, and result generalizability: three noteworthy but somewhat different issues, *Measurement and Evaluation in Counseling and Development* **22**(1), 2–5.
- [7] Thompson, B. (1991). Invariance of multivariate results, *Journal of Experimental Education* **59**, 367–382.
- [8] Thompson, B. (1992). DISCSTRA: a computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids, *Educational and Psychological Measurement* **52**, 905–911.
- [9] Thompson, B. (1994). The pivotal role of replication in psychological research: empirically evaluating the replicability of sample results, *Journal of Personality* **62**, 157–176.
- [10] Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms, *Educational Researcher* **25**(2), 26–30.
- [11] Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes, *Educational Researcher* **31**(3), 24–31.
- [12] Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, American Psychological Association, Washington.
- [13] Thompson, B. (in press). Research synthesis: effect sizes, in *Complementary Methods for Research in Education*, J. Green, G. Camilli & P.B. Elmore, eds, American Educational Research Association, Washington.

BRUCE THOMPSON

Reproduced Matrix

CHARLES E. LANCE

Volume 4, pp. 1743–1744

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Reproduced Matrix

A reproduced matrix (symbolized as $\hat{\Sigma}$) is a matrix of correlations or covariances that is calculated from parameter estimates obtained for a path analytic (see **Path Analysis and Path Diagrams**), factor analytic (see **Factor Analysis: Confirmatory**), or latent variable **structural equation model**. Take, for example, the hypothetical path model shown in Figure 1. If we were to gather data on the X and Y variables shown in this figure, we could estimate this model's parameters in the following three equations:

$$Y_1 = \alpha_1 + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_3 + d_1 \quad (1a)$$

$$Y_2 = \alpha_2 + \beta_{21}Y_1 + \gamma_{23}X_3 + \gamma_{24}X_4 + d_2 \quad (1b)$$

$$Y_3 = \alpha_3 + \beta_{31}Y_1 + \beta_{32}Y_2 + d_3 \quad (1c)$$

and calculate $\hat{\Sigma}$ in part from the estimated model parameters. $\hat{\Sigma}$ actually consists of three conceptually distinct parts, however, correlations (or covariances) among the endogenous variables ($\hat{\Sigma}_{YY'}$), correlations among the exogenous variables ($\hat{\Sigma}_{XX'}$), and correlations between the exogenous and endogenous variables ($\hat{\Sigma}_{XY} = \hat{\Sigma}_{YX}'$) so that overall

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{YY'} & \hat{\Sigma}_{YX} \\ \hat{\Sigma}_{XY} & \hat{\Sigma}_{XX'} \end{bmatrix}. \quad (2)$$

As is shown in Figure 1, the exogenous variables are assumed to be correlated (as is usually the case) so that $\hat{\Sigma}_{XX'}$ is given from the data. However, correlations between the X s and Y s in $\hat{\Sigma}_{XY}$ result from various functions of model parameters, including direct effects of X s on Y s (e.g., the effect of X_1 on $Y_1 - \gamma_{11}$), indirect effects of X s on Y s (e.g., the effect of X_1 on Y_2 through $Y_1 - \gamma_{11}\beta_{21}$), and/or common causal effects (e.g., X_3 affects both Y_1 and $Y_2 - \gamma_{13}\gamma_{23}$). Similarly, correlations among the Y s in

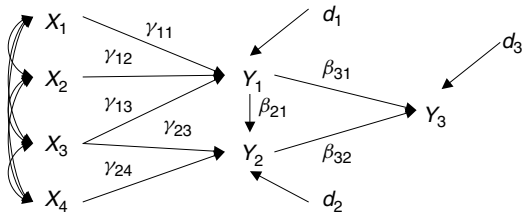


Figure 1 Hypothetical path model

$\hat{\Sigma}_{YY'}$ also arise from direct effects (e.g., the effect of Y_1 on $Y_2 - \beta_{21}$), indirect effects (the effect of Y_1 on Y_3 through $Y_2 - \beta_{21}\beta_{32}$), and common causal effects (Y_1 affects both Y_2 and $Y_3 - \beta_{21}\beta_{31}$). So, once the model's parameters in (1a) through (1c) are estimated from data, they can be used to calculate or *reproduce* the correlations among the variables in the model by using the parameter estimates to solve the equations reflecting the decomposition of effects specified by the model.

An important question is whether the *reproduced* correlations equal (or approximate) the *observed* correlations calculated directly from data. The reproduced matrix will equal (or approximate) the observed correlations if (a) the model being estimated is correct, or (b) as many model parameters are estimated as there are elements in the observed correlation matrix. However, neither of these will generally be the case. First, it is widely accepted that models that are tested in the behavioral sciences are rarely 'true' (see [2]). Second, most models in the behavioral sciences estimate fewer (and often many fewer) parameters than elements in the matrix of correlations among variables in the model, and this is true of the hypothetical model in Figure 1. This can be seen by rewriting (1a) through (1c) as follows:

$$Y_1 = \alpha_1 + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_3 + \mathbf{0}X_4 + d_1 \quad (3a)$$

$$Y_2 = \alpha_2 + \beta_{21}Y_1 + \mathbf{0}X_1 + \mathbf{0}X_2 + \gamma_{23}X_3 + \gamma_{24}X_4 + d_2 \quad (3b)$$

$$Y_3 = \alpha_3 + \beta_{31}Y_1 + \beta_{32}Y_2 + \mathbf{0}X_1 + \mathbf{0}X_2 + \mathbf{0}X_3 + \mathbf{0}X_4 + d_3 \quad (3c)$$

where the bold elements represent 'zero-effect' hypotheses, that is, hypotheses that certain effects that *could be* estimated within a particular model are actually zero. If a model contains no zero-effect hypotheses, then as many parameters are estimated as there are elements in the observed correlation matrix and the reproduced correlation matrix will equal the observed correlation matrix by tautology.¹ For models in which there are one or more 'zero restrictions' as in (3a) through (3c), the reproduced correlation matrix will not necessarily equal the observed correlation matrix and, in fact, it most likely will not. This is because one or more of the zero restrictions may be incorrect (i.e., the true effect is actually nonzero) or

2 Reproduced Matrix

the restriction may hold only approximately. There are a number of ways that the plausibility of zero-effect hypotheses can be tested (see [1]), but the most popular, omnibus statistical test is the χ^2 statistic based on the maximum likelihood fit function ($\chi^2 = (n - 1)F_{ML}$) where n = sample size and

$$F_{ML} = \log |\hat{\mathbf{S}}| + \text{tr}(\mathbf{S} \hat{\mathbf{S}}^{-1}) - \log |\mathbf{S}| - (p + q), \quad (4)$$

where $\log$ refers to the natural logarithm, $\mathbf{S}$ refers to the observed or sample correlation (or covariance) matrix, $|\mathbf{W}|$ and $\text{tr}(\mathbf{W})$ denote the determinant and trace of matrix $\mathbf{W}$, respectively, and p and q refer to the number of Y and X variables, respectively. If the model's zero restrictions are (at least approximately) tenable, then the discrepancy $\log |\hat{\mathbf{S}}| - \log |\mathbf{S}|$ from (4) will be small as will the discrepancy $\text{tr}(\mathbf{S} \hat{\mathbf{S}}^{-1}) - (p + q)$ so that the resulting χ^2 will also tend to be 'small.' If one or more zero restrictions do *not* hold, the discrepancy between $\hat{\mathbf{S}}$ and $\mathbf{S}$ increases, along with the F_{ML} discrepancy function and the χ^2 statistic. As such, the χ^2 statistic is a badness-of-fit index that is inexorably tied to the discrepancy between the observed and reproduced correlation (or covariance) matrices and whose degrees of freedom is closely related to the number

of zero restrictions that are imposed on the model tested. Large and statistically significant χ^2 statistics indicate that one or more zero restrictions imposed in a path, factor analytic, or latent variable structural equation model are implausible.

References

- [1] James, L.R., Mulaik, S.A. & Brett, J.M. (1982). *Causal Analysis: Models, Assumptions, and Data*, Sage Publications, Beverly Hills.
- [2] MacCallum, R.C. (2003). Working with imperfect models, *Multivariate Behavioral Research* **38**, 113–139.

Note

1. The issue of model identification is actually much more complex an issue than can be treated here. Our humble goal is merely to introduce some notions related to model identification as they help understand the important role of the reproduced matrix in assessing model fit.

(See also **Factor Analysis: Confirmatory**)

CHARLES E. LANCE

Resentful Demoralization

PATRICK ONGHENA

Volume 4, pp. 1744–1746

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Resentful Demoralization

Resentful demoralization is a validity threat in intervention studies (*see* **Clinical Trials and Intervention Studies**) in which comparison groups not obtaining a desirable treatment are aware of this inequity or find out about this during the study, become discouraged or retaliatory, and as a result perform worse on the outcome measures [4, 5]. The prototypical setting for this threat to operate is a randomized two-group intervention study with a treatment group and a no-treatment control group in which treatment allocation is obtrusive and participants in the control group get deprived of certain facilities offered to the participants in the treatment group. As a consequence, people in the control group could feel neglected and behave differently than they would have behaved if they had not known about the favorable intervention in the treatment group. The most likely effect of resentful demoralization is that the observed treatment effect gets inflated and that the intervention program looks more effective than it actually is.

As a social interaction threat to validity, this confound has the same basis as **compensatory rivalry** (i.e., perceived inequality between the comparison groups), but the emotional and behavioral reaction that is involved, is quite the opposite. Whereas in case of compensatory rivalry the participants in the deprived group(s) become competitive, in case of resentful demoralization they get dejected or vindictive and bring about inferior performance. For both threats, the treatment effects become confounded by the differential motivation to perform, but in the case of compensatory rivalry, the treatment effect probably gets underestimated, while in the case of resentful demoralization it gets overestimated because of the confound.

Campbell and Stanley [3] already hinted at this problem when they warned us of the possibility of ‘reactive arrangements’ in experimental research, but the term ‘resentful demoralization’ made its first appearance in Cook and Campbell’s list of **internal validity** threats [4]. According to Shadish, Cook, and Campbell [5], resentful demoralization is even so closely associated with the treatment construct itself that it should be included as part of the treatment construct description. Therefore, in their revised list Shadish et al. [5] classified this confound among the threats to construct validity. According to these

authors, internal validity threats are disturbing factors that can occur even without a treatment, while this is obviously not the case with resentful demoralization: ‘The problem of whether these threats should count under internal or construct validity hinges on exactly what kinds of confounds they are. Internal validity confounds are forces that could have occurred in the absence of the treatment and could have caused some or all of the outcome observed’ (p. 95).

Resentful demoralization is clearly related to compensatory rivalry, but is also related to other construct and internal validity threats as well. If participants have the impression that they are treated unfairly, then they could try to get the beneficial treatment anyhow, resulting in **compensatory equalization**. Or if the demoralization is vast, then the participants of the control group might decide to stop participation altogether, resulting in differential **attrition**. Furthermore, notice that, although the prototypical setting described above is a randomized trial in which treatment allocation is conspicuous, also a **quasi-experiment** working with eligible participants for one of the treatment arms is particularly susceptible to this kind of bias.

Resentful demoralization can be avoided or minimized by isolating the comparison groups in time (e.g., using a waiting list condition) or space (e.g., using geographically remote groups), or making the participants unaware of the intervention being applied (e.g., using blinding). If no such design control is possible, poststudy assessment strategies could be useful, for instance, by asking the participants in a debriefing whether they felt uncomfortable being assigned to the control condition, and by relating these responses to the outcome.

A good example of the latter strategy is given in the **prospective study** by Berglund et al. [1, 2] in which 98 of the 199 cancer patients who wanted to participate in the study were randomly assigned to the structured rehabilitation program ‘Starting Again’, and the other patients were assigned to the control condition. This program consisted of 11 two-hour sessions focusing on physical training, information, and training of coping skills for cancer patients, and although the results were generally positive, the researchers were suspicious about the possibility of resentful demoralization because the patients were aware of the treatment assignment procedure by the informed consent. However, their poststudy analysis showed that although a few patients may have been

2 Resentful Demoralization

resentful, this did not significantly affect the outcome variables. Furthermore, they were able to compare the 101 patients assigned to the control condition with 73 patients who did not wish to participate in the study in the first place, and found no negative effects resulting from being randomized to the control condition.

As a conclusion, if assignment-related motivational effects, such as resentful demoralization, in some of the comparison groups are suspected, then researchers should be cautious about any causal generalization of the treatment. However, in intervention studies that have to deal with these potentially confounding motivational effects, the credibility of the results may be strengthened by design control, post-study assessment, and qualified interpretations.

References

- [1] Berglund, G., Bolund, C., Gustafsson, U.-L. & Sjöden, P.-O. (1994). One-year follow-up of the 'Starting Again' group rehabilitation program for cancer patients, *European Journal of Cancer* **30A**, 1744–1751.
- [2] Berglund, G., Bolund, C., Gustafsson, U.-L. & Sjöden, P.-O. (1997). Is the wish to participate in a cancer rehabilitation program an indicator of the need? Comparisons of participants and non-participants in a randomized study. *Psycho-Oncology* **6**, 35–46.
- [3] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.
- [4] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand-McNally, Chicago.
- [5] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

PATRICK ONGHENA

Residual Plot

JEREMY MILES

Volume 4, pp. 1746–1750

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Residual Plot

The **residuals** in a regression analysis (*see* **Multiple Linear Regression**) (or related, e.g., **analysis of variance**) are the difference between the scores on the outcome variable predicted from the regression equation and the observed scores. The distributional (and other) assumptions that are made when using regression analysis relate to the residuals, rather than the outcome variable. Residuals can be transformed in different ways to improve their interpretability (*see* **Residuals**). Here, when we refer to residuals, we generally mean studentized-deleted residuals (also known as externally studentized residuals), unless explicitly stated otherwise.

Draper and Smith [2] suggest that residuals should be plotted

- overall, in a **histogram** or **boxplot**;
- against the time in which the data were collected, in a **scatterplot**;
- against the predicted values of the outcome variable;
- against the independent variable(s);
- in any other way that seems to be sensible.

Distribution of Residuals

The distribution of the residuals should always be examined to check for normality and for **outliers**. If outliers are detected, these may be further examined, and tested for statistical significance. If the data are nonnormal, this may suggest that a transformation of the data would be appropriate (although it is possible for a normally distributed outcome variable to give rise to nonnormal residuals).

The following data were taken from a study examining 51 women who were undergoing a biopsy for breast cancer. Anxiety, depression, and self-esteem were measured one week before undergoing the biopsy for breast cancer, and on the day that they underwent the biopsy. The analysis presented looks at anxiety on the day of the biopsy as an outcome, and predictors are anxiety, depression, and self-esteem, measured one week before the biopsy. The predictors had a large and highly significant effect on the outcome variable ($R^2 = 0.82$, $p < 0.001$). The standardized coefficients were -0.17 ($p = 0.039$) for

self-esteem, 0.192 ($p = 0.030$) for depression, and 0.644 ($p < 0.001$) for anxiety.

The residuals can be examined using a histogram, **probability plots** or boxplots – each gives the same information, but presents it in a slightly different format. Although any of the types of residual can be used, the most useful is probably the studentized-deleted (also known as the externally studentized, or the jack-knifed) residual, because this is the most easily interpreted.

The histogram (Figure 1) shows that distribution is approximately symmetrical, but the tails of the distribution may be heavier than we would expect from a normal distribution – we might consider them to be outliers.

The probability plot (Figure 2) similarly shows that the distribution is symmetrical and deviates away from the normal distribution – probability plots are poor at detecting outlying cases, and so it is not clear that there are potential outliers here. Finally, Figure 3 shows the box plots for three types of residuals, the standardized, studentized (also known as the internally studentized) and the studentized-deleted. The distribution of the studentized-deleted residuals seems to have wider tails than the other distributions. The boxplot is also useful for identifying outliers – there are clearly three points that might be considered to be outliers. The largest one of these has a studentized-deleted residual of 3.77. The Bonferroni-corrected probability of finding a studentized residual with an absolute value of 3.77 is equal to 0.07 and close to the conventional cutoff of 0.05. However, we should not discard data without good reason, and taking a conservative approach would be better in this

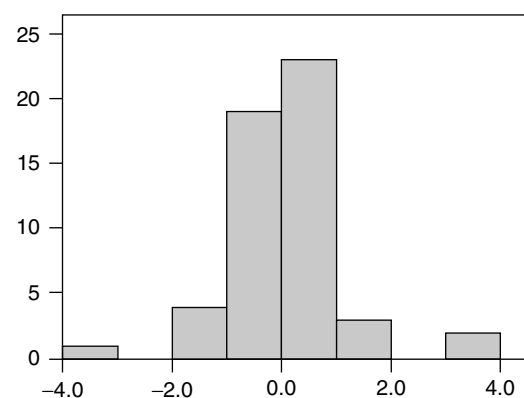


Figure 1 Histogram of studentized-deleted residuals

2 Residual Plot

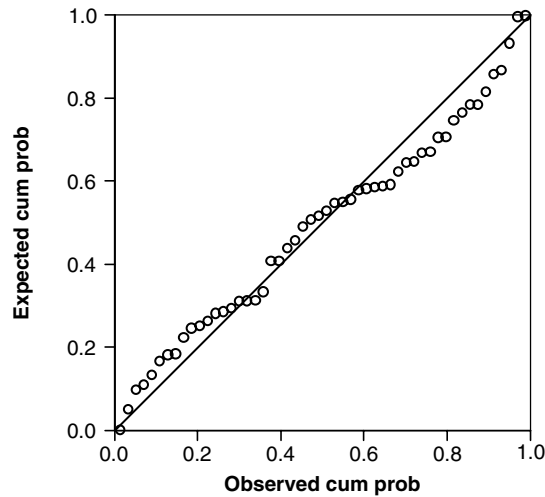


Figure 2 Normal probability plot of studentized-deleted residuals

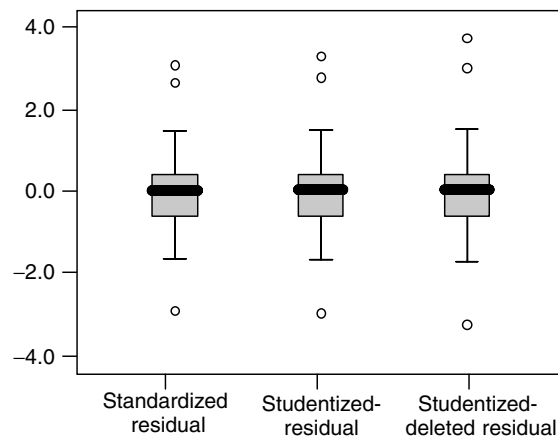


Figure 3 Boxplots of standardized residual, studentized residual, and studentized-deleted residual. Note that the studentized-deleted residual has the larger variance, and the residuals appear to be more extreme

case, so we shall leave the data point in the file. (It is still probably worth treating the case with some suspicion, and ensuring that there is nothing else that leads us to have misgivings about it.)

Residuals Against Time

Plotting residuals against time or participant number in a scatterplot, where participant number reveals the

order in which the data were collected, is the next task (*see Index Plots*). This serves two purposes: first, it reassures us that there is no linear relation with time, and second, it can be used to detect violations of the independence assumption.

If the data were collected over a period (as they almost certainly were), then it is possible that something has changed over that period which would have had an effect on the data. This may need to be taken into account or may invalidate the research. There are a number of ways in which time might have an effect. If the data were collected over the course of a day, it is possible that the conditions changed over that day – the experimenter may have become more tired, a piece of equipment may have gradually lost its calibration, the room temperature may have gradually increased, for example. In physiological studies, it is possible that the substance under examination may have decayed over time – if the control group is analyzed first, this might lead to a spurious result if the effect of time is not taken into account.

Figure 4 shows the scatterplot of the studentized residual against the participant number. Note that here I have added two reference lines, at ± 1.96 , indicating that 95% of the points should lie between these lines. Figure 4 also contains a loess regression line (*see Scatterplot Smoothers*) which can help to identify any trend in the data.

Examining Figure 4 suggests two possible effects that may have occurred in these data. First, there seems to be a trend in that the residuals are increasing in value over time. In this case, where it appears that

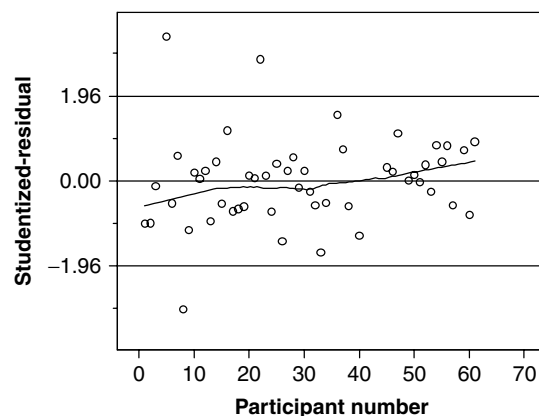


Figure 4 Scatterplot of studentized residuals plotted against participant number with loess regression line

there is a monotonic relationship, we can examine the correlation between the residuals and the participant number. This result is not statistically significant ($r = 0.14$, $p = 0.33$), and therefore does not support the hypothesis.

The second possibility shown by this graph is that the variance is nonconstant at different levels of the predictor variable. This effect is known as heteroscedasticity (see **Heteroscedasticity and Complex Variation**). One possible cause of heteroscedasticity of these residuals is the reliability changing over time – in the case that may occur here, it may be that the measurements of the outcome variable are increasing in reliability over time, and that there is more error at the start of the study than at the end. It could be that the increase in measurement reliability is caused by the experimenter becoming more familiar with the procedure or with a piece of equipment. We can test the heteroscedasticity assumption using White's test, and we find that $\chi^2 = 3.16$, $df = 2$, $p = 0.21$; again, there is no statistical evidence to support the hypothesis that the variance is non-constant.

The second use of this type of residual plot is that it can help us detect violations of the independence assumption in **Generalized Linear Model** analyses.

It is assumed in regression that residuals are independent – that is, knowing the value of one residual should not help us predict the value of the next residual. For example, if we tested our participants in groups or batches, we might find that the groups were in some way similar to each other – if our outcome was alcohol consumption, it is feasible that there might be a conformity effect, and in some groups everyone drank a lot, and in others everyone drank little (it is also possible that the opposite effect might have occurred – one person drinking heavily may have stopped the others from drinking as much). Similarly, this problem may occur if we select participants who are in some way in naturally occurring groups, and where these groupings might affect the outcome variable – common examples of this problem include children in classroom groups, patients treated by the same general practitioner (GP) or hospital and students living in the same hall of residence. Figure 5 shows a scatterplot of participant number against the residual for 35 participants. The chart shows that there are five groups of participants, the first with positive residuals, then a group with negative residuals, and

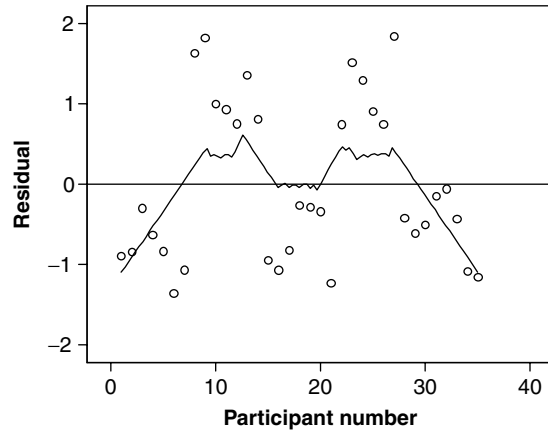


Figure 5 Plot of participant number against residual, where there are clustered data with loess regression line

so on. The loess line has been added, to emphasize this effect. This type of effect is (or should be) relatively rare – researchers are increasingly aware of the problems of clustered data, and either try to avoid the clustering or alternatively take it into account (see [3] and [4]).

The second way in which nonindependence may occur is if the previous measurement is in some way predictive of the next measurement (or any future measurement.) The most common example of this effect is in time-series designs, where instead of measuring multiple people on one occasion, one person (or a small number of people) is measured on multiple occasions. If we are measuring reaction time, it is possible that this will slow down towards the end of a long series of trials. If we are measuring mood over the course of a number of days, my mood yesterday is likely to be a better predictor of my mood today than it is of my mood next Tuesday.

If graphical examination of the data reveals that there is a clustering effect that was not anticipated, then either the Durbin–Watson test or a runs test should be carried out to examine whether the effect is statistically significant.

Residuals Plotted Against Predicted Values

Perhaps the most important plot is that of the residuals plotted against the fitted values, shown in

4 Residual Plot

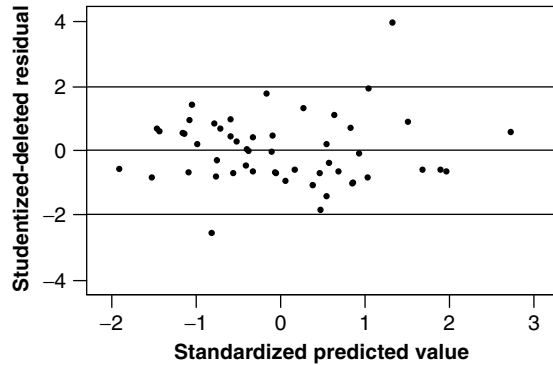


Figure 6 Standardized predicted values plotted against studentized-deleted residuals. Additional horizontal lines have been drawn at ± 1.96 on the y-axis

Figure 6. This can be used to reveal several different potential problems: outliers, nonlinearity, heteroscedasticity, and nonnormality.

Outliers are revealed in this graph as cases that are far from the other cases – the graph has additional gridlines extending from the y-axis and ± 1.96 . We would expect 95% of cases to lie within these lines. The point at the top of the graph, with a residual approaching four is particularly problematic.

The points in the graph should form a straight line across the chart – there should be no tendency of the residuals to be above or below zero at any point. Figure 7 shows the general shape that is indicative of a nonlinear relationship in the data, and hence that the violation of linearity has been violated.

The same plot allows us to examine the assumption of constant variance across the range of predicted values. Here we would be looking for a constant spread of the residuals around the center line. Figure 8 shows the shape of the plot that would be expected if the nonconstant variance assumption were to be violated.

Finally, something that is occasionally disturbing to data analysts is the presence of stripes, which make the residual plot look rather different from the examples generally presented in textbooks. Stripes occur because of the measurement properties of the variables – the measures are treated as continuous, but are actually discrete. In Figure 9, the outcome variable is a test with seven items, which are scored as correct or incorrect. Each individual therefore scores

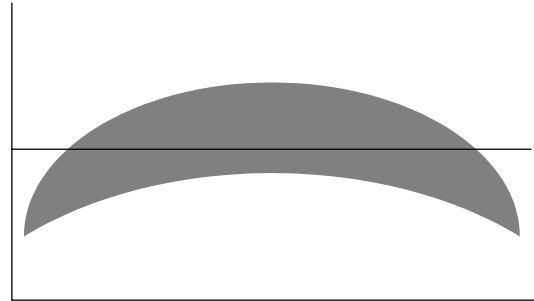


Figure 7 Scatterplot of residuals against predicted values showing nonlinearity

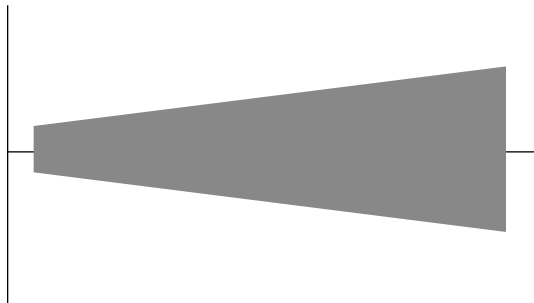


Figure 8 Scatterplot of residuals against predicted values, showing nonconstant variance

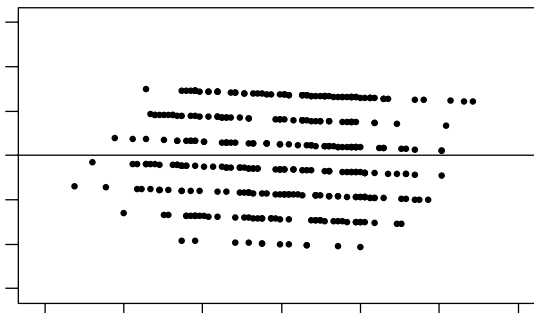


Figure 9 Residual plot showing the presence of stripes

a whole number, from a minimum of zero to a maximum of seven.

Cook and Weisberg [1] provide an excellent guide to graphics in regression; there is a computer program (ARC) to accompany the book.

References

- [1] Cook, R.D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, John Wiley & Sons, New York.
- [2] Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Edition, John Wiley & Sons, New York.
- [3] Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*, Lawrence Erlbaum Associates, Mahwah.
- [4] Leyland, A. & Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*, Wiley Series in Probability and Statistics, Wiley, Chichester.

JEREMY MILES

Residuals

JEREMY MILES

Volume 4, pp. 1750–1754

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Residuals

It is (too) commonly believed that distributional assumptions for many statistical tests are made on the variables – this is not the case: for most statistical techniques, the assumptions that are made are dependent on the distribution of the residuals.

Consider a group of 10 individuals, of different ages, who have been presented with a quiz that measured their knowledge of a range of areas. The age of each person and the number of items that they recalled are shown in the first two columns of Table 1. If we fit a least squares linear **regression model** to these data, we find that

$$\hat{x} = 0.223 \times a + 12.52, \quad (1)$$

where x is the score and a is the age of the individual. The hat on the x indicates that this is a predicted score, not the actual score. The 95% **confidence interval** for the regression coefficient is 0.063 to 0.383, and the associated significance is 0.012.

In other words, we would expect that people who are one year older will score 0.223 points higher on the quiz.

If we wished to use the actual score for each individual, we need to add a term to take account of the difference between the predicted score and the score that each individual achieved – we refer to this as *error* (e_i), and these are the residuals.

$$x_i = 0.223 \times a + 12.52 + e_i \quad (2)$$

Table 1 Data, predicted values, and residuals

Age (a)	Number of items correct (x)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)
20	15	16.98	−1.98
25	21	18.10	2.90
30	19	19.21	−0.21
35	18	20.33	−2.33
40	21	21.44	−0.44
45	25	22.56	2.44
50	22	23.67	−1.67
55	26	24.79	1.21
60	31	25.90	5.10
65	22	27.02	−5.02

(The i subscript has been added to index each individual.) Rearranging this equation gives:

$$e_i = x_i - (0.223 \times a + 12.52) \quad (3)$$

However, note that the part of the equation $0.223 \times a + 12.52$ is equal to $\hat{x}$, therefore we can substitute this into the equation, giving:

$$e_i = x_i - \hat{x}_i \quad (4)$$

These values are shown in Table 1.

In regression and related analyses (**analysis of variance**, t Test, **analysis of covariance**, etc.), we assume that the residuals are sampled from a normally distributed population with a mean equal to zero.

To illustrate the difference between the distribution of the variables and the distribution of the residuals, consider the simple example of an independent samples t Test. Figure 1 shows that the distribution of the outcome variable appears to be positively skewed. Figure 2 shows the distribution of two different groups – here it can be seen that the distribution in Figure 1 is actually comprised of two different distributions, one for each of two groups, and that the group with the higher mean has a smaller sample size. The predicted values in this case are simply the means for each of the two groups, and so the residuals are the difference between each value and the mean for that group. Figure 3 shows the distribution of the residuals – this is clearly from a normal distribution.

In the case of the two group t Test, it was easy to see how the distribution of the variable could make us believe that we may have had a problem with our variable – in a more complex analysis, for example, if we were to carry out an analysis of covariance, in an experiment with a 2×2 design and a covariate,

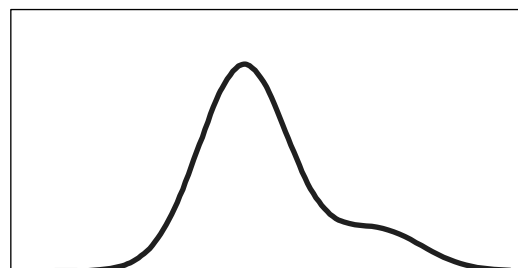


Figure 1 Histogram that shows distribution of outcome variable

2 Residuals

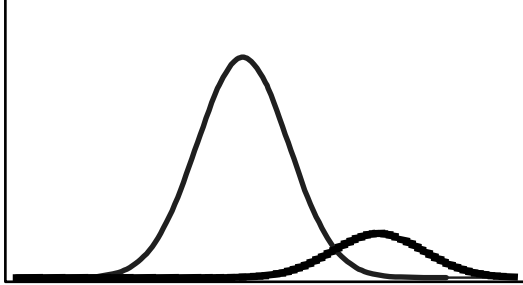


Figure 2 Histogram that shows distribution of two groups

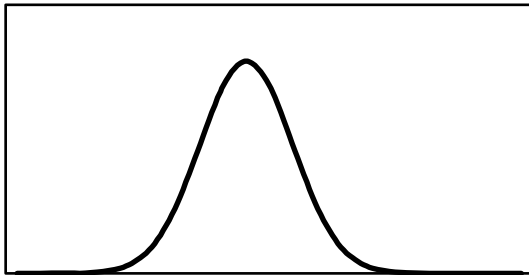


Figure 3 Histogram of residuals

it could be difficult to relate the distribution of the residuals to the distribution of the outcome variable.

One particular reason to examine residuals is to ensure that there are no **outliers** or influential cases. An outlier is a case with an unusual value, or combination of values, in a dataset – we are concerned about outliers because that individual will have undue influence on our parameter estimates.

Univariate outliers can be detected through the usual data cleaning procedures; however, there are other types of outliers, termed **multivariate outliers**, which cannot be detected through such methods (see **Outlier Detection**). Consider the dataset shown in Table 1 – but with one additional subject, a 15-year-old, who scores 33 on the test. The 15-year-old is younger than the other participants, but not excessively younger, and 33 is the highest score on the test, but not excessively high (see Table 2). From the **scatterplot** in Figure 4, we see that this individual (indicated on the chart with the solid black spot) does seem to lie outside of the main set of points.

If we reanalyze our regression with this additional case, we find that the regression coefficient has dropped to 0.067 (95% CI $-0.173, +0.308$;

Table 2 Data including outlier observation, new predicted values and residuals

Age (<i>a</i>)	Number of items correct (<i>x</i>)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)
15	33	21.32	11.68
20	15	21.65	−6.65
25	21	21.99	−0.99
30	19	22.33	−3.33
35	18	22.66	−4.66
40	21	23.00	−2.00
45	25	23.34	1.66
50	22	23.67	−1.67
55	26	24.01	1.99
60	31	24.35	6.65
65	22	24.68	−2.68

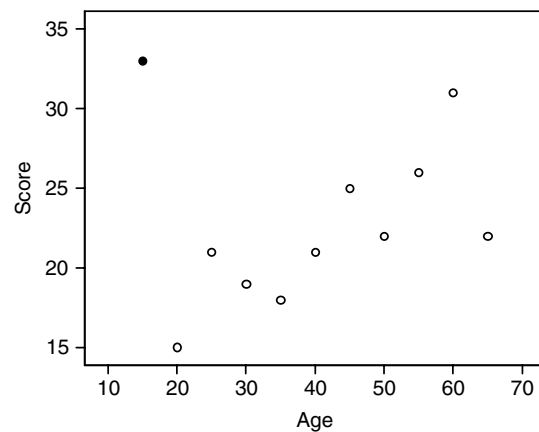


Figure 4 Scatterplot of score versus age. Solid black spot indicates a potential outlier

$p = 0.542$). This individual has obviously had considerable influence on our analysis and has caused us to dramatically alter our conclusions. Table 2 shows the new predicted scores and residuals. Simply scanning the residuals by eye shows that the first case does have a residual that is higher than the others. We could also view these data using a **histogram** or a **boxplot**, such as in Figure 5.

Transformation of Residuals

Standardized Residuals. Residuals in their raw form are difficult to interpret, as the scale is that of the outcome variable – they are not in a common metric that we can interpret. Residuals can be transformed in a number of different ways to a metric that we

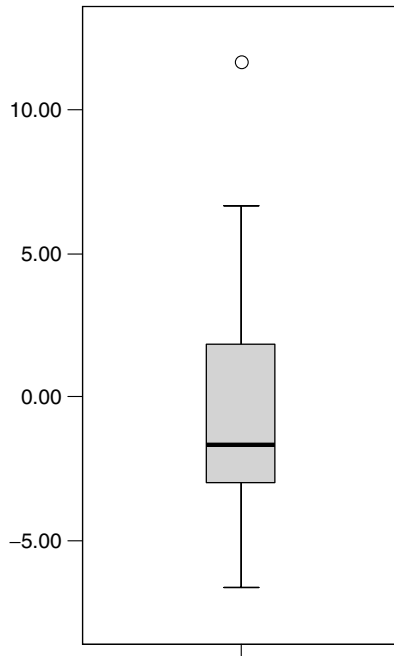


Figure 5 Boxplot of residuals, showing outlier

understand better, for example, to a normal or a t distribution. We should note from the outset that there are differences among the names given for these different transformations by different books and computer programs.

An additional problem is the lack of common names for the different types of residuals, which I shall point out along the way. (A further warning is required here – computer programs are also inconsistent with their naming – different programs will use different names for the same thing, and the same name for different things. It is important to know exactly what the program is doing when using any of these statistics.)

The first type of transformation is to divide the residuals by their standard deviation, to produce residuals with a standard normal distribution (mean = 0, standard deviation = 1) in a large sample, and t distribution (with df equal to $n - k - 2$) in a smaller sample. These are also referred to as *unit normal deviate* residuals by Draper and Smith [2].

The standardized residual (e^s) for each case is given by:

$$e_i^s = \frac{e_i}{s_e} \quad (5)$$

where e_i is the residual for each case, and s_e is the standard deviation of the residuals. (Also be warned that some references, for example, Fox [3], use the term ‘standardized residual’ to refer to what we shall call ‘studentized residuals’.) Note that the standard deviation of the residuals is given by:

$$s_e = \sqrt{\frac{\sum e^2}{n - k - 1}} \quad (6)$$

We must incorporate k , the number of predictor variables, into the equation as well as the sample size. (Although in large samples, this will have a negligible effect on the calculation; given that we are almost certainly using a computer, we might as well do it properly.)

Because standardized residuals follow an approximately standard normal distribution, we can make statements about the likelihood of different values arising. We would expect approximately 1 case in 20 to have an absolute value greater than 2, 1 case in 100 to fall outside an absolute value of 2.6, and 1 in 1000 to fall beyond 3.1. If, therefore, we have a case with an absolute standardized residual of 3, in a sample size of 50, we should consider looking at that case.

In the data of people’s ages and their scores, the sum of squared residuals is equal to 279.5, and the standard deviation is therefore equal to the square root of $279.5/(11 - 1 - 1)$, which gives 5.57. Dividing each of the raw residuals by 5.57 will therefore give the standardized residuals (see column 5 of Table 3).

Studentized Residuals. There is a problem with the use of standardized residuals; some [4] have argued that they should not be called standardized residuals at all. The problem that we need to address is that the variances of the residuals are not equal, as we have considered them to be. The variance of the residual is dependent on the scores on the predictor variables. In the case of analyses with one predictor variable, the variance of the residual depends on the distance of the predictor variable from its mean – extreme scores on the predictor variable are associated with lower variance of the residuals. In the multiple predictor case, the distance from the centroid of all predictor variables is used to ascertain the variance of the residuals.

The distance from the mean or from the centroid of all predictor variables is called the *leverage*, (see

4 Residuals

Table 3 Predicted scores, residuals, and hat (leverage) values for original data

Age (<i>a</i>)	Number of items correct (<i>x</i>)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)	Standardized residual	Leverage (hat value)	Studentized residual	Studentized deleted residual
20	15	16.98	-1.98	2.10	0.32	2.54	3.00
25	21	18.10	2.90	-1.19	0.25	-1.37	-1.28
30	19	19.21	-0.21	-0.18	0.18	-0.20	-0.19
35	18	20.33	-2.33	-0.60	0.14	-0.64	-0.62
40	21	21.44	-0.44	-0.84	0.11	-0.88	-0.85
45	25	22.56	2.44	-0.36	0.10	-0.38	-0.37
50	22	23.67	-1.67	0.30	0.11	0.31	0.32
55	26	24.79	1.21	-0.30	0.14	-0.32	-0.32
60	31	25.90	5.10	0.36	0.18	0.39	0.40
65	22	27.02	-5.02	1.19	0.25	1.37	1.49

Leverage Plot) or the *hat* value, or h , and is the diagonal element of the hat matrix. The minimum value of h_i is $1/n$, the maximum is 1. As might be expected by now, however, there is not complete agreement about the nomenclature – the leverage has a second form, the centred leverage, h^* , which has a minimum value of zero, and a maximum of $(N - 1)/N$.

Most computer programs use the leverage (h); SPSS, however, uses the centred leverage (h^*) but refers to it as ‘leverage’ – the only clue is that the help file says that the minimum value is zero and the maximum is $(N - 1)/N$.

When we have the leverage values, these can be used to correct the estimate of the standard deviation of the residuals and calculate the studentized residual (e'), using the following equation.

$$e'_i = \frac{e_i}{s_e \sqrt{1 - h_i}} \quad (7)$$

And once again, we note the different titles that are given to these – Cohen et al. [1] call these *internally studentized residuals*, Fox [3] calls these *standardized residuals*, and Draper and Smith [2], to ensure that there really is no confusion, call them the $e_i/\{(1 - \mathbf{R}_{ii})s^2\}^{1/2}$ form of the residuals (noting of course, that they choose to refer to the hat matrix as **R** rather than **H**).

The studentized residuals for our example are shown in Table 3.

Studentized Deleted Residuals. We have dealt with one problem with the residuals, but we have yet another. The residuals do not quite follow a t distribution – if we wanted to use the residuals to

ascertain the probability of such a value arising, we need to calculate the studentized deleted residuals.

The standardized residual and the studentized residual are calculated on the basis of the standard deviation of the residuals. However, the residual has an influence on the standard deviation of the residuals, in two ways. First, where the residual is large, the standard deviation will be larger, because that residual will be used in the calculation of the standard deviation – this will have the effect of shrinking the residual. Second, where the case has had an influence on the regression estimates (as is likely if it is an outlier), the regression line will be drawn toward the case, and so the size of the residual will again shrink.

The solution is to remove the case from the dataset, run the regression analysis again, estimate the regression line and the standard deviation of the residuals, then replace the case, and calculate the standardized and studentized residuals. In fact, this can be found using the equation given earlier for studentized residual e' except that s_e is calculated with the offending case omitted. The deleted studentized residuals for our example can be seen in Table 3.

As would be expected by now, the deleted studentized residual is also known by other names, principally as the *externally studentized residual*, but also as the *jackknifed residual*.

The deleted studentized residual is distributed as a t distribution, with $n - k - 2$ degrees of freedom. The null hypothesis that we can test using this value is that a case with a studentized residual that large should have arisen by chance. We can examine the largest absolute studentized deleted residual and calculate the associated probability (it is unlikely that tables will be of much use here, you will need to use

a computer). The probability associated with a value of t of 3.00 is 0.017 – we would therefore consider that there was support for the hypothesis that the case was not sampled from the same population as the other residuals and should consider this case to be an outlier. Using this criterion, and a 5% cut off, we will find that 5% of our cases are outliers – obviously not a useful way to proceed. The alternative is to correct this probability by using Bonferroni correction – that is, we either multiply the probability associated with the case by N , or alternatively (and equivalently) we divide the cutoff that we are using (typically 0.05) by N . Taking the first approach, multiplying the probability by N (which is 11), we find the Bonferroni corrected probability to be $0.017 \times 11 = 0.19$. This is above the cutoff of 0.05, and therefore we do not have sufficient evidence against the null hypothesis that the case is sampled from the same population. (Equivalently, we could divide our cutoff of 0.05 by 11, to give a new value of 0.0045; by this

criteria also, we do not have evidence against the null hypothesis.) This analysis has, of course, been affected by the smallness of the size of the sample, which was purely for illustrative purposes.

References

- [1] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.
- [2] Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Edition, John Wiley & Sons, New York.
- [3] Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, Thousand Oaks.
- [4] Ryan, T.P. (1997). *Modern Regression Methods*, John Wiley & Sons, New York.

JEREMY MILES

Residuals in Structural Equation, Factor Analysis, and Path Analysis Models

TENKO RAYKOV

Volume 4, pp. 1754–1756

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Residuals in Structural Equation, Factor Analysis, and Path Analysis Models

Several types of residuals have been considered for **structural equation, factor analysis, and path analysis models**. They represent different aspects of discrepancies between model and data.

The traditional residuals, which can be referred to as *covariance structure residuals* (CSRs), are defined as element-wise differences between the empirical covariance matrix $\mathbf{S}$ for a given set of observed variables, $Z_1, Z_2, \dots, Z_k$, and the implied (reproduced) covariance matrix $\Sigma(\theta)$ by a fitted model, M , where θ is the vector of model parameters ($k > 1$). Denoting the CSRs by r_{ij} in an empirical application they are obtained as $r_{ij} = s_{ij} - \sigma_{ij}(\hat{\theta})$, where s_{ij} and $\sigma_{ij}(\hat{\theta})$ symbolize the elements in i th row and j th column in $\mathbf{S}$ and $\Sigma(\hat{\theta})$, respectively, and $\Sigma(\hat{\theta})$ designates the model-implied covariance matrix at the parameter estimator ($i, j = 1, \dots, k$). The CSRs play an important role in evaluating the overall fit of the model. In particular, the chi-square goodness-of-fit test statistic (see **Goodness of Fit for Categorical Variables**) when testing adequacy of M – that is, the null hypothesis that there exists a point θ^* in the parameter space of M , such that $\Sigma(\theta^*) = \Sigma^*$, where Σ^* is the population covariance matrix of $Z_1, Z_2, \dots, Z_k$ – can be viewed, for practical purposes, as a weighted sum of the CSRs. Specifically, when **maximum likelihood estimation** (ML) method is used, the fit function $F_{ML} = -\ln |\mathbf{S}(\Sigma(\theta))^{-1}| + tr(\mathbf{S}(\Sigma(\theta))^{-1}) - k$ (e.g., [2]; $|\cdot|$ designates determinant and $tr(\cdot)$ matrix trace) can be approximated by the sum of squared CSRs weighted by corresponding elements of the inverse implied covariance matrix, $[\Sigma_{ML}(\hat{\theta})]^{-1}$; these residuals are similarly related to fit functions with other methods (e.g., [3]). When the mean structure is analyzed – that is, M is fitted to the covariance matrix $\mathbf{S}$ and means $m_1, m_2, \dots, m_k$ of $Z_1, Z_2, \dots, Z_k$, respectively – also *mean structure residuals* (MSRs) can be obtained as $m_i - \mu_i(\hat{\theta})$, where $\mu_i(\hat{\theta})$ is the i th variable mean implied by the model at the parameter estimator ($i = 1, \dots, k$). The CSRs and MSRs contain information about the location and degree of lack of fit of M , and in this sense may be considered local indices of fit (see also

below). Standardized CSRs larger than 2 in absolute value may be viewed as indicative of a considerable lack of fit of M at least with regard to the pair of variables involved in each such residual; for a given model, however, these residuals are not independent of one another, and thus caution is advised when more than a few such residuals are examined in this way. Generally, a large positive (standardized) CSR may suggest that introduction of a parameter(s) further contributing to the relationship between the two variables involved may lead to model fit improvement; similarly, a negative (standardized) CSR with large absolute value may suggest that deleting or modifying the value of a parameter(s) currently involved in the variables' relationship may lead to marked improvement of model fit. Residuals discussed thus far can be routinely obtained for structural equation, factor analysis, and path analysis models with popular structural equation modeling software, such as LISREL, EQS, MPLUS, SEPATH, RAMONA, and SAS PROC CALIS (see **Structural Equation Modeling: Software**).

In addition to these residuals, *individual case residuals* (ICRs) can also be obtained that pertain to each studied individual (case) and dependent variable. In a path analysis model, say $\mathbf{Y} = \mathbf{B}\mathbf{Y} + \mathbf{\Gamma}\mathbf{X} + \mathbf{E}$ – where $\mathbf{Y}$ is a vector of dependent observed variables, $\mathbf{X}$ is that of independent observed variables, $\mathbf{E}$ is the vector of error terms, and $\mathbf{B}$ and $\mathbf{\Gamma}$ are corresponding coefficient matrices (such that $\mathbf{I} - \mathbf{B}$ is invertible, where $\mathbf{I}$ is the identity matrix of appropriate size) – ICRs result as $\mathbf{r}_p = \mathbf{Y}_p - (\mathbf{I} - \mathbf{B})^{-1} \mathbf{\Gamma} \mathbf{X}_p$, where $\mathbf{X}_p$ and $\mathbf{Y}_p$ are the vectors of the p th individual's values on the independent and dependent variables, respectively, and $\mathbf{r}_p$ is the vector of his/her residuals (that is of the same size as $\mathbf{Y}$; $p = 1, \dots, N$, with N denoting sample size). Estimates of these ICRs are furnished by substituting $\mathbf{B}$ and $\mathbf{\Gamma}$ in the last equation with their estimates obtained when fitting the model (cf. [4]). In a factor analysis model, $\mathbf{Y} = \Lambda \boldsymbol{\eta} + \boldsymbol{\epsilon}$ – where $\boldsymbol{\eta}$ is the vector of latent factors (fewer in number than observed variables), Λ is the factor loading matrix, and $\boldsymbol{\epsilon}$ that of error terms – ICRs are obtained as $\mathbf{r}_p = \mathbf{Y}_p - \Lambda \mathbf{f}_p$, where $\mathbf{f}_p$ is the vector of factor scores pertaining to the p th case ($p = 1, \dots, N$; [1, 5]). Estimates of these residuals are furnished when substituting Λ in the last equation with its estimate (along with the factor score – e.g., Bartlett – estimates) obtained when fitting the model. In a structural equation model

(with fewer factors than observed variables), $\mathbf{Y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}$ – where $\mathbf{B}$ is the structural regression matrix and $\boldsymbol{\zeta}$ the vector of latent disturbances – ICRs are obtained as $\mathbf{r}_p = \mathbf{Y}_p - \Lambda(\mathbf{I} - \mathbf{B})^{-1}\mathbf{g}_p$, where $\mathbf{g}_p$ is the vector of the p th individual's factor scores for the model $\mathbf{Y} = \Lambda(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta} + \mathbf{e}$ ($p = 1, \dots, N$; $\mathbf{e}$ being model error term). Estimates of these ICRs are furnished when in the equation $\mathbf{r}_p = \mathbf{Y}_p - \Lambda(\mathbf{I} - \mathbf{B})^{-1}\mathbf{g}_p$ estimates of Λ and $\mathbf{B}$ (along with factor score estimates) are substituted ([1, 5]; $p = 1, \dots, N$). Extended individual case residuals (EICRs) for structural equation and factor analysis models with less latent than observed variables are discussed in Raykov & Penev [5]. The EICRs represent ICRs that are obtained using a nonorthogonal projection with a model error covariance matrix, and have been shown to differ across particular equivalent models [5]. Latent individual residuals (LIRs) are discussed in Raykov & Penev [6], which reflect individual case residuals with regard to a latent relationship, and can be used for purposes of studying latent variable relationships, as exemplified in Raykov & Penev [6].

Acknowledgment

I thank M. W. Browne, R. E. Millsap, and D. Rindskopf for valuable discussions on residual evaluation. This work

was supported by the Research Institute on Addictions with the State University of New York.

References

- [1] Bollen, K.A. & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models, in *Sociological Methodology*, P.V. Marsden, ed., Jossey-Bass, San Francisco, pp. 235–262.
- [2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [3] Browne, M.W., MacCallum, R.C., Kim, C.-T., Andersen, B.L. & Glaser, R. (2002). When fit indices and residuals are incompatible, *Psychological Methods* 7, 403–421.
- [4] McDonald, R.P. & Bolt, D.M. (1998). The determinacy of variables in structural equation models, *Multivariate Behavioral Research* 33, 385–402.
- [5] Raykov, T. & Penev, S. (2001). The problem of equivalent structural equation models: an individual residual perspective, in *New Developments and Techniques in Structural Equation Modeling*, G.A. Marcoulides & R.E. Schumacker, eds, Lawrence Erlbaum, Mahwah, pp. 297–321.
- [6] Raykov, T. & Penev, S. (2002). Exploring structural equation model misspecifications via latent individual residuals, in *Latent Variable and Latent Structure Models*, G.A. Marcoulides & I. Moustaki, eds, Lawrence Erlbaum, Mahwah, pp. 121–134.

TENKO RAYKOV

Resistant Line Fit

PAT LOVIE

Volume 4, pp. 1756–1758

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Resistant Line Fit

A simple way of representing a linear relationship between two variables in a functional form on a **scatterplot** is to fit a *resistant line*. The aim is to find a line that makes the deviations in the y direction (the **residuals**) as small as possible while resisting the influence of any outlying points (**outliers**).

The initial line is found by splitting the n ordered x values (with their associated y values) into three approximately equal sized groups or batches. Any observations with the same x value should be kept together in the same batch and, ideally, the left (small x) and right (large x) batches should each contain the same number of observations. The **medians** for the x and associated y values in the left and right batches (x_L, x_R, y_L, y_R) determine the line through the point cloud. (Note that the pairing of the x and y values is ignored when finding the medians.)

If all we require is a quick visual indication of fit, then the initial resistant line is that which passes through the left and right batch x and y medians on the scatterplot; if the middle batch median deviates markedly from the line, this suggests possible nonlinearity.

Rough values of the slope b and intercept a for the resistant line can be obtained directly from the graph to give an equation for the fitted y values $\hat{y}_i = a + bx_i$ ($i = 1, \dots, n$). Alternatively, the coefficients can be found, using the medians, from the expressions $b = (y_R - y_L)/(x_R - x_L)$ and $a = \text{mean}(y_R - bx_R, y_M - bx_M, y_L - bx_L)$.

We can also assess linearity by looking at the half-slopes instead of ‘by eye’. The left half-slope is the slope of the line joining the medians of the left and middle batches. Similarly, the right half-slope is that of the line between the medians of the middle and right batches. A minor modification of the slope formula given above is all that is required to obtain these values. If one of the half-slopes is more than twice the other, that is, the half-slope ratio is greater than two, we should not be attempting to fit a straight line.

A polishing routine can then be used to adjust the line if the residuals (the differences between the observed and fitted y values) show any distinct pattern on a scatterplot of the residuals against the x values (or indeed on a **stem and leaf plot**). In

essence, what we try to do is to balance up the size of the (median) residuals in the left and right batches.

To polish the line, we go through the same initial steps of the fitting process but this time using the residuals as new y values and find the slope b_{res} . Next, we adjust b by adding the value b_{res} to it, then recalculate the intercept and finally obtain the residuals for this new fitted line. The procedure is repeated if necessary until the residuals show no evidence of a relationship with the x values, that is, they have zero slope. Generally, the slope changes by smaller and smaller amounts at each iteration, eventually converging to a stable value. Clearly, these iterations are tedious to do by hand and thus are best left to statistical packages such as Minitab (see **Software for Statistical Analyses**).

As an illustration, suppose that our data (see Table 1) are Semester 1 and Semester 2 examination scores (each out of a maximum of 50) achieved by a sample of 11 students on a two-semester elementary statistics course. Naturally, we have already inspected a scatterplot (see Figure 1) and are fairly confident that there is a linear relationship between the variables; we note, however, one observation (10, 39) that is distinctly out of line with the remainder of the data. To fit a resistant line with Semester 1 score on the x axis, we split the data into three batches and find the medians for the x and y values:

The initial resistant line determined by the medians is superimposed on the scatterplot in Figure 1. It seems to pass a little too high of center for small values of x (apart from the ‘outlier’), although we have

Table 1 Semester 1 and Semester 2 examination scores on an elementary statistics course

x (Semester 1)		y (Semester 2)	
10	$x_L = 17$	39	$y_L = 21$
15		17	
19		21	
20	$x_M = 28$	21	$y_M = 29$
25		29	
28		37	
30	$x_R = 39$	24	$y_R = 39$
32		33	
38		36	
40		42	
44		43	

2 Resistant Line Fit

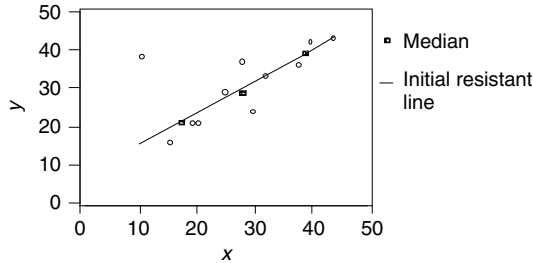


Figure 1 Scatterplot with initial resistant line

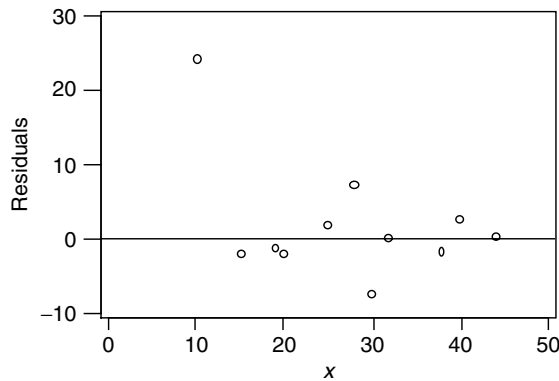


Figure 2 Residual plot for initial resistant line fit

to be careful not to be too dogmatic because our data set is very small. Also, as the line does not miss the middle batch median by much, nonlinearity is not a worry.

The equation of the initial resistant line calculated from the slope and intercept formulae given earlier is $\hat{y}_i = 6.76 + 0.82x_i$ ($i = 1, 2, \dots, 11$). Although it is difficult to discern any pattern in the residual plot (Figure 2), except that the residuals seem to be a little larger for middle sized x values, we shall allow Minitab to do some polishing for us. Minitab takes a few further iterations to come up with a final equation $\hat{y}_i = 3.81 + 0.91x_i$ ($i = 1, 2, \dots, 11$). In passing, we note that the half slope ratio is 1.25, confirming a fairly linear relationship.

Superimposing this polished line on the scatterplot (Figure 3), we see that new line has been shifted downward on the left side and now passes a little closer to the points with smaller x values and even further away from our outlier. This seems a reasonable fit to the data and would allow us to make

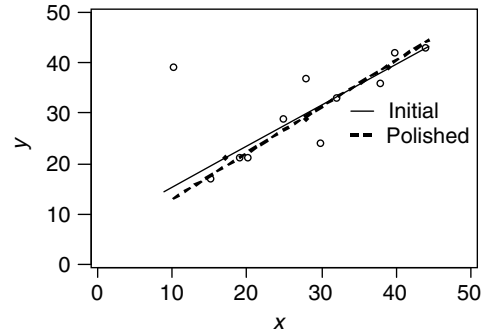


Figure 3 Scatterplot with initial (solid) and polished (dotted) resistant lines

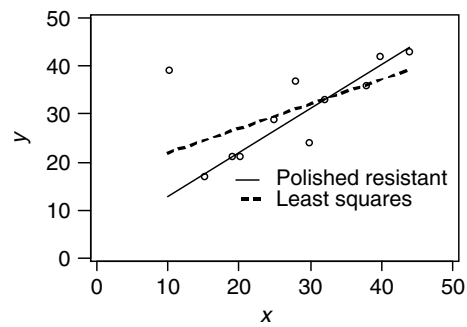


Figure 4 Scatterplot with polished resistant (solid) and least squares (dotted) lines overlaid

some cautious predictions about performance at the end of Semester 2.

Suppose though that we had just pressed ahead with traditional least squares **regression** for these data. The equation of the least squares line is $\hat{y}_i = 17.38 + 0.50x_i$ ($i = 1, 2, \dots, 11$), which has a considerably higher intercept and a smaller slope than our polished resistant line (see Figure 4). The least squares line has been pulled toward the observation (10, 39), which we had already branded as an outlier.

To overcome the effect of this anomalous observation, we shall have to omit it from the analysis. The equation of regression line then becomes $\hat{y}_i = 4.60 + 0.88x_i$ ($i = 1, 2, \dots, 10$), which is quite close to our resistant line solution.

As we have seen, resistant line fitting can be a useful alternative to the least squares method

for describing the relationship between two variables when the presence of outliers makes the latter procedure risky. However, it is essentially an exploratory technique and lacks the inferential framework of traditional regression analysis. Further discussion of resistant line fitting can be found in [1].

Reference

- [1] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

PAT LOVIE

Retrospective Studies

DEAN P. MCKENZIE

Volume 4, pp. 1758–1759

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Retrospective Studies

A retrospective study is a type of **observational study** (no random assignment to groups) in which the events being studied have already occurred, before the study has begun. Information on such events may already have been collected, perhaps for administrative purposes (e.g., employment records), or it may be obtained by questioning individuals. For example, Hart et al. [6] interviewed the carers of patients diagnosed with Alzheimer's disease. The interviewees were asked about the presence and duration of patients' symptoms such as anxiety, which had occurred in the past three years. **Case-control studies** are generally retrospective, indeed, the term retrospective study originally referred to this type of design [8]. The past experiences, (e.g., smoking behavior) of those identified as having a particular disease (e.g., lung cancer) or outcome (the cases), are compared with those who do not (the controls). A further type of retrospective study is the historical, or retrospective, **cohort study** [5]. A group or cohort of individuals is identified, based on characteristics found in historical databases such as accommodation, employment, medical (including case notes and charts), military or school records. Information on exposures and disease or outcome of interest is obtained from these or other records, and the cohort is followed up over 'historical time'. In other words, the cohort will be studied from a point backward in time, up to the more recent past, or present. For example, Zammit et al. [10] studied a historical cohort of over 50 000 Swedish males, who had originally been conscripted for compulsory military training during 1969–1970. A variety of demographic and other information, including self-reported cannabis use, had been collected and stored. Zammit et al. [10] sought to establish whether cannabis use (the exposure) was a risk factor for schizophrenia (the disease). The original records were accessed and linked to historical psychiatric diagnostic information, obtained from the Swedish hospital discharge registry, for the period 1970–1996. The **odds** of developing schizophrenia over this period for persons reporting cannabis use during 1969–1970 were then compared with the odds for those who did not. In contrast, a prospective cohort study (*see Prospective and Retrospective Studies*) is one in which individuals are selected

based on their current, rather than their past, characteristics. Information is then collected from the beginning of the study until some point in the future. Prospective studies readily allow the directionality of events to be examined (e.g., is cannabis use a consequence of psychiatric illness rather than a probable cause? [2]), but suffer from the problem of drop-outs (a particular problem in studies involving illicit drug users [3]). Prospective studies can be extremely expensive and time-consuming. This is especially true if a large cohort has to be followed up over many years until the event of interest (e.g., Alzheimer's disease) occurs [1, 9, 4], in which case retrospective studies may be the only practical option. Retrospective studies readily and (comparatively) inexpensively allow the analysis of many thousands of individuals, over several decades, provided that the necessary historical information has been collected, or can be remembered. Memories may be highly inaccurate [1], and database information incomplete. Changes in measurement and diagnostic methods may also have occurred over time. For events such as suicide, however, there may be no alternative to the use of retrospective studies [1]. In practice, the distinction between prospective and retrospective studies can be somewhat blurred. A prospective cohort study might have a retrospective component, for example, individuals may be initially asked about their childhood experiences. Finally, either a prospective, or retrospective, cohort study, may provide data for a nested case-control study [5, 9] in which cohort members identified as having a particular diagnosis (e.g., depression) or outcome are compared with those who do not. Further information about retrospective studies can be found in [1], [4], [5], [8] and [9] with a light-hearted example given in [7].

References

- [1] Anstey, K.J. & Hofer, S.M. (2004). Longitudinal designs, methods and analysis in psychiatric research, *Australian and New Zealand Journal of Psychiatry* **38**, 93–104.
- [2] Arseneault, L., Cannon, M., Poulton, R., Murray, R., Caspi, A. & Moffitt, T.E. (2002). Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study, *British Medical Journal* **325**, 1212–1213.
- [3] Bartu, A., Freeman, N.C., Gawthorne, G.S., Codde, J.P. & Holman, C.D.J. (2004). Mortality in a cohort of opiate and amphetamine users in Perth, Western Australia, *Addiction* **99**, 53–60.

2 Retrospective Studies

- [4] de Vaus, D.A. (2001). *Research Design in Social Research*, Sage Publications, London.
- [5] Doll, R. (2001). Cohort studies: history of the method II. retrospective cohort studies, *Sozial – und Präventivmedizin / Social and Preventive Medicine* **46**, 152–160.
- [6] Hart, D.J., Craig, D., Compton, S.A., Critchlow, S., Kerrigan, B.M., McIlroy, S.P. & Passmore, A.P. (2003). A retrospective study of the behavioural and psychological symptoms of mid and late phase Alzheimer's disease, *International Journal of Geriatric Psychiatry* **18**, 1037–1042.
- [7] Nelson, M.R. (2002). The mummy's curse: historical cohort study, *British Medical Journal* **325**, 1482–1484.
- [8] Paneth, N., Susser, E. & Susser, M. (2002). Origins and early development of the case-control study: part 1, early evolution, *Sozial – und Präventivmedizin / Social and Preventive Medicine* **47**, 282–288.
- [9] Tsuang, M.T. & Tohen, M. (2002). *Textbook in Psychiatric Epidemiology*, 2nd Edition, Wiley-Liss, Hoboken.
- [10] Zammit, S., Allebeck, P., Andreasson, S., Lundberg, I. & Lewis, G. (2002). Self reported cannabis use as a risk factor for schizophrenia in Swedish conscripts of 1969: historical cohort study, *British Medical Journal* **325**, 1199–1201.

DEAN P. MCKENZIE

Reversal Design

JOHN FERRON

Volume 4, pp. 1759–1760

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Reversal Design

Reversal designs [1] are a type of **single-case design** used to examine the effect of a treatment on the behavior of a single participant. The researcher measures the behavior of the participant repeatedly during what is referred to as the baseline phase. After the baseline has been established, the researcher implements the treatment and continues to repeatedly measure the behavior during the treatment phase. The researcher then removes the treatment and reestablishes the baseline condition. Since the treatment is withdrawn during this third phase, many refer to this type of design as a withdrawal design.

One typically expects the behavior will be stable during the baseline phase, improve during the treatment phase, and then reverse, or move back toward baseline levels during the second baseline phase. When improvement is seen during the treatment phase, some may question whether this improvement was the result of the treatment or whether it resulted from maturation or some event that happened to coincide with treatment implementation. If we see the behavior return to baseline levels during the second baseline phase, these alternative explanations seem less plausible, and we feel more comfortable attributing the behavior changes to the treatment. Put another way, the reversal increases the **internal validity** of the design.

The minimum number of phases in a reversal design is three: a baseline phase (A), followed by a treatment phase (B), followed by the second baseline phase (A). It is possible, however, to extend the basic ABA design to include more phases, creating other phase designs, such as an ABAB design or an ABABAB design.

Reversal designs also vary in how the phase shifts are determined. In some cases, the assignment is systematic, for example, a researcher may decide that there will be eight observations in each phase. This method works well when baseline observations are constant, which allows one to assume temporal

stability and use what has been referred to as the scientific solution to causal inference. When baseline observations are not constant, inferences become more difficult and one may alter the method of assigning phase shifts to facilitate drawing treatment–effect inferences.

One option is to use a response-guided strategy in which the data are viewed and conditions are changed after the researcher judges the data within a phase to be stable. If the researcher is able to identify and control the factors leading to variability in the baseline data, the variability can be reduced, hopefully leading to constancy in the later baseline observations and relatively straightforward inferences about treatment effects (*see Multiple Baseline Designs*).

When baseline variability cannot be controlled, one may turn to statistical methods for making treatment–effect inferences, which may motivate the use of some restricted form of **randomization** to choose the intervention points. For example, the intervention points could be chosen randomly under the constraint that there are at least five observations in each phase. A randomization test (*see Randomization Based Tests*) [3] could then be constructed to allow inference about the presence of a treatment effect. To make inferences about the size of a treatment effect when confronted with variable baseline data, one could turn to statistical modeling options (*see Statistical Models*) [2].

References

- [1] Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **1**, 91–97.
- [2] Huitema, B.E. & McKean, J.W. (2000). Design specification issues in time-series intervention models, *Educational and Psychological Measurement* **60**, 38–58.
- [3] Onghena, P. (1994). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder, *Behavioral Assessment* **14**, 153–171.

JOHN FERRON

Risk Perception

BRIAN S. EVERITT

Volume 4, pp. 1760–1764

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Risk Perception

Alistair Cooke in *A Letter from America* during the Gulf War of 1990 told the sad story of an American family of four who cancelled their planned holiday to Europe because of the fears of terrorist attacks on the country’s airlines. They decided to drive to San Francisco instead. At the last junction, before leaving their small Midwest town, they collided with a large truck with fatal consequences for them all.

Life is a risky business and deciding which risks are worth taking and which should be avoided has important implications both for an individual’s lifestyle, and the way society operates. The benefits gained from taking a risk need to be weighed against the possible disadvantages. An acceptable risk is proportional to the amount of benefits. For the individual, living life to the fullest means achieving a balance between reasonable and unreasonable risk, and this balance is dependent largely on the individual’s personality. But, in society as a whole, where the same balancing act is required, it has to be achieved through political action and legislation. If risks could be assessed and compared in a calm and rational manner, it would benefit both individuals and society. There is, however, considerable evidence that such assessment and comparison is not straightforward.

Defining and Quantifying Risk

A dictionary definition of risk is ‘the possibility of incurring misfortune or loss’. (The word risk is derived from the Greek word ‘rhiza’, which refers to such hazards of sailing as: too near the cliffs, contrary winds, turbulent downdraughts, and swirling tides (see **Relative Risk**)). Quantifying risk and assessing risk involves the calculation and comparison of probabilities, although most expressions of risk are compound measures that describe both the probability of harm and its severity. Americans, for example, run a risk of about 1 in 7000 of dying in a traffic accident. This probability is derived by noting that in the year 2000 there were about 40 000 traffic accident deaths among a population of about 280 000 in the United States. The figure of 1 in 7000 expresses the overall risk to society. The risk to any particular individual clearly depends on her exposure: how much she is

on the road, where she drives and in what weather, whether she is psychologically accident prone, what mechanical condition the vehicle is in, and so on. Because gauging risk is essentially probabilistic, it follows that a risk estimate can assess the overall chance that an untoward event will occur, but it is powerless to predict any specific event. Just knowing that the probability of tossing a head with a fair coin is one-half leaves one unable to predict which particular tosses will result in heads and which in tails.

Risk Perception

Risk perception is one’s opinion of the likelihood of the risk that is associated with a certain activity or lifestyle. Risk perception is influenced by sociological, anthropological, and psychological factors, with the result that people vary considerably in which risks they consider acceptable and which they do not, even when they agree on the degree of risk involved. For example, many people with no fear of traveling large distances by car or train consider the prospect of flying, even with a well-respected commercial airline, to be a nightmare, often requiring several trips to the airport bar before being able to board the airplane. For others, air travel represents the very model of a low-risk form of transport. As Table 1 shows, flying is actually one of the safest forms of transport.

Perhaps one reason for some people’s excessive and clearly misguided fear of flying is the general view that being killed in a plane crash must be a particularly nightmarish way to die. Another possibility is that the flying phobic considers the sky an alien environment, and this consideration distorts the perception of the risk involved. A third possible reason is that flying accidents are more prominent in the media than those involving automobiles, although the latter are far more common. Perception of risk is also likely to be influenced by whether we feel in control of a perceived risk.

Table 1 Causes of death and their probability

Cause of death	Probability
Car trip across the United States	1 in 14 000
Train trip across the United States	1 in 1 000 000
Airline accident	1 in 10 000 000

2 Risk Perception

Research shows that people tend to overestimate the probability of unfamiliar, catastrophic, and well-publicized events and to underestimate the probability of unspectacular or familiar events that claim one victim at a time. For example, in one study [2], respondents rated 90 hazards, each with respect to 18 qualitative characteristics such as whether the risk was voluntary or involuntary, personally controllable or not, and known to those exposed or not. A **principal component analysis** of the data identified two major components, and the location of the 90 hazards with respect to those two components. The first factor was labeled as 'dread' risk. This factor related judgments of scales such as uncontrollability, fear, and involuntariness of exposure. Hazards that rate high on this factor include nuclear weapons, nerve gas, and crime. Hazards that rate low on this factor include home appliances and bicycles. A second factor, labeled 'unknown risk' related judgments of the observability of risks, such as whether or not the effects are delayed in time, the familiarity of the risk, and whether or not the risks are known to science. Hazards that rate high on this dimension include solar electric power, DNA research and satellites; those that rate low include motor vehicles, fire fighting, and mountain climbing.

Slovic, Fischhoff, and Lichtenstein [2] conclude that perceptions of risk are clearly related to the position of an activity in the principal component space, particularly in respect of the 'dread' factor. The higher a hazard's score on this factor, the higher its perceived risk, the more people want to see its current risks reduced, and the more they want to see strict regulation employed to achieve the desired reduction in risk. It seems that the risks that kill people and the risks that scare people are different.

The findings in [2] are supported by the results of polls of college students and members of the League of Women Voters in Oregon. Both groups considered nuclear power their number-one 'present risk of death', far riskier than motor vehicle accidents, which kill almost 42 000 Americans each year, or cigarette smoking, which kills 150 000, or handguns, which kill 17 000. Experts in risk assessment in the same poll considered motor vehicle accidents their number-one risk, with nuclear power below the risk of swimming, railroads, and commercial aviation. Here, the experts seem to have the most defensible conclusions. Average annual fatalities expected from nuclear power, according to most scientific estimates, are fewer than

10. Nuclear power does not appear to merit its risk rating of number one. It appears that the two well-educated and influential segments of the American public polled in Oregon seem to have been misinformed. The misinformants are not difficult to identify. Journalists report almost every incident involving radiation. A truck containing radioactive material is involved in an accident, a radioactive source is temporarily lost, a container leaks radioactive materials – all receive nationwide coverage, whereas the 300 Americans killed each day in other types of accidents are largely ignored. Reports in the media concentrate on issues and situations that frighten – and therefore interest – readers and viewers. The media fills its coverage with opinions (usually from interested parties) rather than facts or logical perspectives. In terms of nuclear power, for example, phrases such as *deadly radiation* and *lethal radioactivity* are common, but the corresponding *deadly cars* and *lethal water* would not sell enough newspapers, although thousands of people are killed each year in automobile accidents and by drowning. The problem is highlighted by a two-year study of how frequently different modes of death become front-page stories in the New York Times. It was found that the range was from 0.02 stories per 1000 annual deaths from cancer to 138 stories per 1000 annual deaths from airplane crashes.

Misperception of risk fueled by the media can lead to unreasonable public concern about a hazard, which can cause governments to spend a good deal more to reduce risk in some areas and a good deal less in other areas. Governments may, for example, spend huge amounts of money protecting the public from, say, nuclear radiation, but are unlikely to be so generous in trying to prevent motor vehicle accidents. They react to loudly voiced public concern in the first case and to the lack of it in the second. But, if vast sums of money are spent on inconsequential hazards, little will be available to address those that are really significant.

Examples of disparities that make little sense are not hard to find. In the late 1970s, for example, the United States introduced new standards on emissions from coke ovens in steel mills. The new rules limited emissions to a level of no more than 0.15 mg/m³ of air. To comply with this regulation, the steel industry spent \$240 million a year. Supporters of the change estimated that it would prevent about 100 deaths from cancer and respiratory disease each year, making the

average annual cost per life saved \$2.4 million. It is difficult to claim that this is money well spent when a large scale program to detect lung cancer in its earliest stages, for example, might be expected to extend the lives of 7000 cancer victims an average of one year for \$15 000 each, and when the installation of special cardiac-care units in ambulances could prevent 24 000 premature deaths a year at an average cost of a little over \$200 for each year of life.

Politicians often sanction huge expenditures to save an identifiable group of trapped miners, but not to improve mine safety or to reduce deaths among the scores of anonymous miners who die from preventable work related causes each year. Politically, at least, Joseph Stalin might have got it just about right when he mused that a single death is a tragedy, but a million deaths is a statistic.

Risk Presentation

The ability to make a rational assessment of risk is important for the individual, but also for a society that hopes to be governed by sensible, justifiable policies and legislation. It is unfortunate that, in general, people have both a limited ability to interpret probabilistic information and a mistrust of experts (sadly, statisticians in particular). But, rather than dismissing public understanding of technical issues as being insufficient for 'rational' decision making, experts (including statisticians) need to make a greater effort to increase the public's appreciation of how to evaluate and compare risks. Risks can be presented in ways that make them more transparent. For example, risks presented as annual fatality rates per 100 000 persons fail to reflect the fact that some hazards such as pregnancy and motor cycle accidents cause death at a much earlier age than other hazards such as lung cancer caused by smoking. One way to overcome this problem is to calculate the average loss of life expectancy due to exposure to the hazard. Table 2 gives some examples of risks presented in this way.

So, for example, the average age at death for unmarried males is 3500 days younger than the corresponding average for men who are married. This does not, of course, imply a cause (marrying) and effect (living 10 years longer) relationship that is applicable to every individual, although it does, in general terms at least, imply that the institution of marriage is 'good' for men. And, the ordering in Table 2 should

Table 2 Life expectancy reduction from a number of hazards

Risk	Days lost
Being unmarried – male	3500
Cigarette smoking – male	2250
Heart disease	2100
Being unmarried – female	1600
Cancer	980
Being 20% overweight	900
Low socioeconomic status	700
Stroke	520
Motor vehicle accidents	207
Alcohol	130
Suicide	95
Being murdered	90
Drowning	41
Job with radiation exposure	40
Illicit drugs	18
Natural radiation	8
Medical X rays	7
Coffee	6
Oral contraceptives	5
Diet drinks	2
Reactor accidents	2
Radiation from nuclear industry	0.02

largely reflect society's and government's ranking of priorities for increasing the general welfare of its citizens. Thus, rather than introducing legislation about the nuclear power industry or diet drinks, a rational government should set up computer dating services that stress the advantages of marriage (particularly for men) and encouraging people to control their eating habits. It is hard to justify spending any money or effort on reducing radiation hazards or dietary hazards such as saccharin.

Perhaps the whole problem of the public's appreciation of risk evaluation and risk perception would diminish if someone could devise a simple scale of risk akin to the *Beaufort scale* for wind speed or the *Richter scale* for earthquakes. Such a 'riskometer' might provide a single number that would allow meaningful comparisons among risks from all types of hazards, whether they be risks due to purely voluntary activities (smoking and hang gliding), risks incurred in pursuing voluntary, but virtually necessary, activities (travel by car or plane, eating meat), risks imposed by society (nuclear waste, overhead power lines, legal possession of guns), or risks due to acts of God (floods, hurricanes, lightning strikes).

4 Risk Perception

Table 3 Risk index values based on Paulos [1]. The lower the number, the greater the risk

Event or activity	Risk index
Playing Russian roulette once a year	0.8
Smoking 10 cigarettes a day	2.3
Being struck by lightning	6.3
Dying from a bee sting	6.8

One such scale is described by Paulos [1], and is based on the number of people who die each year pursuing various activities. If one person in N dies, the associated risk index is set at $\log_{10} N$; 'high' values indicate hazards that are not of great danger, whereas 'low' values suggest activities to be avoided if possible. (A logarithmic scale is used because the risks of different events and activities differ by several orders of magnitudes.) If everybody who took part in a particular pursuit or was subjected to a particular exposure died, then Paulos's risk index value becomes zero, corresponding to certain death. (Life is an example of such a deadly pursuit.) In the United Kingdom, 1 in every 8000 deaths each year results from motor vehicle accidents; consequently, the risk index value for motoring is 3.90. Table 3 shows examples of values for other events.

Paulos's risk index would need refinement to make it widely acceptable and practical. Death, for example, is not the only concern; injury and illness are also important consequences of exposure to hazards and would need to be quantified in any worthwhile 'index'. But, if such an index could be devised, it might help prevent the current, often sensational approach to hazards and risks that is adopted by most journalists. A suitable riskometer rating might help improve both media performance and the general public's perception of risks.

Summary

The general public's perceptions of risk are often highly inaccurate, but by underestimating common

risks while exaggerating exotic ones, we may end up protecting ourselves against unlikely perils while failing to take precautions against those that are far more dangerous. For example, people may be persuaded by sensational news stories that chemicals and pesticides considerably increase the risk of certain types of cancer. Perhaps they do, but the three main causes of cancer remain smoking, dietary imbalance, and chronic infections. Statisticians, psychologists, and others should put more effort into finding ways of presenting risks so that they may be more rationally appraised and compared. This is unlikely to be easy because people tend to form opinions rather quickly, usually in the absence of strong supporting evidence. Strong beliefs about risk, once formed, change very slowly and are extraordinarily persistent in the face of contrary evidence. Risk communication research must take up this challenge if the public is ever going to be persuaded to be rational about risk.

References

- [1] Paulos, J. (1994). *Innumeracy*, Penguin Books, London.
- [2] Slovic, P., Fischhoff, B. & Lichtenstein, S. (1980). Facts and fears: understanding perceived risk, in *Societal Risk Assessment: How Safe is Safe Enough?* R.C. Schwing & W.A. Albers, eds, Plenum Press, New York.

Further Reading

- Royal Society Study Group. (1983). *Risk Assessment*, The Royal Society, London.
- BMJ. (1990). *The BMA Guide to Living with Risk*, Penguin, London.
- Walsh, J. (1996). *True Odds*, Merritt Publishing, Santa Monica.

(See also **Odds and Odds Ratios; Probability: An Introduction**)

BRIAN S. EVERITT

Robust Statistics for Multivariate Methods

MARIA-PIA VICTORIA-FESER

Volume 4, pp. 1764–1768

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Robust Statistics for Multivariate Methods

Robust statistics, as a concept, probably dates back to the prehistory of statistics. It has, however, been formalized in the sixties by the pioneering work of Huber [9, 10] and Hampel [6, 7]. Robust statistics is an extension of classical statistics, which takes into account the fact that models assumed to have generated the data at hand are only approximate. It provides tools to investigate the robustness properties of a statistic T (such as estimators, test statistics) as well as robust estimators and robust testing procedures (see **Robust Testing Procedures**).

Although one would easily agree that models can only describe approximately the reality, what is more difficult to understand is the effect of this fact on the properties of classical statistics T for which it is assumed that models are exact. Suppose that the hypothetical (multivariate) model is denoted by F but that the data at hand have been generated by the general mixture $F_\varepsilon = [1 - \varepsilon]F + \varepsilon H$, with H a contamination distribution. Assuming F_ε means that the data have been generated by the model F with probability $[1 - \varepsilon]$ and by a contamination distribution H with probability ε . Note that a particular case for H is a distribution assigning a probability of one to an arbitrary point, that is, producing so-called outliers. If ε is large, the contamination distribution has an important weight in the mixture distribution and an analysis based on F_ε (assuming F as the central model) is meaningless. On the other hand, if ε is small, an analysis based on F_ε should not be entirely determined by the contamination. It is, therefore, important to find or construct statistics T that are not entirely determined by data contamination, that is, robust under slight model deviations (see **Finite Mixture Distributions**).

A well-known tool to assess the effect (on the bias of T) of infinitesimal amounts ε of contamination is the *influence function* (*IF*) introduced by Hampel [6, 7] and further developed in [8]. Another tool is the *breakdown point* (*BDP*), which measures the maximal amount ε of (any type of) contamination that T can withstand before it ‘breaks down’ or gives unreliable results (see for example [8]). A statistic T with bounded *IF* is said to be robust (in the infinitesimal sense). It should be stressed that most

classical procedures are not robust, and, in particular, all classical procedures for models based on the multivariate normal distribution (see **Catalogue of Probability Density Functions**) are not robust. This is the case, for example, for regression models (see **Multiple Linear Regression**), factor analysis, structural equation models, linear multilevel models (which include **repeated measures analysis of variance**).

In practice, to detect observations from a contamination distribution (i.e., contaminated data) is not an obvious task. For models based on the p -variate normal distribution $F_{\mu, \Sigma}$, a useful measure is the **Mahalanobis distance** d_i defined on each (multivariate) observation $\mathbf{x}_i$ by

$$d_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \quad (1)$$

The d_i takes into account the covariance structure of the data, which is very important in multivariate settings (see **Multivariate Analysis: Overview**). Indeed, as an example we consider scores on psychological tests collected for the study of age differences in working memory (see [1] for more details), which is presented as a multi **scatterplot** in Figure 1. A close look at the scatterplot between the variables ML1TOT and ML2TOT reveals that there is a minority of subjects not ‘fitting’ the covariance structure described by the bulk of data (i.e., the majority). On the other hand, on the univariate level, that is, when looking at the scores only on one of each variable, this minority of subjects has not so extreme scores. The point here is that, when dealing with multivariate models, the screening of the data at the univariate level is not sufficient to detect contaminated data (see **Multivariate Outliers**).

Unfortunately, scatter plots show only the behavior of the data at the bivariate level, and (exact) bivariate normality does not imply (exact) normality of higher orders. It is, therefore, important to be able to rely on general measures such as (1). However, (1) supposes the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ to be known, which, in practice, is never the case. If nonrobust estimators are used, then they are biased in the presence of data contamination, which means that the d_i will in the best case not reveal the right contaminated data (masking effect), and, in the worst, reveal false contaminated data.

Robust statistics for multivariate models have first been used for the estimation of multivariate mean (location) and covariance (scatter). In this setting, it is desirable for the robust estimators to

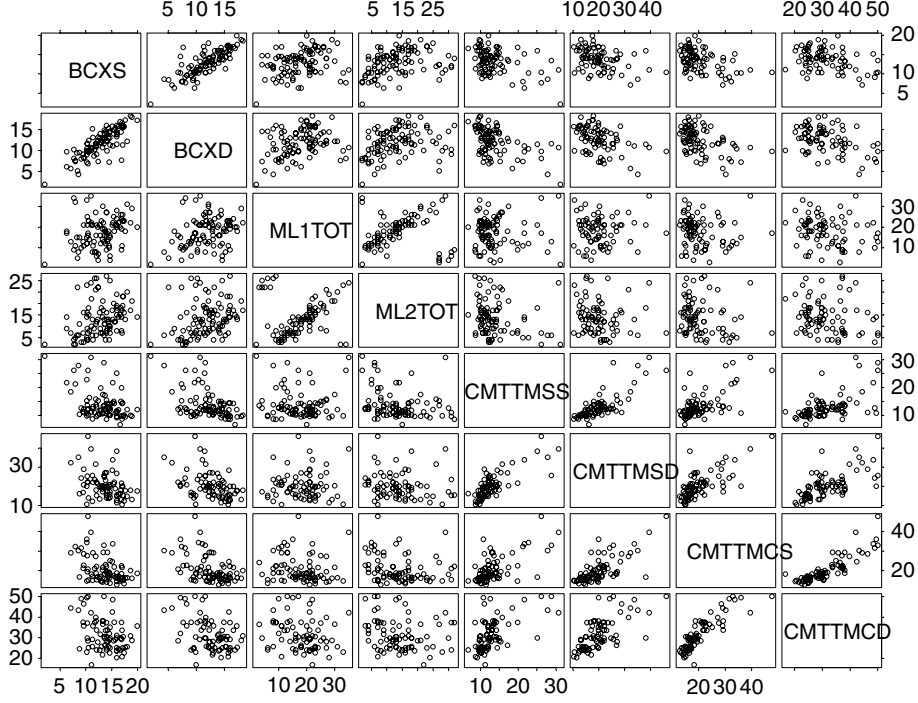


Figure 1 Multiscatter plot of the working memory study data

be affine equivariant (a linear transformation of the data results in a known transformation of the estimates), to have relatively high *BDP* (see [13]) and to be computationally efficient. The first high *BDP* affine equivariant estimator is the *minimum volume ellipsoid* (MVE) proposed by [17]. The ellipsoid (of dimension p) containing at least half of the data with minimum volume is found and the sample mean and covariance of these data define the MVE. The latter is very computationally intensive, and is known to have poor efficiency. However, it is used, for example, to detect contaminated data or as a starting point for more efficient estimators based on weighted means and covariances.

A general class of estimators in which one can find robust ones is the class of *M*-estimators (see [10]) that generalize **maximum likelihood estimators (MLE)**. *M*-estimators (see **M Estimators of Location**) are defined for general parametric models F_θ as the solution in θ of

$$\frac{1}{n} \sum_{i=1}^n \psi(\mathbf{x}_i, \theta) = 0 \quad (2)$$

When the ψ -function is the score function $s(\mathbf{x}, \theta) = (\partial/\partial\theta) \log f(\mathbf{x}, \theta)$, one gets the MLE. Such estimators, under very mild conditions, have known asymptotic properties that can be used for inference (see e.g., [8]). For the multivariate normal model, another popular class of estimators is the class of *S*-estimators (see [18]), which can be computed iteratively by means of

$$\frac{1}{n} \sum_{i=1}^n w_i^\mu (\boldsymbol{\mu} - \mathbf{x}_i) = 0 \quad (3)$$

$$\frac{1}{n} \sum_{i=1}^n [w_i^\delta \boldsymbol{\Sigma} - w_i^\eta (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] = 0 \quad (4)$$

where the weights $w_i^\mu, w_i^\eta, w_i^\delta$ are decreasing functions of the Mahalanobis distances d_i . Note that when the former are equal to 1 for all i , one gets the classical sample means and covariances. The choice for the weights define different estimators (see e.g., [16]). When there are missing data, [1] proposed an adaptation of (3) and (4) as an alternative to the EM algorithm (see [4]). For the working memory

data (which include missing data), the correlation between ML1TOT and ML2TOT was found to be 0.84 (robust estimation), whereas it is equal to 0.20 when using the EM algorithm. Other robust estimators for multivariate location and scatter (and their statistical properties) can be found in, for example, [3], [5], [11], [12], [14], [19], [20], [21], [22] and [23].

Although the multivariate normal distribution (*see Catalogue of Probability Density Functions*) is the central distribution for several models, the covariance matrix is not always present in a free form. Indeed, like in **structural equations models** or in mixed linear models (*see Linear Multilevel Models*), the true covariance matrix is structured. For example, it could be supposed that the variances are all equal, and the covariances are all equal (one-way ANOVA with repeated measures). In these cases, it is important to estimate the covariance matrix by taking into account its structure, and not just estimate it freely, and then ‘plug-in’ the estimate in the model to estimate the other parameters. [2] proposed a general class of S -estimators for constrained covariance matrices that can be used for example with mixed linear models.

When the models are not based on the multivariate normal distribution, robust statistics become more complex. The Mahalanobis distance does not play anymore a role, and another measure for detecting contaminated data needs to be specified. For M -estimators, [10] proposed a weighting scheme based on the score function itself, that is,

$$\psi(\mathbf{x}, \theta) = w_c(\mathbf{x}, \theta)s(\mathbf{x}, \theta) \quad (5)$$

with

$$w_c(\mathbf{x}, \theta) = \min \left\{ 1; \frac{c}{\|s(\mathbf{x}, \theta)\|} \right\} \quad (6)$$

where $\|\mathbf{x}\| = \left(\sum_{j=1}^p x_j^2 \right)^{1/2}$ denotes the Euclidian norm. Observations corresponding to large (absolute) value of the score function are hence downweighted. The score function in a sense replaces the Mahalanobis distance for multivariate normal models. The parameter c can be chosen for efficiency arguments. With nonsymmetric models, (5) leads to inconsistent estimators, and, therefore, a shift needs to be added to (5) to make the M -estimator consistent (see for example, [8] and [15]). This can make the robust estimator computationally nearly unfeasible. With nonnormal multivariate models, robust statistics, therefore, still need to be further developed.

References

- [1] Cheng, T.-C. & Victoria-Feser, M. (2002). High breakdown estimation of multivariate mean and covariance with missing observations, *British Journal of Mathematical and Statistical Psychology* **55**, 317–335.
- [2] Copt, S. & Victoria-Feser, M.-P. (2003). *High Breakdown Inference in the Mixed Linear Model*. Cahiers du département d’économétrie no 2003.6, University of Geneva.
- [3] Davies, P.L. (1987). Asymptotic behaviour of S -estimators of multivariate location parameters and dispersion matrices, *The Annals of Statistics* **15**, 1269–1292.
- [4] Dempster, A.P., Laird, M.N. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–22.
- [5] Donoho, D.L. (1982). Breakdown properties of multivariate location estimators, Ph.D. qualifying paper, Department of Statistics, Harvard University.
- [6] Hampel, F.R. (1968). Contribution to the theory of robust estimation, Ph.D. thesis, University of California, Berkeley.
- [7] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**, 383–393.
- [8] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley, New York.
- [9] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.
- [10] Huber, P.J. (1981). *Robust Statistics*, John Wiley, New York.
- [11] Kent, J.T. & Tyler, D.E. (1996). Constrained M -estimation for multivariate location and scatter, *The Annals of Statistics* **24**, 1346–1370.
- [12] Lopuhaä, H.P. (1991). τ -estimators for location and scatter, *Canadian Journal of Statistics* **19**, 307–321.
- [13] Maronna, R.A. (1976). Robust M -estimators of multivariate location and scatter, *The Annals of Statistics* **4**, 51–67.
- [14] Maronna, R.A. & Zamar, R.H. (2002). Robust estimates of location and dispersion for high-dimensional datasets, *Technometrics* **44**, 307–317.
- [15] Moustaki, I. & Victoria-Feser, M.-P. (2004). *Bounded-Bias Robust Inference for Generalized Linear Latent Variable Models*. Cahiers du département d’économétrie no 2004.02, University of Geneva.
- [16] Rocke, D.M. & Woodruff, D.L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association* **91**, 1047–1061.
- [17] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

4 Robust Statistics for Multivariate Methods

- [18] Rousseeuw, P.J. & Yohai, V.J. (1984). Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis*, J.W., Franke & R.D., Martin editors, Hardle W, eds, Springer-Verlag, New York, pp. 256–272.
- [19] Stahel, W.A.(1981). Breakdown of covariance estimators, Technical Report 31, Fachgruppe für Statistik, ETH, Zurich.
- [20] Tamura, R. & Boos, D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, *Journal of the American Statistical Association* **81**, 223–229.
- [21] Tyler, D.E. (1983). Robustness and efficiency properties of scatter matrices, *Biometrika* **70**, 411–420.
- [22] Tyler, D.E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics, *Annals of Statistics* **22**, 1024–1044.
- [23] Woodruff, D.L. & Rocke, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association* **89**, 888–896.

MARIA-PIA VICTORIA-FESER

Robust Testing Procedures

RAND R. WILCOX

Volume 4, pp. 1768–1769

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Robust Testing Procedures

The term robust testing procedure roughly refers to hypothesis testing methods that are relatively insensitive to violations of the assumptions upon which they are based. This means, in particular, that a robust testing procedure should achieve two fundamental goals. The first is that the actual type I error probability is reasonably close to the nominal level. So if some hypothesis is tested at the .05 level, for example, the actual probability of a type I error should be reasonably close to .05. The second is that small shifts or changes in a distribution should not have an undue influence on **power**, the probability of detecting situations where the null hypothesis is false. In particular, if a hypothesis testing procedure has high power under normality, power should remain reasonably high when data are sampled from a slightly nonnormal distribution.

The mathematical tools for studying and understanding robustness issues have advanced considerably during the last forty years. The mathematical foundations of these methods are summarized by Hampel, Ronchetti, Rousseeuw, and Stahel [1] and Huber [2]. A description of these methods, written at a more elementary level, is provided by Staudte and Sheather [3]. Briefly, these tools provide a way of studying and characterizing the effect that small changes in a distribution have on measures of location (such as the mean) and dispersion (such as the usual variance). This has led to an understanding of why conventional hypothesis testing methods, such as the two-sample Student's t Test, and ANOVA F test, (see **Catalogue of Parametric Tests**) are not robust, contrary to what was initially thought. Not only do they suffer from problems when trying to control the probability of a type I error but also arbitrarily small departures from normality can substantially lower power relative to other techniques that might be used (see **Robustness of Standard Tests**).

Also, these mathematical tools have formed the foundation for new inferential methods that deal with the problems now known to plague conventional techniques. For example, a common way of improving efficiency relative to the sample mean is to downweight **outliers**. But this leads to a technical problem: the usual method for estimating standard

errors no longer applies. If, for example, outliers are simply removed and standard methods for means are applied to the remaining data, the wrong standard error is being used, which can result in poor control over the probability of a type I error. But, owing to the new mathematical and statistical tools that have emerged, theoretically sound estimates are now available.

When comparing two or more groups of participants, there are many ways of improving control over the probability of a type I error versus conventional methods based on means. Each approach has advantages and disadvantages, which are discussed in Wilcox [4]. For a more detailed description written at a slightly more advanced level, see Wilcox [5]. Some of these methods deal directly with measures of location, roughly referring to a value intended to reflect what is typical. The mean and median are the best-known examples, but today other measures of location have been found to have practical value such as trimmed means and **M-estimators of location** (see **Trimmed Means**).

Conventional methods for means can suffer from low power for three general reasons: unequal variances, skewness, and sampling from distributions where **outliers** are relatively common. (The term outliers refers to values that are unusually small or large.) All three create serious concerns, but perhaps outliers are particularly troublesome. One reason is that modern outlier detection methods suggest that outliers are rather common. Another reason is that outliers can inflate the usual sample variance, which in turn can mean low power (see **Robustness of Standard Tests**). But even if no outliers are detected, skewness and unequal variances can result in relatively low power as well.

One general approach when trying to avoid low power, due to nonnormality, is to replace means with a measure of location that provides reasonably high power under normality, but unlike methods based on means, relatively high power is achieved when sampling from nonnormal distributions. A crucial feature of these alternative estimators is that they deal directly with outliers. Methods based on medians can offer improved control over the probability of a type I error, but their power is relatively unsatisfactory when the data are normal. But other measures of location, such as trimmed means and M-estimators, satisfy both criteria. No single method dominates, but there are several inferential techniques that appear

2 Robust Testing Procedures

to perform well for a broad range of situations (e.g., [5]).

As for regression and correlation, conventional hypothesis testing methods inherit all of the practical problems associated with conventional methods for comparing groups based on means, and new problems are introduced. Again, vastly improved methods have been derived [5]. For example, even under normality, it is known that heteroscedasticity is a serious practical problem when using the ordinary least squares estimator (*see **Least Squares Estimation***), but methods that deal with this problem are available. Also, both heteroscedasticity and nonnormality can result in relatively poor power when using ordinary least squares, but many modern estimators provide relatively high power under both normality and homoscedasticity as well as nonnormality and

heteroscedasticity. Like methods intended to improve on the sample mean, they are based on techniques that are relatively insensitive to outliers.

References

- [1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.
- [2] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [3] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [4] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.
- [5] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.

RAND R. WILCOX

Robustness of Standard Tests

RAND R. WILCOX

Volume 4, pp. 1769–1770

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Robustness of Standard Tests

Standard hypothesis-testing methods make two crucial assumptions: normality and homoscedasticity (equal variances). These methods include student's t , the **analysis of variance (ANOVA)** F test, and basic methods for testing hypotheses about least squares regression parameters and Pearson's correlation (*see Multiple Linear Regression; Catalogue of Parametric Tests*). If either or both of these assumptions are violated, it is questionable whether these methods provide an adequate control over the probability of type I errors, and whether they have high **power** relative to other methods that might be used. Early investigations into the robustness of the ANOVA F by Box [1] seemed to suggest that it is indeed relatively insensitive to violations of assumptions. But more recent results summarized by Hampel, Ronchetti, Rousseeuw and Stahel [3], Huber [4], Staudte and Sheather [5], and Wilcox [7–9] paint a decidedly different picture. These newer results might appear to contradict results reported by Box, but this is not the case. New theoretical results have helped improve our understanding of where to look for problems, the result being the realization that any method for comparing groups based on means can be expected to have serious practical difficulties for a broad range of situations.

Consider, for example, the usual one-way ANOVA F test for independent groups. Box [1] reported theoretical and empirical results on how unequal variances affect the probability of a type I error, assuming normality. Let R be the largest (population) standard deviation among the groups divided by the smallest standard deviation. Box limited his numerical results to situations where $R \leq \sqrt{3}$. No reason was given for not considering larger ratios, perhaps because at the time there were few, if any, studies looking into how much the standard deviations might differ in practice. But, various empirical studies summarized by Wilcox and Keselman [10] suggest that much larger ratios are common. When comparing two groups only, and when sampling from a normal distribution, reasonably good control over the probability of a type I error is achieved except possibly for very small sample sizes. However, for unequal sample sizes, or when sampling from a nonnormal distribution, this is no

longer the case as indicated, for example, in [7–10]. Even under normality with four or more groups, unequal variances can result in poor control over the probability of a type I error, and nonnormality exacerbates this problem.

A basic requirement of any method is that under random sampling, it should converge to the correct answer as the sample sizes get large. For example, when computing a 0.95 confidence interval, the actual probability coverage should approach 0.95. With unequal sample sizes, there are general conditions where student's t does not satisfy this criterion [2].

A rough rule is that, as an experimental design becomes more complicated, standard hypothesis testing methods become more sensitive to violations of assumptions. For example, under normality and with equal sample sizes, student's t is relatively robust in terms of type I errors when the variances are unequal, but this is no longer the case when using the ANOVA F with four or more groups.

There are at least three general reasons why conventional methods for means can have relatively low power: unequal variances, **skewness**, and sampling from distributions where **outliers** are relatively common. Outliers are values that are unusually large or small. Outliers are of particular concern because modern outlier detection methods suggest they are rather common and because they can inflate the usual sample variance, which in turn can mean low power.

As an illustration, consider the following data.

Group 1 : 6 9 19 9 8 12 14 11 14
Group 2 : 4 7 2 10 15 11 1 3.

The sample means are 11.3 and 6.6, respectively, and student's t rejects at the 0.05 level. Now suppose the largest value in the first group is increased to 34. Then, the mean for the first group increases from 11.3 to 13, so the difference between the two means has increased from 4.7 to 6.4, yet we no longer reject; the P value is 0.08. Increasing the largest value to 80, the mean for the first group increases to 18, yet the P value increases to 0.19. The reason is that the sample variance has increased as well. The value 80, for example, is an outlier among the data for the first group, and it inflates the sample variance to the point that we no longer reject, even though the difference between the sample means has increased as well. Even small departures from normality can cause

2 Robustness of Standard Tests

problems, a result that became evident with the publication of a seminal paper by Tukey [6]. Theoretical results were developed during the 1960s with the goal of dealing with this problem [3–5], and they form the basis of a wide range of modern inferential techniques [7–10]. Modern **robust testing procedures** not only deal with low power due to nonnormality, they substantially reduce problems associated with skewness and heteroscedasticity (see **Heteroscedasticity and Complex Variation**).

References

- [1] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way model, *Annals of Mathematical Statistics* **25**, 290–302.
- [2] Cressie, N.A.C. & Whitford, H.J. (1986). How to use the two sample *t*-test, *Biometrical Journal* **28**, 131–148.
- [3] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.
- [4] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [5] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [6] Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in *Contributions to Probability and Statistics*, I. Olkin, et al., eds, Stanford University Press, Stanford.
- [7] Wilcox, R.R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy*, Springer, New York.
- [8] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, San Diego, CA: Academic Press.
- [9] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.
- [10] Wilcox, R.R. & Keselman, H.J. (2003). Modern robust data analysis methods: measures of central tendency, *Psychological Methods* **8**, 254–274.

RAND R. WILCOX

Runs Test

CLIFFORD E. LUNNEBORG

Volume 4, pp. 1771–1771

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Runs Test

Two-sample Runs Test

The Wald–Wolfowitz [3] runs test dates from 1940, making it one of the earliest nonparametric tests. It provides a test of a common distribution for two independent random samples. However, the test has low power relative to such alternatives as the Kolmogorov–Smirnov or Cramér–von Mises two-sample tests and has declined in popularity, as attested to by [1] and [2].

Let $(x_1, x_2, \dots, x_n)$ and $(y_1, y_2, \dots, y_m)$ be independent random samples of sizes n and m from the two random variables, X and Y . The scale of measurement of the two random variables is at least ordinal and, to avoid problems with ties, ought to be strictly continuous (*see Scales of Measurement*).

The hypothesis to be nullified is that the two random variables have a common distribution. The alternative is that the two distributions differ.

Arrange the combined samples from smallest to largest and identify each observation with its source. As an example, Table 1 gives the ranked algebra achievement scores of two samples of students, one taught by the present method (P), and one by a proposed new method (N).

Table 1 Ranked algebra achievement scores

Score	60	64	67	68	69	70	71	72	73	79	80	84
Method	N	N	P	N	P	N	N	P	N	P	P	P

Count the number of runs of the two sources. Here, the number is eight, beginning with a run of two ‘N’s,

followed by a run of one ‘P’, and finishing with a run of three ‘P’s. If the distribution of achievement scores is the same under the two methods of instruction, the scores should be well mixed, leading to many short runs. If the distributions differ, the number of runs will be small. Is eight a small enough number of runs as to be unlikely under the null hypothesis?

The runs test is a permutation or randomization test (*see Permutation Based Inference; Randomization Based Tests*). The null reference distribution consists of the number of runs for all 924 possible permutations of the observations, six attributed to the present method, and six to the new method. This test is implemented, for example, in the XactStat and SC (www.mole-soft.demon.co.uk) packages. The exact Wald–Wolfowitz runs test in SC reports a probability of eight or fewer runs under the null hypothesis at 759/924, approximately .82. No evidence is provided by this test against the null hypothesis.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition., Wiley, New York.
- [2] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Edition, Chapman & Hall/CRC, London.
- [3] Wald, A. & Wolfowitz, J. (1940). On a test whether two samples are from the same population, *Annals of Mathematical Statistics* **11**, 147–162.

CLIFFORD E. LUNNEBORG

Sample Size and Power Calculation

KEVIN R. MURPHY

Volume 4, pp. 1773–1775

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sample Size and Power Calculation

If a random sample of N observations is drawn from a population, the precision with which sample statistics (e.g., sample means) estimate corresponding population parameters is determined largely by the number of observations. When population distributions are at least approximately normal, the precision of these estimates can be calculated from simple formulas in which N plays a prominent role.

For example, the formula for the standard error of the mean (SE_M) is

$$SE_M = \sqrt{\frac{\sigma^2}{N}}. \quad (1)$$

The standard error of the difference between two independent sample means (SE_{M1-M2}) is

$$SE_{M1-M2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}. \quad (2)$$

If you know, or can estimate, both the level of precision you wish to attain and the variability of scores in the population, it is easy to solve for the sample size needed to attain that level of precision. For example, suppose you would like to be 95% certain that the mean response on a survey that uses a five-point response scale is within 0.25 scale points of the population mean, and your best estimate of the population standard deviation is $\sigma = 0.92$. You can easily rearrange the formula for the standard error of the mean to find the N needed to attain this level of accuracy, using

$$N_{\text{needed}} = 1/[(\text{desired precision level}/1.96)^2/\sigma^2]. \quad (3)$$

If $\sigma = 0.92$, a sample with $N = 52$ (i.e., $1/[(0.25/1.96)^2/0.92^2] = 52$) will provide an estimate of the population mean that achieves the desired level of precision. Similarly, if you would like to be 95% certain that the difference between mean responses in two independent samples (with SDs of 0.90 and 0.85, respectively) is within 0.25 scale points of the population difference, you can rearrange the formula for the standard error of the difference between

sample means. In this example, you will need samples of $N = 94$ (i.e., $1/[(0.25/1.96)^2/(0.90^2 + 0.85^2)] = 94$) to attain the desired level of precision.

Hypothesis Testing – Comparisons of Means

The **power** of a statistical test of a null hypothesis is defined as the probability that a test statistic will lead you to reject this null hypothesis when it is in fact false. For example, if you use **analysis of variance** to compare the means in k samples drawn from populations with different means, power is the probability that you will correctly conclude that these means differ. Statistical power is determined by three key parameters: (a) the population difference in means, or the **effect size**, (b) the decision criteria used to define results as statistically significant, and (c) the number of observations [1–5]. Power is highest when there are in fact large differences in population means, when less stringent criteria are used to define statistical significance (e.g., $p < .05$ vs. $p < .01$), and/or when large samples are used.

One of the primary applications of power analysis is to determine the sample size needed to have a reasonable chance of rejecting the null hypothesis. Standards and conventions vary somewhat across fields, but power must usually be substantially above 0.50 to be judged adequate [5], and power levels of 0.80 or above are usually sought [1]. A number of approaches have been suggested for estimating power; models based on the noncentral F distribution [6, 7] can be applied to a wide range of data-analytic techniques [5]. For example, the test statistic used to evaluate differences in sample means in the analysis of variance ($F = MS_{\text{between}}/MS_{\text{within}}$) is distributed as a noncentral F , where the degree of noncentrality reflects the size of the difference in population means. The null hypothesis that population means are identical would produce a noncentrality parameter of zero. Thus, comparing the observed F with the values obtained from a table of the simple (central) F distribution provides a test of the plausibility of the null hypothesis. The greater the difference in population means (i.e., the larger the value of the noncentrality parameter), the more the population distribution of F values will shift upward, and the greater the likelihood that the obtained F will exceed the critical value of F used to define a

2 Sample Size and Power Calculation

Table 1 Sample size required to attain power = 0.80($\alpha = 0.05$)

Effect size: % of variance explained	k = number of means	N = sample size	n_j = number of subjects per cell
0.01	2	777	389
	3	956	318
	4	1077	270
	5	1266	253
0.10	2	74	37
	3	92	31
	4	104	26
	5	116	24
0.20	2	34	17
	3	44	15
	4	51	13
	5	57	11

statistically significant result. Effect sizes are often expressed in terms of statistics such as the standardized mean difference (d) or the percentage of variance accounted for by differences in group means (η^2 , or its equivalent, R^2).

Power increases as a nonlinear function of N . As the standard error formulas shown earlier suggest, power functions more closely track the square root of N than the absolute value of N . Table 1 lists the sample size needed to obtain a power of 0.80 for rejecting the null hypothesis as a function of the effect size and the number of means compared (k), given the decision criterion $\alpha = 0.05$. In this table, the effect size is indexed by the proportion of variance accounted for by differences between group means in the population [5]; both the total sample size needed and the number of subjects per cell needed are shown in Table 1 (Tables in Cohen [1] are presented in terms of n_j).

For example, if you are comparing the means of three groups, and you expect that differences between groups account for 10% of the variance in scores in the population, you will need at least $N = 92$ observations to attain power of 0.80. If you expect a smaller effect size, for example that group differences account for 1% of the variance in scores, a much larger sample will be needed to achieve power of 0.80 (here, $N = 956$).

Power analyses in analysis of variance designs with multiple factors (see **Factorial Designs**) or repeated measures (see **Repeated Measures Analysis of Variance**) follow the same general principle, that power is higher when the population effects are large or when the sample sizes are larger. However,

in a complex design, you might have substantially different levels of power for different questions that are asked in the study. In multifactor designs, you will generally have more power when asking questions about main effects (i.e., the simple effects of noise and illumination) than about interactions, but the level of power of each in an analysis of variance model is influenced by both the population effect size and the number of observations that go into calculating each of the sample means to be compared.

Power analyses can be extended to complex multivariate procedures (see **Multivariate Analysis: Overview**). Stevens [8] discusses applications of power analysis in **multivariate analysis of variance**. Cohen [1] discusses power analyses for a wide range of statistical techniques. Both sources include extensive tables for estimating power, and these can be used to determine the number of cases needed to reach power of 0.80 (or whatever other convention is used to define adequate power) for the great majority of statistical procedures used in the behavioral and social sciences.

References

- [1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.
- [2] Kraemer, H.C. & Thiemann, S. (1987). *How Many Subjects?* Sage Publications, Newbury Park.
- [3] Labovitz, S. (1968). Criteria for selecting a significance level: a note on the sacredness of 05, *American Sociologist* 3, 220–222.
- [4] Lipsey, M.W. (1990). *Design Sensitivity*, Sage Publications, Newbury park.

- [5] Murphy, K. & Myors, B. (2003). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, 2nd Edition, Erlbaum, Mahwah.
- [6] Patnaik, P.B. (1949). The non-central χ^2 - and F-distributions and their applications, *Biometrika* **36**, 202–232.
- [7] Pearson, E.S. & Hartley, H.O. (1951). Charts of a power function for analysis of variance tests, derived from the non-central F-distribution, *Biometrika* **38**, 112–130.
- [8] Stevens, J.P. (1980). Power of multivariate analysis of variance tests, *Psychological Bulletin* **86**, 728–737.

KEVIN R. MURPHY

Sampling Distributions

SABINE LANDAU

Volume 4, pp. 1775–1779

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sampling Distributions

A *statistic* is a summary measure that can be calculated from a sample (a subset of a population). A sampling distribution is the probability distribution of a statistic. The goal of *inferential statistics* is to use such summary measures to make inferences about parameters that describe the population from which the sample is drawn. Knowledge of the sampling distribution of particular statistics is essential to derive these inferences, and much effort in statistical science has been devoted to achieving this (for an introduction to statistical inference see, for example, [1] and entries on **Classical Statistical Inference: Practice versus Presentation**, **Deductive Reasoning and Statistical Inference**, and **Neyman–Pearson Inference**). To appreciate the concept of the sampling distribution, we need to first consider the frequentist’s model (see **Probability: Foundations of**) of how data come about.

Population and Sample

The collection of individuals about whom information is desired is usually referred to as the *population* or more specifically the *target population*. For example, we might be interested in the prevalence of depression in a clearly defined clinical population (patients who have been given a certain diagnosis, are within a specified age range, are members of a particular ethnic group, etc.). Since it is impractical and often impossible to measure the outcome of interest (here absence/presence of depression) for every member of the target population, a subset of the population (sample) is selected for study. Statistical inferences depend on the assumption of randomness where a *random sample* of size n is a subset of n of the members from the relevant population where the subset is chosen in such a way that every possible subset of size n has the same chance of being selected as any other.

A *parameter* is a characteristic that describes the target population in the same way a statistic describes a sample. For example, in our depression example, the outcome of interest is a binary variable with possible values ‘1’ (depressed) and ‘0’ (not depressed) and its distribution in the population is characterized completely by the proportion of

individuals who have depressive symptoms. (When relating to the population, a proportion is also often referred to as a *probability*.) Where a statistic is used to find out about or as a stand in for a population parameter, it is specifically called an *estimator statistic* or, for short, an *estimator*. If it forms the basis of a *statistical test*, it is known as a *test statistic*. Here, we shall look at statistics in the context of estimation. For example, an intuitive (and as it turns out a ‘good’, see later) estimator of the proportion of individuals with depression in the target population is the proportion of individuals with such symptoms in the sample. Note that in the frequentist’s approach to statistical inference, the population (true) parameter is assumed to be a fixed quantity. In contrast, the estimator statistic is a random quantity since it is a function of the data collected.

Sampling Variability

While sampling individuals saves time and money, the value calculated for the sample will almost certainly differ from the population parameter since not all individuals are sampled; the statistic is said to be affected by *sampling error*. For example, if the true prevalence of depression was 20%, and we find that in a sample of 50 patients, 12 had depressive symptoms, then our sample estimate of $12/50 = 24\%$ would be affected by a sampling error of 4%. Each sample might have a different sampling error introducing what is known as *sampling variability* of the statistic. The amount of sampling variability will reduce as the sample size increases and more of the population elements are investigated. To visualize this, we can employ a computer to repeatedly draw random samples of sizes $n = 50$ and $n = 100$ from a population with true prevalence 20% (i.e., from a *Bernoulli distribution* with success probability 0.2 (see **Catalogue of Probability Density Functions**) and calculate the proportion of individuals with depression (= successes) for each sample. Figure 1 illustrates the observed sample proportions by means of **histograms**. We see that most proportion estimates are near the true parameter value of 0.2 (say within a distance of 0.1) with the average deviation from the true parameter value being larger for the smaller ($n = 50$) samples.

2 Sampling Distributions

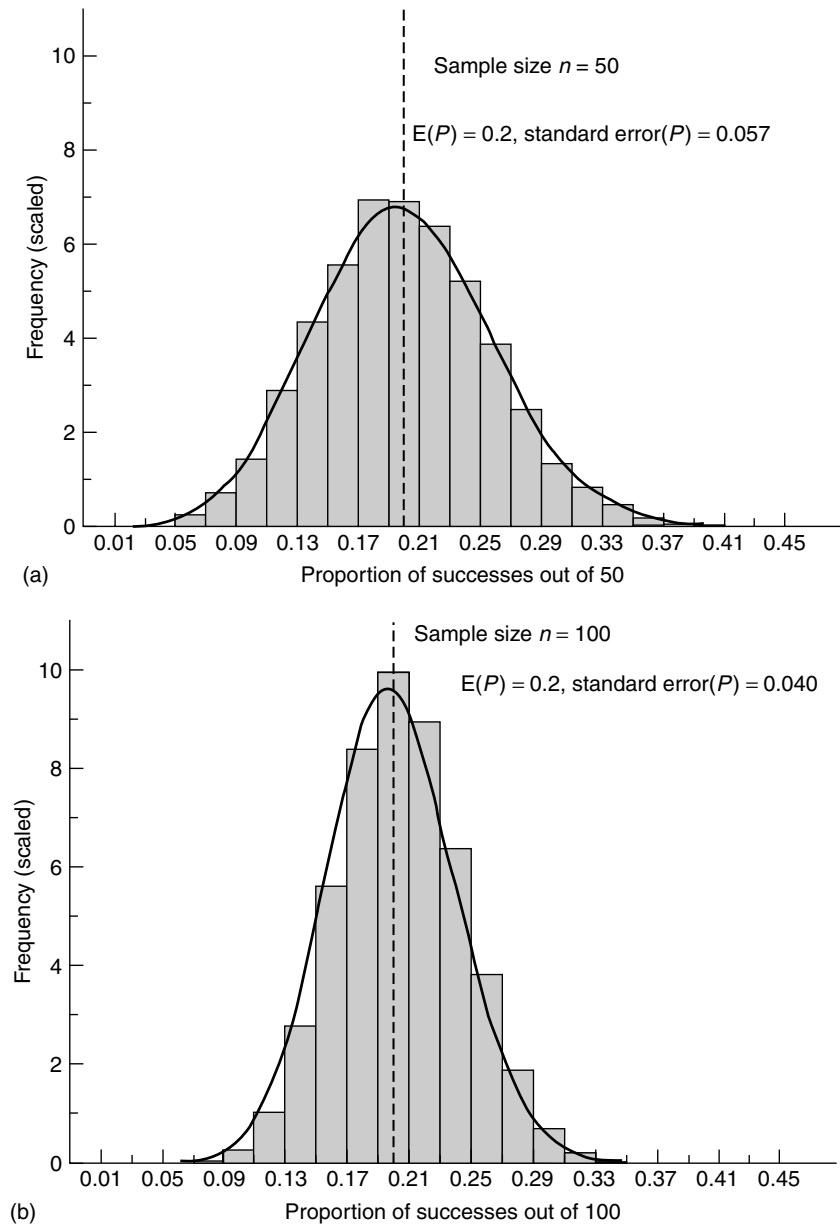


Figure 1 Simulated distribution of sample proportion of successes for two different sample sizes (10 000 simulation runs). The curves show normal density approximations to the histograms

Standard Error and Properties of Estimators

Figure 1 shows the sampling distribution of the proportion of successes based on simulations that mimic the data-generating process. As with any

probability distribution, two important characteristics are its mean and variance (or first and second moments (*see Expectation; Moments*)). The expected value of the sampling distribution is also called the *expected value* or *mean of the statistic*. The standard deviation of the sampling distribution

measures the average departure of the estimator from its long-term average and serves to quantify the *precision* of the estimator statistic. To distinguish it from the standard deviation of the population distribution, it is referred to as the *standard error of the statistic* or, for short, the standard error. Note that in contrast to the standard deviation of the population the standard error is affected by sample size. From our simulation, we calculate the expected value of the proportion statistic as $E(P) \approx 0.2$; that is, almost the true parameter 0.2. The value of the standard error of the proportion statistic P_n varies with sample size and we calculate that approximately $\text{s.e.}(P_{50}) \approx 0.057$ and $\text{s.e.}(P_{100}) \approx 0.040$; that is, if we draw a sample of size $n = 50$ and estimate the proportion, we will on average be 5.7% away from the expected value of the proportion statistic, while if we were to increase the sample size to $n = 100$, our average imprecision would reduce to 4%.

The sampling distribution provides a means by which to compare different estimators. Intuitively good estimators should hit the true population parameter on average. More formally, a desirable property of an estimator is that its expectation should equal the population parameter that it is aiming to estimate, irrespective of what that value is, that is, that it is *unbiased* for the parameter of interest. For example, our simulation suggests that the sample proportion is unbiased for the population proportion. (Unbiasedness can be shown to hold theoretically for any value of the population proportion.) When the estimator is unbiased, its standard error is the square root of the average squared deviation of the sample estimates from the population parameter; or in other words, an average sampling error. The sampling error generated by two alternative unbiased estimators can therefore be compared by their standard errors. Under some circumstances, unbiased estimators can be shown to attain the smallest variance that is possible within a particular probability model (see **Information Matrix; Estimation**).

The estimators discussed above are more specifically known as *point estimators*. Knowledge of the sampling distribution of relevant statistics also allows the construction of *interval estimators*, which are more commonly referred to as **confidence intervals**. The second strand of classical statistical inference – the testing of hypotheses about the parameters of

the population (or populations) at a given *significance level* – also requires knowledge of the sampling distribution of a test statistic under the relevant null hypothesis. Finally, since quality measures of inferential methods such as the standard error of an estimator, the width of a confidence interval, or the *power* of a test, all depend on the sample size, knowledge of the sampling distribution is essential to plan the appropriate size of a study (see **Power**).

How to Derive the Sampling Distribution?

Having made the case that the sampling distribution is an essential tool for statistical inference, the question might be asked ‘How can the sampling distribution of a statistic be derived from a *single* observed sample?’. After all, the considerations above assumed that we knew the population parameter of interest and could therefore mimic the repeated random sampling from the population. In practice, we do not know the true parameter and we only have one (random) sample. In general, the answer is that we have to specify the probability distribution in the population so that statistical theory or empirical resampling techniques can provide the sampling distribution of a statistic.

Parametric statistical methods assume a parameterized probability distribution in the population. Preferably, such a model assumption should be based on theory or derived empirically from the observed data. However, when theoretical results are not available and the data are sparse, a parameterized shape of the distribution is often simply assumed for convenience. Once a parametric distribution model has been specified, analytic results often facilitate expression of the sampling distribution as a function of the (unknown) population parameters. Alternatively, computing intensive methods such as the *parametric bootstrap* that resample from the estimated population distribution can be used to generate the sampling distribution (see **Bootstrap Inference**). In contrast, *nonparametric statistical methods* derive sampling distributions without parameterization of the population distribution. Resampling methods, in particular, the *nonparametric bootstrap*, which resamples from the *empirical* distribution of a sample, are useful in this respect (see **Bootstrap Inference**).

The theoretical derivations of sampling distributions can involve considerable amounts of probability theory, and here we just mention two well-known principles for derivation.

1. Let μ denote the population mean of a continuous variable of interest and σ the population standard deviation. If the population has a bell-shaped distribution (a *normal distribution*, see **Catalogue of Probability Density Functions**), then the sample mean $\bar{X} = (1/n) \sum_{i=1}^n X_i$, where X_i denotes the observation on the i th object in the sample of size n , also has a normal distribution with mean μ and standard error $\sigma/\sqrt{n}$. In other words, when normality can be assumed in the target population, then the sample mean is an unbiased estimator of the population mean and as the sample size n increases, its standard error decreases by the factor $1/\sqrt{n}$. The standard error of the sample mean is commonly estimated from the sample by replacing the population standard deviation in $\sigma/\sqrt{n}$ by a suitable estimator. For a normal population, the sample standard deviation $s = (1/(n-1))\sqrt{\sum_i (X_i - \bar{X})^2}$ is an unbiased estimator for σ .
2. When the sample size is sufficiently large, an important statistical theorem, the **central limit theorem**, states that the sample mean has mean μ and standard error $\sigma/\sqrt{n}$, irrespective of the distributional shape in the population. The larger the sample size, the better the approximation of the sampling distribution of the sample mean by the normal distribution. We can apply the central limit theorem to the depression example, where we were sampling from a population with a binary outcome. Since the proportion of '1s' in a set with binary elements is the same as the arithmetic mean of the '0' and '1' values, the unknown prevalence of depression is

the population mean. We can therefore estimate it by the unbiased sample mean (the sample proportion of '1s') and know that the standard error of this estimator is approximately $\sigma/\sqrt{n}$. The population standard deviation is a function of the population distribution and for binary outcomes is known to be $\sigma = \sqrt{p(1-p)}$, where p denotes the probability of observing a '1'. We can therefore estimate the standard error of the proportion statistic by $\sqrt{(P_n(1-P_n))/n}$. For example, if 12 out of a sample of 50 subjects showed depressive symptoms, we would estimate the prevalence of depression in the clinical target population as 24% and the standard error of the proportion estimator as $\sqrt{(0.24(1-0.24)/50)} = 0.0604$ or approximately 6%. Comparison of this to the true standard error of the sample proportion of 5.7% shows that the procedure has performed reasonably well.

To conclude, sampling distributions are needed to carry out statistical inferences. They describe the effects of sampling error on statistics as a function of sample size. Frequently encountered sampling distributions are the normal distribution, *Student's t distribution*, the *chi-square distribution* and the *F distribution* (see **Catalogue of Probability Density Functions**).

Reference

- [1] Howell, D.C. (2002). *Statistical Methods in Psychology*, 5th Edition, Duxbury Press, Belmont.

SABINE LANDAU

Sampling Issues in Categorical Data

SCOTT L. HERSHBERGER

Volume 4, pp. 1779–1782

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sampling Issues in Categorical Data

Introduction

A *categorical* variable is one for which the measurement scale consists of a set of categories [5]. Categorical variables may have categories that are naturally ordered (*ordinal* variables), or have no natural order (*nominal* variables). For example, the variable ‘health status’ with categories ‘excellent’, ‘good’, ‘satisfactory’, and ‘poor’ is an ordinal variable, as is age with categories ‘young’, ‘middle age’, and ‘old’. Alternatively, variables such as political party affiliation, with categories ‘Democratic’, ‘Republican’, ‘Libertarian’, and ‘Independent’, or sex with categories ‘male’ and ‘female’ are examples of nominal variables (*see Scales of Measurement*).

In most studies with categorical data, the *sampling units* (e.g., people) are classified simultaneously on the levels of the categorical variables. For instance, we might categorize people simultaneously by health status, age, party affiliation, and sex. One particular unit might then be described as a Democratic, young male in good health. The results of cross-classifying the sampling units are frequently arranged as counts in a **contingency table**. The simplest example of a contingency table is the 2×2 cross-classification of the sampling units into one of the four cells defined by the two levels of the two variables. When expressed in terms of observed frequencies, a 2×2 table might be represented as shown in Table 1.

When expressed in terms of probabilities, a 2×2 table might be represented as shown in Table 2.

Table 1 Observed frequencies

f_{11}	f_{12}		$f_{1.}$
f_{21}	f_{22}		$f_{2.}$
$f_{.1}$	$f_{.2}$		$f_{..}$

Table 2 Probabilities

p_{11}	p_{12}		$p_{1.}$
p_{21}	p_{22}		$p_{2.}$
$p_{.1}$	$p_{.2}$		$p_{..}$

Example of a 2×2 Contingency Table

In this example, the results are from an experiment concerned with the association between the true length of a line and the length as perceived by the subjects. Subjects were shown two lines, one line was longer than 12 in. and one line was shorter than 12 in. The subjects were to decide whether the line they were shown was actually longer or shorter than 12 in. The contingency table is shown in Table 3.

The null hypothesis is that a subject’s perception of a line’s length is not related to its true length. Stated more formally, if correct, this null hypothesis implies that the conditional probability of being in column 1, given that an observation belongs to a known row, is the same for both rows:

$$\frac{p_{11}}{p_{1.}} = \frac{p_{21}}{p_{2.}}. \quad (1)$$

This also implies that

$$\frac{p_{11}}{p_{1.}} = \frac{p_{12}}{p_{2.}}. \quad (2)$$

Taken together, these two equalities result in the **odds ratio** (α) (also known as the cross-products ratio) [6]:

$$\alpha = \frac{\left(\frac{p_{11}}{p_{12}}\right)}{\left(\frac{p_{21}}{p_{22}}\right)} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1. \quad (3)$$

When $\alpha > 1$, the two variables are positively associated; when $\alpha < 1$, the two variables are negatively associated. However, odds ratios are not symmetric around one: An odds ratio larger than one by a given amount indicates a smaller effect than an odds ratio smaller than one by the same amount. While the magnitude of an odds ratio is restricted to range between zero and one, it is literally unrestricted above one,

Table 3 Results of perception of line length experiment

Perceived length	True length		Total
	$\leq 12''$	$> 12''$	
$\leq 12''$	6	1	7
$> 12''$	1	6	7
Total	7	7	14

2 Sampling Issues in Categorical Data

allowing the ratio to potentially take on any value. If the natural logarithm ($\ln$) of the odds ratio is taken, the odds ratio is symmetric above and below one, with $\ln(1) = 0$.

Under the null hypothesis of independence ($\alpha = 1$), the expected frequency in cell i, j is

$$e_{ij} = \frac{f_{i.}f_{.j}}{f_{..}}. \quad (4)$$

Thus, the χ^2 goodness-of-statistic can be used as a test of independence:

$$\begin{aligned} \chi^2 &= \frac{\sum_i \sum_j (f_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{f_{..}(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1.}f_{2.}f_{.1}f_{.2}} \\ &= \frac{14\{(6)(6) - (1)(1)\}^2}{(7)(7)(7)(7)} \\ &= 7.143, \quad df = 1, \quad p = .01. \end{aligned} \quad (5)$$

The odds ratio also indicates a substantial relationship between perceived and true line length:

$$\begin{aligned} \alpha &= \frac{\left(\frac{6}{14}\right)\left(\frac{6}{14}\right)}{\left(\frac{1}{14}\right)\left(\frac{1}{14}\right)} = 35.999, \\ \ln(35.99) &= 3.583. \end{aligned} \quad (6)$$

Subjects were more than three-and-a-half times more likely to correctly judge the line length than not.

Effects of Sampling Method

Statistical inferences concerning a contingency table require knowledge of how the observations were sampled. The theoretical probability distribution that best models how the data were sampled should ideally be identified, although in the case of a contingency table of any dimension, the χ^2 goodness-of-fit test is valid under a wide range of sampling schemes [1, 3]. Nonetheless, it is still useful to explicitly define the method of sampling, if for no other reason than the influence of the sampling model on our interpretation of the data [4]. Therefore, we consider three different sampling methods that

have been used to elicit subjects' responses to the line lengths.

Sampling Method 1: The Hypergeometric Distribution

Consider the situation in which there were seven lines longer than 12 in. and seven lines shorter than 12 in. When asked to judge line length, subjects were informed that there would be seven of both types. This constrains $f_{1.}$, $f_{2.}$, $f_{.1}$, and $f_{.2}$ to each equal a specific number, in this case seven. When all of the marginal totals are fixed by design, the underlying distribution of responses is best described by the *hypergeometric* distribution (see **Catalogue of Probability Density Functions**).

The hypergeometric distribution is defined as follows for a 2×2 contingency table [6]. Given a sample space containing a finite number of elements, suppose that the elements are divided into $K = 4$ mutually exclusive and exhaustive cells, with f_{11} in cell 1, f_{12} in cell 2, f_{21} in cell 3, and f_{22} in cell 4. A sample of $f_{..}$ observations is drawn at random without replacement. Then the probability of a *specific configuration* of a contingency table is given by the hypergeometric probability

$$p(f_{11}, f_{12}, f_{21}, f_{22}; f_{..}) = \frac{f_{1.}!f_{2.}!f_{.1}!f_{.2}!}{f_{..}!f_{11}!f_{12}!f_{21}!f_{22}!}. \quad (7)$$

Using the line length response data, the probability of this contingency table's configuration is

$$p(6, 1, 1, 6; 14) = \frac{7!7!7!7!}{14!6!1!1!6!} = .01. \quad (8)$$

When the sampling model required by the experimental design follows a hypergeometric distribution, the marginal totals are fixed. This type of experimental design is frequently used to test the *homogeneity* of distributions; that is, the distribution of responses is the same across two levels of a variable [7]. In our example, a test of homogeneity would examine whether the probability of correctly identifying the line length category was the same for both line length categories.

Sampling Method 2: The Binomial Distribution

Consider the situation again in which there were seven lines longer than 12 in. and seven lines shorter

than 12 in. When asked to judge line length, subjects are *not* informed that there are seven of both types. Although subjects are aware that there will be a total of 14 lines to judge, and that there are two types of lines, no information is given as to the distribution of the lines in the two categories. Thus, no constraints are placed on the specific values of the $f_{1.}$, $f_{2.}$, $f_{.1}$, and $f_{.2}$ marginal totals. In this situation, where $f_{..}$ is fixed, and where each observation can result in one of two choices (i.e., the line is longer or shorter than 12 in.; the probability of selecting either is .5), the probability of the contingency table is given by the *binomial* probability

$$p(f_{+}; f_{..}, p) = \frac{f_{..}!}{f_{+}!(f_{..} - f_{+})!} (p)^{f_{+}} (1 - p)^{f_{..} - f_{+}}, \quad (9)$$

where f_{+} = the number of correct classifications [3]. Thus, for our data, the binomial probability of the contingency table is

$$p(12; 14, .5) = \frac{14!}{12!(2)!} (.5)^{12} (.5)^2 = .01. \quad (10)$$

Binomial distributions occur when the number of classes = 2 (the line is longer or shorter than 12 in.), each class having a known probability p of selection. Here, $p = 7/14 = .5$.

Under the same experimental conditions, where subjects are unaware of how many there are of each type of line, but know the types of lines is greater than two, each with a known probability of selection, the contingency table's configuration follows the *multinomial* probability distribution. Thus, the multinomial distribution is a generalization of the binomial distribution.

Sampling Method 3: The Negative Binomial Distribution

Now we consider the situation in which there are still only two types (classes) of lines, but the total number $f_{..}$ of lines is not fixed. Here, we are interested in the $f_{..}$ required to successfully judge

a certain number (f_{+}) of lines. Thus, $f_{..}$ is left free to vary but f_{+} is fixed. Let us assume that the line length experiment was stopped when we found that subjects had correctly judged 12 lines. Therefore, 14 judgments were required to produce 12 successful ones. Observations generated from this experimental design follow a *negative binomial* distribution [2, 3] (see **Catalogue of Probability Density Functions**):

$$p(f_{..}; f_{+}, p) = \frac{(f_{..} - 1)!}{(f_{+} - 1)!(f_{..} - f_{+})!} (p)^{f_{+}} (1 - p)^{f_{..} - f_{+}}. \quad (11)$$

The negative binomial probability of our contingency table's specific configuration is

$$p(14; 12, .5) = \frac{(13)!}{(11)!(2)!} (.5)^{12} (.5)^2 = .01. \quad (12)$$

As noted earlier, under all three sampling methods, the hypergeometric, the binomial, and the negative binomial, the probability of our specific contingency table is the same.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.
- [2] Kudô, A., & Tarumi, T. (1978). 2×2 tables emerging out of different chance mechanisms, *Communications in Statistics, Part A – Theory and Methods* 7, 977–986.
- [3] Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks.
- [4] Pearson, E.S. (1947). The choice of statistical tests illustrated on the interpretation of data classified in a 2×2 table. *Biometrika* 34, 139–167.
- [5] Powers, D.A. & Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*, Academic Press, San Diego.
- [6] Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer, New York.
- [7] Wickens, T.D. (1986). *Multiway Contingency Tables Analysis for the Social Sciences*, Erlbaum, Hillsdale.

(See also **Measures of Association**)

SCOTT L. HERSHBERGER

Saturated Model

SCOTT L. HERSHBERGER

Volume 4, pp. 1782–1784

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Saturated Model

A *saturated model* contains as many parameters as there are data points, providing a perfect fit to the data [1]. Consider as an example a **log-linear model** fit to a **contingency table**. A loglinear model that specifies all possible main effects and interactions is saturated because the number of parameters equals the number of cells of the table. Saturated models have no residual variance – the *deviance* is zero – and are most useful for comparing the fit of *hierarchically nested models* (see **Hierarchical Models**).

A model's *identification status* is an issue especially relevant to the concept of a saturated model. Models can be (a) *under-identified*, (b) *just-identified*, or (c) *over-identified* [4]. Under-identification occurs when not enough relevant data is available to obtain unique parameter estimates. Note that when the degrees of freedom of a model is negative, at least one its parameters is under-identified. Just-identified models are always identified in a trivial way: Just-identification occurs when the number of data elements equals the number of parameter to be estimated. This is the saturated model. If the model is just-identified, a solution can always be found for the parameter estimates that will result in perfect fit – a discrepancy function equal to zero. Over-identification occurs when the number of data points available is greater than that which is needed to obtain a unique solution for all of the parameters. In fact, with an over-identified model, the degrees of freedom are always positive so that model fit can be explicitly tested. An over-identified model also implies that, for at least one of the model parameters, there is more than one model equation that the solution to the parameter must satisfy. The number of additional equations the solution must satisfy is generally referred to as the number of *over-identifying constraints*.

Most of the familiar statistical models within the family of the **Generalized Linear Model** are saturated or just-identified owing to restrictions that are placed on the parameters of the models [3]. Without these restrictions, the models would be under-identified. For example, let us consider the standard **analysis of variance (ANOVA)** model. For $i = 1, \dots, n$ and $j = 1, \dots, m$, with m fixed, let

$$y_{ij} \sim N(\mu_j, \sigma^2),$$

$$\mu_j = \alpha + \theta_j, \quad (1)$$

where y_{ij} is person i 's outcome in treatment j , μ_j is the population mean, θ_j is the difference between the mean of group j and the population mean, and α is the model's intercept. The parameters of interest are $\alpha, \theta_1, \dots, \theta_m$; however, the model is under-identified. One restriction among several that can be introduced to just-identify the analysis of variance (ANOVA) model is to impose the restriction is that $\sum_{j=1}^m \theta_j = 0$. Because $\theta_j = \alpha - \mu_j$, it follows that $\alpha = (1/m) \sum_{j=1}^m \mu_j$. Therefore, α represents a constant effect for the population, which is an average of all the cell means, whereas $\theta_j = \mu_j - (1/m) \sum_{j=1}^m \mu_j$ represents the deviation of the cell mean μ_j from the average of all of the cell means. It is, therefore, the main effect due to the j th level of the factor. Importantly, the statistical meaning of the parameters α and θ depends on the identification restriction.

A fully specified log-linear model is another example of a saturated, just-identified model. For instance, suppose that we wish to investigate the relationships between two categorical variables, X and Y , where X has I categories and Y has J categories. Then the saturated ('full') loglinear model is

$$\log(m_{ij}) = \lambda + \lambda_i X + \lambda_j Y + \lambda_{ij} XY, \quad (2)$$

for each combination of the $I \times J$ levels of the m cells, $i = 1, 2, \dots, I$, and $j = 1, 2, \dots, J$. $\log(m_{ij})$ is the log of the expected cell frequency of the cases for cell ij in the contingency table; μ is the overall mean of the natural log of the expected frequencies; $\lambda_i X$ is the main effect for variable X , $\lambda_j Y$ is the main effect for variable Y , and $\lambda_{ij} XY$ is the interaction effect for variables X and Y .

In order for the saturated loglinear model be just-identified, constraints must be imposed on the parameters. Several alternative constraint specifications will accomplish this [5]. For example, as in the analysis of variance (ANOVA) model, we may require that the sum of the parameters over all categories of each variable be zero. For a 2×2 table in which two variables, X (two categories) and Y (two categories):

$$\lambda_1 X + \lambda_2 X = 0,$$

$$\lambda_1 Y + \lambda_2 Y = 0,$$

$$\lambda_{11} XY + \lambda_{12} XY = 0,$$

2 Saturated Model

$$\begin{aligned}\lambda_{21}XY + \lambda_{22}XY &= 0, \\ \lambda_{11}XY + \lambda_{21}XY &= 0,\end{aligned}\quad (3)$$

which implies that

$$\begin{aligned}\lambda_1X &= -\lambda_2X, \\ \lambda_1Y &= -\lambda_2Y,\end{aligned}\quad (4)$$

and

$$\lambda_{11}XY = -\lambda_{21}XY = -\lambda_{12}XY = \lambda_{22}XY. \quad (5)$$

These restrictions result in the estimation of four parameters: One parameter estimated for λ , one parameter estimated for X (corresponding to the first category), one parameter for Y (corresponding to the first category of Y), and one parameter for the interaction of X and Y (corresponding to the first categories of X and Y). The values of the remaining parameters are derived from the four estimated parameters. The model is saturated – just-identified – because the model, which has four cells, has four estimated parameters. The expected cell frequencies will exactly match the observed frequencies. In order to find a more parsimonious model that is explicitly testable, an unsaturated model must be specified that introduces one or more over-identifying constraints on the parameter estimates. Such a model has degrees of freedom greater than zero, and can be achieved by setting some of the effect parameters to zero.

For example, if both categorical variables are mutually independent, then the following *independence model* describes the relationship between X and Y :

$$\log(m_{ij}) = \lambda + \lambda_iX + \lambda_jY. \quad (6)$$

Further, we may decide that Y is not a significant predictor of the cell frequencies:

$$\log(m_{ij}) = \lambda + \lambda_iX, \quad (7)$$

or alternatively, that X is not:

$$\log(m_{ij}) = \lambda + \lambda_jY. \quad (8)$$

Conceivably, neither X nor Y may be useful, resulting in the most restricted baseline model:

$$\log(m_{ij}) = \lambda. \quad (9)$$

The examples of unsaturated loglinear models given above are hierarchically nested models. Hierarchical models include all lower terms composed from variables in the highest terms in the model. Therefore, the model

$$\log(m_{ij}) = \lambda + \lambda_iX + \lambda_{ij}XY \quad (10)$$

would not be considered a nested model – for the $\lambda_{ij}XY$ term to be present in the model, both of its constituent variables must be as well. The choice of a preferred model is typically based on the formal comparison of goodness-of-fit statistics associated with hierarchically nested models: the likelihood ratios and/or deviances of nested models are compared to determine whether retaining the more parsimonious model of the two results in a significant decrement in the fit of the model to the data [2]. A significant decrement in fit implies that the expected frequencies generated by the more parsimonious model are significantly less close to the observed frequencies.

References

- [1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.
- [2] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, 2nd Edition, Chapman & Hall, Boca Raton.
- [3] Green, S.B., Marquis, J.G., Hershberger, S.L., Thompson, M.S. & McCollam, K.M. (1999). The overparameterized analysis of variance model, *Psychological Methods* 4, 214–233.
- [4] Hershberger, S.L. (2005). The problem of equivalent structural models, in *A Second Course in Structural Equation Modeling*, G.R. Hancock & R.O. Mueller, eds, Information Age Publishing, Greenwich.
- [5] Wickens, T.D. (1986). *Multiway Contingency Tables Analysis for The Social Sciences*, Lawrence Erlbaum, Hillsdale.

SCOTT L. HERSHBERGER

Savage, Leonard Jimmie

SANDY LOVIE

Volume 4, pp. 1784–1785

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Savage, Leonard Jimmie

Born: November 20, 1917, in Michigan, USA.

Died: November 1, 1971, in Connecticut, USA.

Although L.J. Savage is not a name that many psychologists would instantly recognize, his achievement in producing a coherent system for Bayesian thinking and inference lay behind much of the 1960s movement in psychology to replace classical **Neyman Pearson inference** with its Bayesian equivalent (see **Bayesian Statistics**) (see [2], for both an instance of the approach and one of its early key documents, and the classic textbook by [4]). This also led to a lively experimental program in behavioral decision making founded in part on Bayesian inferential premises, one which was even taken up by **Tversky** and **Kahneman**, the inventors of *Biases and Heuristics*, with their long-running investigation into so-called base rate problems using the Green and Blue Taxi Cab story and variants as illustrative material (see [3] for an overview of the early material). Of course, the motivation for much of the experimental work in psychology was to see how far human decisions approximated the normative ones prescribed by Bayes theory, but there was also a commitment to Bayes theory itself as a better way of making inferences in the face of uncertainty, the avowed aim of classical statistical inference, which leads us back to Savage.

He was educated initially at the University of Michigan's Ann Arbor campus and received a B.S. in mathematics in 1938, with a Ph.D. in 1941 on an aspect of pure mathematics, specifically differential geometry. During his year at Princeton's Institute of Advanced Study (1941–1942) he came to the attention of John von Neumann, whose own work on the theory of games became a later inspiration for Savage. Neumann encouraged Savage to turn to statistics, and he joined the Statistical Research Group at Columbia University in 1943. It was during his tenure at the University of Chicago that he wrote his most influential book *The Foundation of Statistics* [5]. Somewhat in the same frame of mind as Freud who, in his *Project for a Scientific Psychology* of 1895, attempted to characterize his emerging ideas about psychoanalysis in terms of current medical

and physiological knowledge, Savage first of all laid out his ideas on personal probability and utility and then attempted to reinterpret classical statistical inference using these new concepts. Again, like Freud, he admitted defeat over the project, but only in the preface to the second edition! The book also shows the significance of **Bruno de Finetti's** work on subjective probability and his key notion of *exchangeability* to the power and novelty of the approach. From a realization of the drawbacks of all nonpersonalistic orientations to probability and inference, Savage gradually developed an alternative, Bayesian form of inference, where the personally assessed *prior* probabilities of events are transformed into probabilities *a posteriori* in the light of new information, also personally assessed. The implications of this new form of inference, and the statistical novelties that flow from it, have not yet been fully realized or worked through and are likely to absorb the energies of statisticians for some time to come.

For Savage himself, the *Foundations* represented something of a high point, although he was to publish an important book with Dubins, which recast gambles as a form of stochastic or probabilistic process [1]. He also moved to the University of Michigan in 1960, and then finally to Yale where he stayed until his comparatively early death at the age of 53. Savage is important not only for his own work but also because he introduced American (and British) statistics to de Finetti and hence to the possibility of an alternative and powerful form of statistical modeling and inference.

References

- [1] Dubins, L.E. & Savage, L.J. (1965). *How to Gamble if you Must: Inequalities for Stochastic Processes*, McGraw-Hill, New York.
- [2] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research, *Psychological Review* **70**(3), 193–242.
- [3] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
- [4] Phillips, L.D. (1973). *Bayesian Statistics for Social Scientists*, Nelson, London.
- [5] Savage, L.J. (1954, 1972). *The Foundation of Statistics*, Wiley, New York.

SANDY LOVIE

Scales of Measurement

KARL L. WUENSCH

Volume 4, pp. 1785–1787

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scales of Measurement

The behavioral scientist who speaks of ‘scales of measurement’ almost certainly is thinking of the hierarchy of measurement scales proposed by psychophysicist **S. S. Stevens** [5.] A scale was said to be created by specifying the rules by which numbers are assigned to objects or events in such a way that there will be a one-to-one correspondence between some of the properties of the measured things and some of the properties of the measurements. Stevens defined four different types of scales, nominal, ordinal, interval, and ratio, based on the extent to which empirical relationships among the measured objects or events correspond to numerical relationships among the measurements (*see Measurement: Overview*).

If the measurement rules are such that the measurements can be used to establish the equivalence of objects with respect to the measured characteristic, then the scale is said to be nominal. As an example of a nominal scale, consider the assignment of the number 1 to all pets that are cats, 2 for dogs, and 3 for other types of pets.

If, in addition to the ability to establish equivalence, the rules allow one to establish order of the measured characteristic, then the scale is considered to be an ordinal scale. In Table 1, imagine that each ‘o’ in the first row represents a bead and that each of the beads is identical in mass. The numbers in the rows below represent measurements of the mass of the strings of beads. For Scale O, the order of the assigned numbers is identical to the order of the objects’ masses, so the scale is ordinal.

Notice that Scale O does not allow one to establish equivalence of differences. Objects A and B differ in mass by the same amount as objects B and C, but the difference in the numerical measurements for objects A and B is not the same as that for B and C. With Scale I, one can determine equivalence of differences. The difference in mass between objects

A and B is equal to the difference in mass between objects B and C, just as the difference between the measurements of objects A and B is equal to the difference in measurements between objects B and C. A scale that allows one to establish equivalence, order, and equivalence of differences is called an *interval scale*.

Scale I allows one to establish equivalence of differences but not of ratios. Object D has twice the mass of object C, and object E has twice the mass of the object D, but the ratio of the measurements for objects D/C does not equal that for objects E/D. Equivalence of ratios can, however, be determined with Scale R. A scale that allows one to establish equivalence, order, equivalence of differences, and equivalence of ratios is called a *ratio scale*.

Stevens opined that the four scales are best characterized by the types of transformations that can be applied to them without distorting the structure of the scale. With a nominal scale, one can substitute any new set of numbers (or other symbols) for the old set without destroying the ability to establish equivalence. For example, instead of using ‘1’, ‘2’, and ‘3’ to represent cats, dogs, and other types of pets, we could use ‘A’, ‘B’, and ‘C’. With an ordinal scale, one can apply any order-preserving transformation (such as ranking) and still have an ordinal scale. With an interval scale, only a positive linear transformation will produce another interval scale. With a ratio scale, only multiplying the measurements by a positive constant will produce another ratio scale.

Stevens’s scales of measurement can be thought of in terms of the nature of the relationship between the observed measurements and the true scores (the true amounts of the measured characteristic, that is, scores on the underlying construct or **latent variable**). Winkler and Hays [8] presented the following list of criteria used to determine scale of measurement:

1. Two things will receive different measurements only if they are truly nonequivalent on the measured characteristic.
2. Object A will receive a larger score than does object B only if object A truly has more of the measured characteristic than does object B. This will be the case when the measurements are related to the true scores by a positive monotonic function.
3. Where T_i represents the true amount of the measured characteristic for object i , and M_i

Table 1 Illustration of Stevens’ Measurement Rules

Object	A	B	C	D	E
Beads		o	oo	oooo	oooooooo
Scale O	–1.2	0	2	3	6
Scale I	–2	0	2	6	14
Scale R	0	2	4	8	16

2 Scales of Measurement

represents the score obtained when measuring object i , $M_i = a + bT_i$, $b > 0$. That is, the measurements are a positive linear function of the true scores.

4. $M_i = a + bT_i$, $b > 0$, $a = 0$. That is, the measurements are a positive linear function of the true scores, and the intercept is zero. The zero intercept is often called a ‘*true zero point*’ – an object that receives a score of zero has absolutely none of the measured attribute.

To be ratio, a scale must satisfy all four of the criteria listed above; to be interval, only the first three; to be ordinal, only the first two; to be nominal, only the first.

In addition to defining four scales of measurement, Stevens argued that the type of statistics that are ‘permissible’ on a set of scores is determined, in part, by the scale of measurement. Consider Stevens’s recommendations regarding measures of location. The mode is permissible for any scale, even the nominal scale. If there are 20 cats, 15 dogs, and 12 pets of other types in the shop, cats are the modal pet whether we represent cats, dogs, and others with the numerals 1, 2, and 3, or 1, 4, and 9, or any other three numerals. The median is permissible only for scales that are at least ordinal. It does not matter whether we measure the mass of the pets in the shop in grams, kilograms, or simple ranks – the computed median will represent the same amount of mass with any positive monotonic transformation of the true masses. The **mean** is permissible only for scales that are at least interval. Imagine that we have five pets, named A, B, C, D, and E. Their true masses are 1, 2, 3, 6, and 18, respectively, for a mean of 6. Pet D has a mass exactly equal to the mean mass. Suppose our interval data for these pets is defined by $M = 10 + 2T$. The observed scores are 12, 14, 16, 22, and 46, respectively, for a mean of 22. Again, pet D has a mass exactly equal to the mean mass. Now suppose we have ordinal data, such as simple ranks. The observed scores are 1, 2, 3, 4, and 5, respectively, for a mean of 3. Now it is pet C that appears to have a mass exactly equal to the mean mass, but that is not true.

Stevens’s belief that the scale of measurement should be considered when choosing which statistical analysis to employ (the measurement view) was embraced by some and rejected by others. Some of the former authored statistics texts that taught social

scientists to consider scale of measurement of great importance when selecting an appropriate statistical analysis [2]. Most controversial was the suggestion that parametric statistics require at least interval level data but that nonparametric statistics were permissible with ordinal data. Many statisticians attacked the measurement view [2, 7], while others defended it [3, 6]. Those opposed to the measurement view argued that the only assumptions necessary when using parametric statistics are mathematical, such as normality and homogeneity of variance. Those favoring the measurement view argued that behavioral researchers are interested in drawing conclusions about underlying constructs, not just observed variables, and accordingly they must consider the scale of measurement, that is, the nature of the relationship between true scores and observed scores.

Imagine that we are interested in testing the hypothesis that the mean aggressiveness of cats is identical to the mean aggressiveness of dogs and that this hypothesis is absolutely true with respect to the latent variable. If the relationship between our measurements and the true scores is not linear, the population means on our measurement variable may well not be equivalent. Accordingly, testing hypotheses about the means of latent variables seems more risky with noninterval data than with interval data. The real fly in the ointment here is that one never really knows with certainty the nature of the relationship between the latent variable and the measured variable – for my example, what is the nature of the function relating common measures of animal aggressiveness with the ‘true’ amounts of aggressiveness? How can one ever know with confidence if such measurements represent interval data or not? In some circumstances, one need not worry about whether or not the data are interval. If one assumes normality and homogeneity of variance, then differences in the means of an observed variable do indicate that the means on the latent variable also differ, regardless of whether the measurements are interval or merely ordinal [1].

One’s attitude to the relationship between scale of measurement and the choice of an appropriate statistical analysis may be determined by one’s more basic ideas about the nature of measurement [4]. Someone who believes that useful measurements are those that capture interesting empirical relationships among the measured objects or events (representational theory) will argue that scale of measurement is

an important characteristic to consider when choosing a statistical analysis, at least when conclusions are to be scale-free. The operationist, by contrast, believes that measurements are always scale-specific, and thus choice of statistical analysis is unrelated to scale of measurement.

Ultimately, one's decision about whether a set of data represents interval or merely ordinal measurement is largely a matter of faith. When one counts number of bites, aggressive postures, and submissive postures of fighting mice and combines them into a composite measure of aggressiveness, what is the nature of the relationship between these measurements and the 'true' amounts of aggressiveness displayed by these animals in 'concrete reality'? Is the relationship linear or not? How could one ever answer such a metaphysical question with certainty? One way to evade this dilemma is to treat reality as being constructed or invented rather than discovered and then argue that the results of the parametric statistical analysis apply only to that 'abstract reality' which is a linear function of our measurements. Such a defining of 'reality' in terms of the observed variables is not much different from a similar device often employed by applied statisticians, defining a population from a sample – the population for which these statistical inferences are made is that population for which this sample could be considered to be a random sample.

References

- [1] Davison, M.L. & Sharma, A.R. (1988). Parametric statistics and levels of measurement, *Psychological Bulletin* **104**, 137–144.
- [2] Gaito, J. (1980). Measurement scales and statistics: resurgence of an old misconception, *Psychological Bulletin* **87**, 564–567.
- [3] Maxwell, S.E. & Delaney, H.D. (1985). Measurement and statistics: an examination of construct validity, *Psychological Bulletin* **97**, 85–93.
- [4] Michell, J. (1986). Measurement scales and statistics: a clash of paradigms, *Psychological Bulletin* **100**, 398–407.
- [5] Stevens, S.S. (1951). Mathematics, measurement, and psychophysics, in *Handbook of Experimental Psychology*, S.S. Stevens, ed., Wiley, New York, pp. 1–49.
- [6] Townsend, J.T. & Ashby, F.G. (1984). Measurement scales and statistics: the misconception misconceived, *Psychological Bulletin* **96**, 394–401.
- [7] Velleman, P.F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading, *The American Statistician* **47**, 65–72.
- [8] Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, Inference, and Decision*, 2nd Edition, Holt Rinehart & Winston, New York.

KARL L. WUENSCH

Scaling Asymmetric Matrices

YOSHIO TAKANE

Volume 4, pp. 1787–1790

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scaling Asymmetric Matrices

Tables with the same number of rows and columns are called square tables. In square tables, corresponding rows and columns often represent the same entities (objects, stimuli, variables, and so on). For example, the i th row of the table represents stimulus i , and the i th column also represents the same stimulus i . Let x_{ij} denote the element in the i th row and the j th column of the table. (We often call it ij th element of the table.) We use X (in matrix form) to denote the entire table collectively. Element x_{ij} indicates (the strength of) some kind of relationship between the row entity (stimulus i) and the column entity (stimulus j). Square tables, in which $x_{ij} \neq x_{ji}$ for some combinations of i and j , are called asymmetric tables. In matrix notation, this is written as $X' \neq X$, where X' indicates the transpose of X .

Asymmetric tables arise in a number of different guises. In some cases, the kind of relationship represented in the table is antisymmetric. For example, suppose you have a set of stimuli, and you ask a group of subjects whether they prefer stimulus i or j for each pair of stimuli. Since j cannot be preferred to i if i is preferred to j , the preference choice constitutes an antisymmetric relationship. Let x_{ij} denote the number of times i is preferred to j . Tables representing antisymmetric relationships are usually asymmetric. These types of tables are often skew-symmetric, or can easily be turned into one by a simple transformation (e.g., $y_{ij} = \log(x_{ij}/x_{ji})$). In the skew symmetric table, $y_{ji} = -y_{ij}$ ($Y' = -Y$). Skew symmetric data, such as the one just described, are often represented by the difference between the preference values of the two stimuli involved. Let u_i represent the preference value of stimulus i . Then, $y_{ij} = u_i - u_j$. Case V of Thurstone's law of comparative judgment [8], and Bradley-Terry-Luce (BTL) model [1, 7] are examples of this class of models. The scaling problem here, is to find estimates of u_i 's, given a set of observed values of y_{ij} 's.

Here is an example: The top panel of Table 1 gives observed choice probabilities among four music composers, labeled as B, H, M and S. Numbers in the table indicate the proportions of times row composers are preferred to column composers. Let us apply the BTL model to this data set. The second panel of

Table 1 The BTL model applied to preference choice data involving four music composers

	Observed choice probabilities			
	B	H	M	S
B	.500	.895	.726	.895
H	.105	.500	.147	.453
M	.274	.853	.500	.811
S	.105	.547	.189	.500
Matrix of $y_{ij} = \log(x_{ij}/x_{ji})$				
B	0	2.143	.974	2.143
H	-2.143	0	-1.758	-.189
M	-.974	1.758	0	1.457
S	-2.143	.189	-1.457	0
Estimated preference values				
	1.315	-1.022	.560	-0.853

the table gives the skew symmetric table obtained by applying the transformation, $y_{ij} = \log(x_{ij}/x_{ji})$, to the observed choice probabilities. Least squares estimates of preference values for the four composers are obtained by row means of this skew symmetric table. B is the most preferred, M the second, then S, and H the least. Something similar can also be done with Thurstone's Case V model. The only difference it makes is that normal quantile (deviation) scores are obtained, when the matrix of the observed choice probabilities is converted into a skew symmetric matrix. The rest of the procedure remains essentially the same as in the BTL model.

Asymmetric tables can also arise from proximity relationships, which are often symmetric. In some cases, they exhibit asymmetry, however. For example, you may ask a group of subjects to identify the stimulus presented out of n possible stimuli, and count the number of times stimulus i is 'confused' with stimulus j . This is called stimulus recognition (or identification) data, and it is usually asymmetric. There are a number of other examples of asymmetric proximity data such as mobility tables, journal citation data, brand loyalty data, discrete panel data on two occasions, and so on. In this case, the challenge is in explaining the asymmetry in the tables.

A variety of models have been proposed for asymmetric proximity data. Perhaps the simplest model is the quasi-symmetry model (see **Quasi-symmetry in Contingency Tables**). The quasi-symmetry is characterized by $x_{ij} = a_i b_j c_{ij}$, where a_i and b_j are row and column marginal effects, and $c_{ij} = c_{ji}$ indicates

2 Scaling Asymmetric Matrices

a symmetric similarity between i and j . This model postulates that after removing the marginal effects, the remaining relation is symmetric. (The special case, in which $a_i = b_i$ for all i , leads to a full symmetric model.) The quasi-symmetry also satisfies the cycle condition stated as $x_{ij}x_{jk}x_{ki} = x_{ji}x_{kj}x_{ik}$. In some cases, the symmetric similarity parameter, c_{ij} , may further be represented by a simpler model, $c_{ij} = \exp(-d_{ij})$, or $c_{ij} = \exp(-d_{ij}^2)$, where d_{ij} is the Euclidean distance between stimuli i and j , represented as points in a multidimensional space.

DEDICOM (DEcomposing Directional COMponents, [4]) attempts to explain asymmetric relationships between n stimuli by a smaller number of asymmetric relationships. The DEDICOM model is written as $X = ARA'$, where R is a square asymmetric matrix of order r (capturing asymmetric relationships between r components, where r is assumed much smaller than n), and A is an n by r matrix that relates the latent asymmetric relationships among the r components to the observed asymmetric relationships among the n stimuli. Several algorithms have been developed to fit the DEDICOM model. To illustrate, the DEDICOM model is applied to a table of car switching frequencies among 16 types of cars [4]. (This table indicates frequencies with which a purchase of one type of car is followed by a purchase of another type by the same consumer.) Table 2 reports the analysis results [5]. Labels of the 16 car types consist of two components. The first three characters mainly indicate size (SUB = subcompact, SMA = small specialty, COM = compact, MID = midsize, STD = standard, and LUX = luxury), and the fourth character indicates mainly origin or price (D = domestic, C = captive imports, I = imports, L = low price, M = medium price, and S = specialty). The top portion of the table gives the estimated A matrix (normalized so that $A'A = I$), from which we may deduce that the first component (dimension) represents plain large and mid-size cars, the second component represents fancy large cars, and the third component represents small/specialty cars. The bottom portion of the table represents the estimated R matrix that captures asymmetry relationships among the three components. There are more switches from 1 to 3, 1 to 2, and 2 to 3 than the other way round. This three-component DEDICOM model captures 86.4% of the total SS (sum of squares) in the original data.

Table 2 DEDICOM applied to car switching data

Matrix A			
Car Class	Dimension		
	1	2	3
SUBD	.13	-.02	.36
SUBC	.02	.00	.03
SUBI	.03	.01	.30
SMAD	.01	.03	.53
SMAC	.00	.00	.00
SMAI	.00	.01	.09
COML	.24	-.11	.17
COMM	.10	-.01	.06
COMI	.02	.00	.03
MIDD	.54	.00	.12
MIDI	.02	.00	.02
MIDS	.09	.24	.58
STDL	.68	-.08	-.18
STDm	.32	.67	-.27
LUXD	-.23	.69	.05
LUXI	.00	.02	.01
Matrix R (divided by 1000)			
dim. 1	127	57	78
dim. 2	26	92	23
dim. 3	17	12	75

Any asymmetric table can be decomposed into the sum of a symmetric matrix (X_s), and a skew symmetric matrix (X_{sk}). That is, $X = X_s + X_{sk}$, where $X_s = (X + X')/2$, and $X_{sk} = (X - X')/2$. The two parts are often analyzed separately. X_s is often analyzed by a symmetric model (such as the inner product model or a distance model like those for c_{ij} described above). X_{sk} , on the other hand, is either treated like a skew symmetric matrix arising from an antisymmetric relationship, or by CASK (Canonical Analysis of SKew symmetric data, [3]). The latter decomposes X_{sk} in the form of AKA' , where K consists of 2 by 2 diagonal blocks of the form $\begin{pmatrix} 0 & k_l \\ -k_l & 0 \end{pmatrix}$ for the l th block. This representation can be analytically derived from the singular value decomposition of X_{sk} .

Generalized GIPSCAL [6] and HCM (Hermitian Canonical Model, [2]) analyze both parts (X_s and X_{sk}) simultaneously. The former represents X by $B(I_r + K)B'$ (where the BB' part represents X_s and the BKB' part represents X_{sk}), under the assumption that the skew symmetric part of R (that is, $(R - R')/2$) in DEDICOM is positive definite. The HCM

first forms an hermitian matrix, H , by $H = X_s + i X_{sk}$ (where i is a symbol for an imaginary number, $i = \sqrt{-1}$), and obtains the eigenvalue-vector decomposition of H .

References

- [1] Bradley, R.A. & Terry, M.E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons, *Biometrika* **39**, 324–345.
- [2] Escoufier, Y. & Grorud, A. (1980). in *Data Analysis and Informatics* E. Diday, L. Lebart, J.P. Pages & Tomassone, eds, North Holland, Amsterdam, pp. 263–276.
- [3] Gower, J.C. (1977). The analysis of asymmetry and orthogonality, in J.R. Barra, F. Brodeau, G. Romier & B. van Cuten, eds, *Recent Developments in Statistics*, North Holland. Amsterdam, pp. 109–123.
- [4] Harshman, R.A. Green, P.E. Wind, Y. & Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research, *Marketing Science* **1**, 205–242.
- [5] Kiers, H.A.L. & Takane, Y. (1993). Constrained DEDICOM, *Psychometrika* **58**, 339–355.
- [6] Kiers, H.A.L. & Takane, Y. (1994). A generalization of GIPSCAL for the analysis of nonsymmetric data, *Journal of Classification* **11**, 79–99.
- [7] Luce, R.D. (1959). *Individual Choice Behavior: a Theoretical Analysis*, Wiley, New York.
- [8] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.

(See also **Multidimensional Scaling**)

YOSHIO TAKANE

Scaling of Preferential Choice

ULF BÖCKENHOLT

Volume 4, pp. 1790–1794

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scaling of Preferential Choice

Choice data can be collected by either observing choices in the daily context of the decision makers or by asking a person directly to state their preferences for a single or multiple sets of options. Both data types, which are referred to as revealed and stated preference data (Louviere et al. [5]), may yield similar outcomes. For instance, in an election votes for political candidates represent revealed choice data. Rankings of the same candidates in a survey shortly before the election are an example for stated choice data. The stated preference data may prove useful in predicting the election outcome and in providing more information about the preference differences among the candidates than would be available from the election results alone.

Scaling models serve the dual purpose to summarize stated and revealed choice data and to facilitate the forecasting of choices made by decision makers facing possibly new or different variants of the choice options (Marshall & Bradlow [9]). The representation determined by the scaling methods can provide useful information for identifying both option characteristics that influence the choices and systematic sources of individual differences in the evaluation of these option characteristics. For example, in the election study, voters may base their decision on a set of attributes (e.g., integrity, leadership) but differ in the weights they assign to the attribute values for the different candidates. An application of scaling models to the voting data may both reveal these attributes and provide insights about how voters differ in their attribute assessments of the candidates.

The relationship between the attribute values and the corresponding preference judgments may be monotonic or nonmonotonic. Thus, decision makers may assess ‘higher’ attribute values as more favorable than ‘lower’ ones (e.g., quality of a product), or they may prefer a certain quantity of an attribute and dislike deviations in either directions from it (e.g., sweetness of a drink). In the latter case, individuals choose the option that is closest to their ‘ideal’ option where closeness is a function of the distance between the choice option and the person – specific ideal (De Leuw [2]). Distance between choice options may be defined in various ways. Applications in the literature

include approaches based on the Euclidean measure in a continuous attribute space and tree structures in which both choice and ideal options are represented by nodes (Carroll & DeSoete [1]). In either case, the interpretation of individual preference differences is much simplified provided decision makers use the same set of attributes in assessing the choice options.

Thurstone’s [14] random-utility approach has been highly influential in the development of many scaling models. Noting that choices by the same person may vary even under seemingly identical conditions, Thurstone argued that choices could be described as realizations of random variables that represent the options’ effects on a person’s sensory apparatus. According to this framework, a choice of an option i by decision maker j is determined by an unobserved utility assessment, v_{ij} , that can be decomposed into a systematic and a random part: $v_{ij} = \mu_{ij} + \epsilon_{ij}$. The person-specific item mean μ_{ij} is assumed to stay the same in repeated evaluations of the item but the random contribution ϵ_{ij} varies from evaluation to evaluation according to some distribution. Thurstone (1927) postulated that the ϵ_{ij} ’s follow a normal distribution. The assumption that the ϵ_{ij} ’s are independently Gumbel or Gamma distributed leads to scaling models proposed by Luce [6] and Stern [12]), respectively. Marden [8] and Takane [13] provide general discussions of these different specifications.

According to the latent utility framework, choosing the most preferred option is equivalent to selecting the option with the largest utility. Thus, an important feature of this choice process is that it is comparative in nature. A selected option may be the best one out of a set of available options but it may be rejected in favor of other options that are added subsequently to the set of options. Because choices are inherently comparative, the origin of the utility scale cannot be identified on the basis of the choices alone. One item may be preferred to another one but this result does not allow any conclusions about whether either of the items are attractive or unattractive.

One may question the use of randomness as a device to represent factors that determine the formation of preferences but are unknown to the observer. However, because, in general, it is not feasible to identify or measure all relevant choice determinants (such as all attributes of the choice options of the person choosing, or of environmental factors), it is not possible to answer conclusively the question of

2 Scaling of Preferential Choice

whether the choice process is inherently random or is determined by a multitude of different factors. Fortunately, for the development of scaling models this issue is not critical because either position arrives at the same conclusion that choices are described best in terms of their probabilities of occurrence (Manski & McFadden [7]).

In recent years, a number of scaling models have been developed that by building on Thurstone's random-utility approach take into account systematic time and individual-difference effects (Keane [4]). Concurrently, with these developments experimental research in judgment and decision making demonstrated that choice processes are subject to many influences that go beyond the simple evaluations of items. For example, different framings of the same choice options may trigger different associations and evaluations with the results that seemingly minor changes in the phrasing of a question or in the presentation format can lead to dramatic changes in the response behavior of a person (Kahneman [3]). One major conclusion of this research is that the traditional assumption of respondents having well-defined preferences should be viewed as a hypothesis that needs to be tested as part of any modeling efforts.

Preference Data

Typically, stated choice data are collected in the form of incomplete and/or partial rankings. Consider a set of J choice alternatives ($j = 1, \dots, J$) and n decision makers ($i = 1, \dots, n$). For each decision maker i and choice alternative j , a vector $\mathbf{x}_{ij}$ of observed variables is available that describe partially the pair (i, j) . Incomplete ranking data are obtained when a decision maker considers only a subset of the choice options. For example, in the method of paired comparison, two choice options are presented at a time, and the decision maker is asked to select the more preferred one. In contrast, in a partial ranking task, a decision maker is confronted with all choice options and asked to provide a ranking for a subset of the J options. For instance, in the best–worst method, a decision maker is instructed to select the best and worst options out of the offered set of choice options. Both partial and incomplete approaches can be combined by offering multiple distinct subsets of the choice options and obtain partial or complete rankings for each of them. For instance, a judge may

be presented with all $\binom{J}{2}$ option pairs sequentially and asked to select the more preferred item in each case.

Presenting choice options in multiple blocks has several advantages. First, the judgmental task is simplified since only a few options need to be considered at a time. Second, it is possible to investigate whether judges are consistent in their evaluations of the choice options. For example, if options are presented in pairs, one can investigate whether respondents are transitive in their comparisons, that is, whether they prefer j to l when they choose j over k and k over l . Third, obtaining multiple judgments from each decision maker simplifies analyses of how individuals differ in their preferences for the choice options. Individual-difference analyses are discussed in more detail in the next section. These advantages need to be balanced with possible boredom and learning effects that may affect a person's evaluation of the choice options when the number of blocks is large.

Revealed choice data differ from stated choice data in a number of ways. Perhaps, most importantly the set of choice alternatives may be unknown and may vary among decision makers in systematic ways. The lack of knowledge of the considered choice set complicates any inferences about the relative advantages of the selected options. Moreover, only top choices are observed typically which provide little information about the nonchosen options. Finally, the timing and context of the revealed choices may vary from person to person which reduces the interindividual comparability of the results. For these reasons, it is useful frequently to combine revealed with stated choice data to obtain a richer and more informative understanding of the underlying preferences of the decision makers.

Thurstonian Models for Preference Data

The choices made by person i for a single choice set can be summarized by an ordering vector $\mathbf{r}_i$. For instance, $\mathbf{r}_i = (h, j, \dots, l, k)$ indicates that choice option h is judged superior to option j which in turn is judged superior to the remaining options, with the least preferred option being k . The probability of observing this ordering vector can be written as

$$\begin{aligned} \Pr(\mathbf{r}_i = (h, j, \dots, l, k) | \xi) \\ = \Pr[(v_{ih} - v_{ij} > 0) \cap \dots \cap (v_{il} - v_{ik} > 0)], \end{aligned} \quad (1)$$

where ξ contains the parameters of the postulated distribution function for v_{ij} . Let $\mathbf{C}_i$ be a $(J-1) \times J$ contrast matrix that indicates the sign of the differences among the ranked items for a given ranking $\mathbf{r}_i$ of J items. For example, for $J = 3$, and the ordering vectors $\mathbf{r}_i = (j, l, k)$ and $\mathbf{r}_{i'} = (k, j, l)$, the corresponding contrast matrices take on the form

$$\mathbf{C}_i = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{C}_{i'} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \quad (2)$$

where the three columns of the contrast matrices correspond to the items j , k , and l , respectively. (1) can then be written as

$$\Pr(\mathbf{r}_i | \xi) = \Pr(\mathbf{C}_i \mathbf{v}_i > \mathbf{0}), \quad (3)$$

where $\mathbf{v}_i = (v_{i1}, \dots, v_{iJ})$ contains the option utility assessments of person i . If, as proposed by Thurstone [14], the rankers' judgments of the J items are multivariate normal (see **Catalogue of Probability Density Functions**) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (see **Correlation and Covariance Matrices**), the distribution of the pairwise differences of the $\mathbf{v}$'s is also multivariate normal. Consequently, the probability of observing the rank order vector $\mathbf{r}_i$ can be determined by evaluating an $(J-1)$ -variate normal distribution,

$$\Pr(\mathbf{r}_i | \xi) = \frac{|\boldsymbol{\Gamma}_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{(J-1)}{2}}} \int_0^\infty \dots \int_0^\infty \exp\{-\frac{1}{2}(\boldsymbol{\delta}_i - \mathbf{x})' \boldsymbol{\Gamma}_i^{-1} (\boldsymbol{\delta}_i - \mathbf{x})\} d\mathbf{x}, \quad (4)$$

where $\boldsymbol{\delta}_i = \mathbf{C}_i \boldsymbol{\mu}$ and $\boldsymbol{\Gamma}_i = \mathbf{C}_i \boldsymbol{\Sigma} \mathbf{C}_i'$. Both the mean utilities and their covariances may be related to observed covariates $\mathbf{x}_{ij}$ to identify systematic sources of individual differences in the evaluation of the choice options.

When a decision maker chooses among the options for T choice sets, we obtain a $(J \times T)$ ordering matrix $\mathbf{R}_i = (\mathbf{r}_{i1}, \dots, \mathbf{r}_{iT})$ containing person's i rankings for each of the choice set. With multiple-choice sets, it becomes possible to distinguish explicitly between within- and between-choice set variability in the evaluation of the choice options. Both sources of variation are confounded when preferences for only a single choice set are elicited. For

example, when respondents compare sequentially all possible pairs of choice options, $T = \binom{J}{2}$, the probability of observing the ordering matrix $\mathbf{R}_i$ is obtained by evaluating a $\binom{J}{2}$ -dimensional normal distribution function with mean vector $\mathbf{A}_i \boldsymbol{\mu}$ and covariance matrix $\mathbf{A}_i \boldsymbol{\Sigma} \mathbf{A}_i' + \boldsymbol{\Psi}$. The rows of $\mathbf{A}_i$ contain the contrast vectors $\mathbf{c}_{ik}$ corresponding to the choice outcome for the k -th choice set and $\boldsymbol{\Psi}$ is a diagonal matrix containing the within-choice-set variances.

For large J and/or T , the evaluation of the normal distribution function by numerical integration is not feasible with current techniques. Fortunately, alternative methods are available based on Monte Carlo Markov chain methods (see **Markov Chain Monte Carlo and Bayesian Statistics**) (Yao et al. [16], Tsai et al. [15]) or limited-information approximations (Maydeu-Olivares [10]) that can be used for estimating the mean and covariance structure of the choice options. Especially, limited-information methods are sufficiently convenient from a computational perspective to facilitate the application of Thurstonian scaling models in routine work.

An Application: Modeling the Similarity Structure of Values

An important issue in value research is to identify the basic value dimensions and their underlying structure. According to a prominent theory by Schwartz [11] the primary content aspect of a value is the type of goal it expresses. Based on extensive cross-cultural research, this author identified 10 distinct values: [1] power, [2] achievement, [3] hedonism, [4] stimulation, [5] self-direction, [6] universalism, [7] benevolence, [8] tradition, [9] conformity, and [10] security. The circular pattern in Figure 1 displays the similarity and polarity relationships among these values. The closer any two values in either direction around the circle, the more similar their underlying motivations. More distant values are more antagonistic in their underlying motivations. As a result of these hypothesized relationships, one may expect that associations among the value items exhibit systematic increases and decreases depending on their closeness and their degree of polarity.

To test this hypothesized value representation, binary paired comparison data were collected from a random sample of 338 students at a North-American university. The students were asked to indicate for

4 Scaling of Preferential Choice

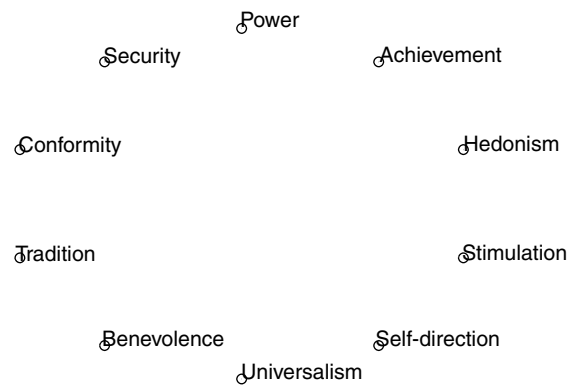


Figure 1 Hypothesized circumplex structure of ten motivationally distinct types of values

each of the 45 pairs formed on the basis of the 10 values, which one was more important as a guiding principle in their life. The respondents were highly consistent in their importance evaluations with less than 5% of the pairwise judgments being intransitive.

Figure 2 displays the first two **principal components** of the estimated covariance matrix $\hat{\Sigma}$ of the ten values. Because the origin of the value scale cannot be identified on the basis of the pairwise judgments,

the estimated coordinates may be rotated or shifted in arbitrary ways, only the distances between the estimated coordinates should be interpreted. Item positions that are closer to each other have a higher covariance than points that are further apart. Consistent with the circular value representation, the first component contrasts self-enhancement values with values describing self-transcendence and the second component contrasts openness-to-change with conservation values. However, the agreement with the circumplex structure is far from perfect. Several values, most notably ‘self-direction’ and ‘security’, deviate systematically from their hypothesized positions.

Concluding Remarks

Scaling models are useful in providing parsimonious descriptions of how individuals perceive and evaluate choice options. However, preferences may not always be well-defined and may depend on seemingly irrelevant contextual conditions (Kahneman [3]). Diverse factors such as the framing of the choice task and the set of offered options have been shown to influence strongly choice outcomes. As a result, generalizations of scaling results to different choice situations and

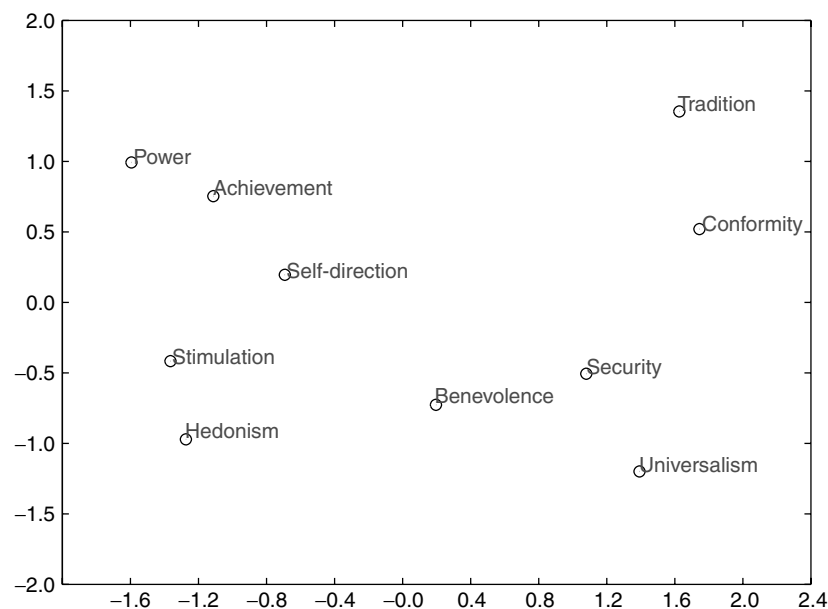


Figure 2 The two major dimensions of the covariance matrix $\hat{\Sigma}$ estimated from pairwise value judgments

options require much care and frequently need to be based on additional validation studies.

References

- [1] Carroll, J.D. & DeSoete, G. (1991). Toward a new paradigm for the study of multiattribute choice behavior, *American Psychologist* **46**, 342–352.
- [2] DeLeeuw, J. (2005). Multidimensional unfolding, *Encyclopedia of Behavioral Statistics*, Wiley, New York.
- [3] Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics, *American Economic Review* **93**, 1449–1475.
- [4] Keane, M.P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior, *Journal of Business and Economic Statistics* **15**, 310–327.
- [5] Louviere, J.J., Hensher, D.A. & Swait, J.D. (2000). *Stated Choice Methods*, Cambridge University Press, New York.
- [6] Luce, R.D. (1959). *Individual Choice Behavior*, Wiley, New York.
- [7] Manski, C. & McFadden, D., eds (1981). *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge.
- [8] Marden, J.I. (2005). The Bradley-Terry model, *Encyclopedia of Behavioral Statistics*, Wiley, New York.
- [9] Marshall, P. & Bradlow, E.T. (2002). A unified approach to conjoint analysis models, *Journal of the American Statistical Association* **97**, 674–682.
- [10] Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data, *Mathematical Social Sciences* **43**, 467–483.
- [11] Schwartz, S.H. (1992). Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries, *Advances in Experimental Social Psychology* **25**, 1–65.
- [12] Stern, H. (1990). A continuum of paired comparison models, *Biometrika* **77**, 265–273.
- [13] Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data, *Cognition and Communication* **20**, 45–62.
- [14] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.
- [15] Tsai, R.-C. & Böckenholt, U. (2002). Two-level linear paired comparison models: estimation and identifiability issues, *Mathematical Social Sciences* **43**, 429–449.
- [16] Yao, G. & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler, *British Journal of Mathematical and Statistical Psychology* **52**, 79–92.

(See also **Attitude Scaling; Multidimensional Scaling; Multidimensional Unfolding**)

ULF BÖCKENHOLT

Scatterplot Matrices

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

Volume 4, pp. 1794–1795

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scatterplot Matrices

When researchers are interested in the relationships between pairs of several continuous variables, they often produce a series of **scatterplots** for each of the pairs. It can be convenient to view these together on a single screen or page using what is usually called a *scatterplot matrix* (though sometimes referred to as a *draftman's plot*). Many statistical packages have this facility. With k variables, there are $k(k - 1)/2$ pairs, and therefore for even small numbers of variables the number of scatterplots can be large. This means each individual scatterplot on the display is small. An example is shown in Figure 1.

Scatterplot matrices are useful for quickly ascertaining all the bivariate relationships, but because of the size of the individual scatterplot it may be difficult to fully understand the relationship. Some of the extra facilities common for two variable scatterplots, such as adding symbols and including confidence limits on regression lines, would create too much clutter in a scatterplot matrix. Here, we have included a line for the linear regression and the univariate **histograms**. Any more information would be difficult to decipher.

Figure 1 just shows bivariate relationships. Sometimes, it is useful to look at the bivariate relationship between two variables at different values or levels of a third variable. In this case, we produce a **trellis display** or casement display. Consider the following study [3] in which participants heard lists of semantically associated words and were played a

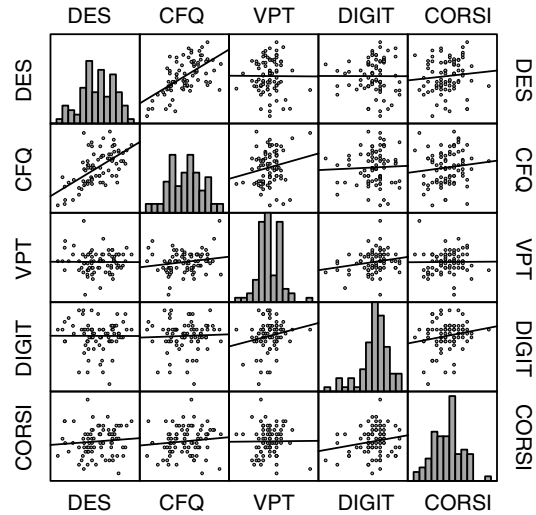


Figure 1 A scatterplot matrix that shows the bivariate relationships between two personality measures (DES – dissociation, CFQ – cognitive failures questionnaire) and impairment from secondary tasks on three working memory tasks (VPT – visual patterns, DIGIT – digit span, CORSI – Corsi block test). Data from [2]

piece of music. Later, they were asked to recall the words, and how many times the participant recalled a semantically related word that was not originally presented (a lure) was recorded. Figure 2 shows the relationship between the number of lures recalled and how much the participant liked the music. There were two experimental conditions. In the first, participants were told to recall as many as words as

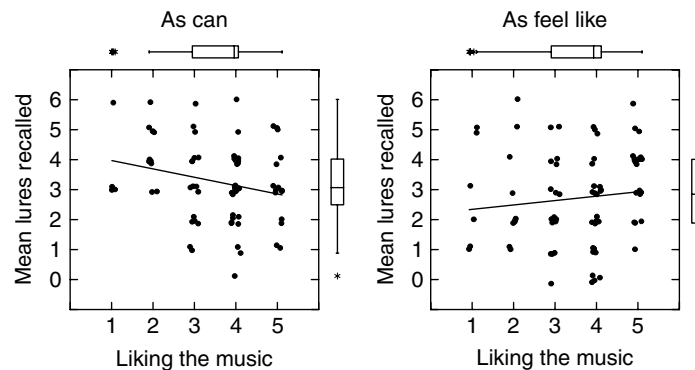


Figure 2 The relationship between liking the music and the number of critical lures recalled depends on the recall task. A random jitter has been added to the points that all are visible. Data from [3]

2 Scatterplot Matrices

they could. The more the participant liked the music, the fewer lures were recalled. The argument is that the music put these people in a good mood so they felt satisfied with their recall so did not try as hard. In the second condition, participants were told to recall as many as words as they felt like. Here, the more people liked the music, the more lures they recalled, that is, if they were happy because of the music, they continued to feel like recalling words.

Trellis scatterplots can be used with more than one conditioning variable. However, with more than two conditioning variables, they can be difficult to interpret. If multivariate relationships are of interest, other techniques, such as **three-dimensional scatterplots** and **bubble plots**, are more appropriate.

A useful source for further information on scatterplot matrices is [1].

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.
- [2] Wright, D.B. & Osborne, J. (2005 in press). Dissociation, cognitive failures, and working memory, *American Journal of Psychology*.
- [3] Wright, D.B., Startup, H.M. & Mathews, S. (under review). Dissociation, mood, and false memories using the Deese-Roediger-McDermott procedure, *British Journal of Psychology*.

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

Scatterplot Smoothers

BRIAN S. EVERITT

Volume 4, pp. 1796–1798

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scatterplot Smoothers

The **scatterplot** is an excellent first exploratory graph with which to study the dependence of two variables. Often, understanding of the relationship between the two variables is aided by adding the result of a simple linear fit to the plot. Figure 1 shows such a plot for the average oral vocabulary size of children at various ages.

Here, the linear fit does not seem adequate to describe the growth in vocabulary size with increasing age, and some form of polynomial curve might be more appropriate. Since the plotted observations show a tendency to an ‘S’ shape, a cubic might be a possibility. If such a curve is fitted to the data, it appears to fit the available data well but between the observations, it rises and then drops again. Consequently, as a model of language acquisition, it leads to the absurd implication that newborns have large vocabularies, which they lose by the age on one, then their vocabulary increases until the age of six, when children start to forget words rather rapidly! Not a very sensible model.

An alternative to using parametric curves to fit bivariate data is to use a nonparametric approach in which we allow the data themselves to suggest the appropriate functional form. The simplest of these alternatives is to use a locally weighted regression or loess fit, first suggested by Cleveland [1]. In essence, this approach assumes that the variables x and y are related by the equation

$$y_i = g(x_i) + \epsilon_i, \quad (1)$$

where g is a ‘smooth’ function and the ϵ_i are random variables with mean zero and constant scale. Values $\hat{y}_i$ used to ‘estimate’ the y_i at each x_i are found by fitting polynomials using weighted least squares with large weights for points near to x_i and small weights otherwise. So smoothing takes place essentially by local averaging of the y -values of observations having predictor values close to a target value. Adding such a plot to the data is often a useful alternative to the more familiar parametric curves such as simple linear or polynomial regression fits (*see Multiple Linear Regression; Polynomial Model*) when the bivariate data plotted is too complex to be described by a simple parametric family. Figure 2 shows the result of fitting a locally weighted regression curve to the vocabulary data. The locally weighted regression fit is able to follow the nonlinearity in the data although the difference in the two curves is not great.

An alternative smoother that can often usefully be applied to bivariate data is some form of spline function. (A spline is a term for a flexible strip of metal or rubber used by a draftsman to draw curves.) Spline functions are polynomials within intervals of the x -variable that are connected across different values of x . Figure 3, for example, shows a linear spline function, that is a piecewise linear function, of the form

$$f(x) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+$$

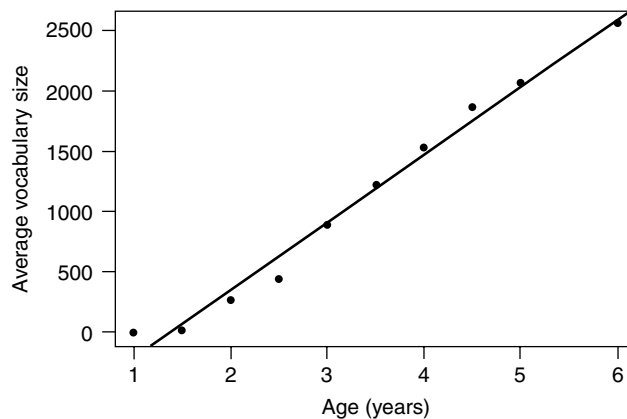


Figure 1 Scatterplot of average vocabulary score against age showing linear regression fit

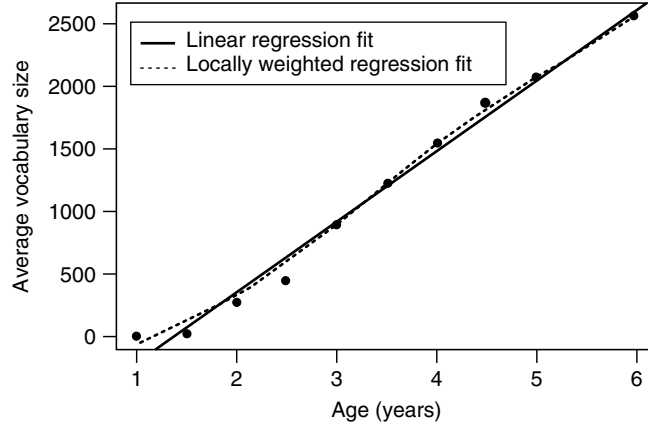


Figure 2 Scatterplot of vocabulary score against age showing linear regression fit and locally weighted regression fit

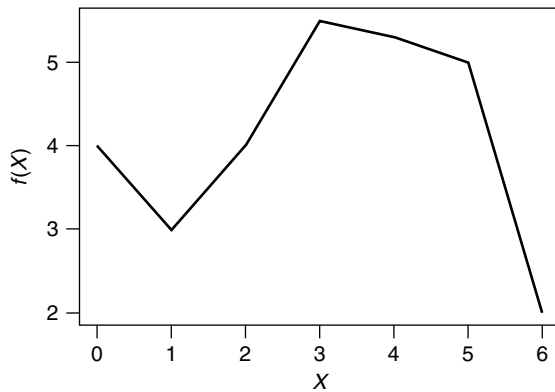


Figure 3 A linear spline function with knots at $a = 1$, $b = 3$, $c = 5$. Taken with permission from [3]

$$\begin{aligned} \text{where } (u)_+ &= u & u > 0 \\ &= 0 & u \leq 0, \end{aligned} \quad (2)$$

The interval endpoints, a , b , and c are called *knots*. The number of knots can vary according to the amount of data available for fitting the function.

The linear spline is simple and can approximate some relationships, but it is not smooth and so will not fit highly curved functions well. The problem is overcome by using piecewise polynomials, in particular, cubics, which have been found to have nice properties with good ability to fit a variety of complex relationships. The result is

a cubic spline that arises formally by seeking a smooth curve $g(x)$ to summarize the dependence of y on x , which minimizes the rather daunting expression:

$$\sum [y_i - g(x_i)]^2 + \lambda \int g''(x)^2 dx, \quad (3)$$

where $g''(x)$ represents the second derivative of $g(x)$ with respect to x . Although when written formally this criterion looks a little formidable, it is really nothing more than an effort to govern the trade-off between the goodness-of-fit of the data (as measured by $\sum [y_i - g(x_i)]^2$) and the ‘wiggliness’ or departure of linearity of g measured by $\int g''(x)^2 dx$; for a linear function, this part of (3) would be zero. The parameter λ governs the smoothness of g , with larger values resulting in a smoother curve.

The solution to (3) is a cubic spline, that is, a series of cubic polynomials joined at the unique observed values of the explanatory variable, x_i . (For more details, see [2]). Figure 4 shows a further scatterplot of the vocabulary data now containing linear regression, locally weighted regression, and spline smoother fits. When interpolating a number of points, a spline can be a much better solution than a polynomial interpolant, since the polynomial can oscillate wildly to hit all the points; polynomial fit the data globally, while splines fit the data locally.

Locally weighted regressions and spline smoothers are the basis of **generalized additive models**.

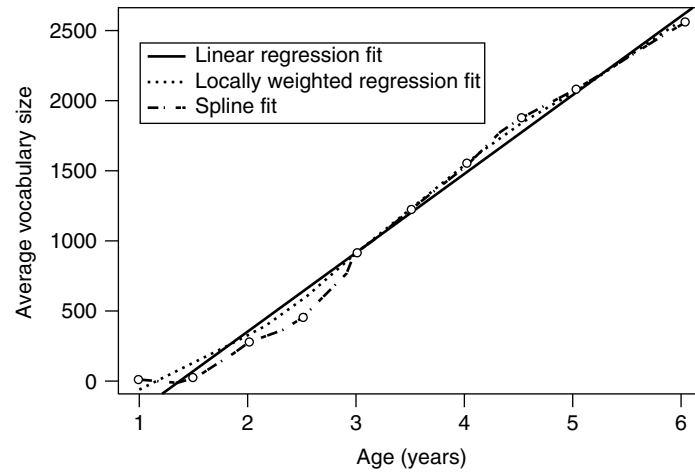


Figure 4 Scatterplot of vocabulary score against age showing linear regression, locally weighted regression, and spline fits

References

- [1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.
- [2] Friedman, J.H. (1991). Multiple adaptive regression splines, *Annals of Statistics* **19**, 1–67.
- [3] Harrell, F.E. (2001). *Regression Modelling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer, New York.

BRIAN S. EVERITT

Scatterplots

SIÂN E. WILLIAMS AND DANIEL B. WRIGHT

Volume 4, pp. 1798–1799

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scatterplots

Scatterplots are typically used to display the relationship, or association, between two variables. Examples include the relationship between *age* and *salary* and that between inches of *rainfall* in a month and the number of *car accidents*. Both variables need to be measured on some continuum or scale. If there is a natural response variable or a predicted variable, then it should be placed on the y-axis. For example, age would be placed on the x-axis and salary on the y-axis because it is likely that you would hypothesize that salary is, in part, dependant on age rather than the other way round.

Consider the following example. The estimated number of days in which students expect to take to complete an essay is compared with the actual number of days taken to complete the essay. The scatterplot in Figure 1 shows the relationship between the estimated and actual number of days.

Most statistical packages allow various options to increase the amount of information presented. In Figure 1, a diagonal line is drawn, which corresponds to positions where estimated number of days equals actual number of days. Overestimators fall below the diagonal, and underestimators fall above the diagonal. You can see from this scatterplot that most students underestimated the time it took them to complete the essay. **Box plots** are also included here to provide univariate information.

Other possible options include adding different regression lines to the graph, having the size of the

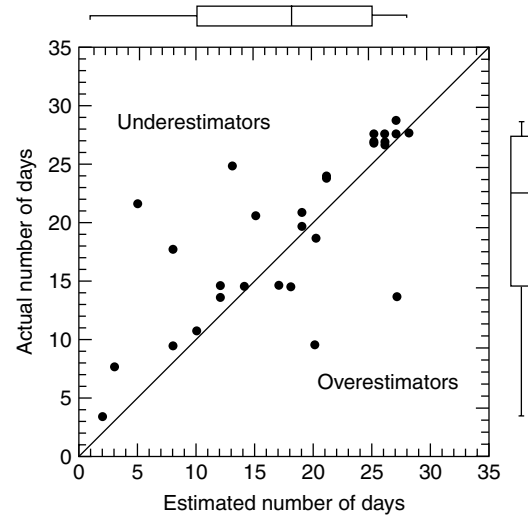


Figure 1 A scatterplot of estimated and actual essay completion times with overlapping case points

points represent their impact on the regression line, using 'sunflowers', and 'jittering'. The use of the sunflowers option and jittering option allow multiple observations falling on the same location of the plot to be counted. Consider Figures 2(a), (b), and (c). Participants were presented with a cue event from their own autobiography and were asked whether that event prompted any other memory [2]. Because participants did this for several events, there were 1865 date estimates in total. If a standard scatterplot is produced comparing the year of the event with the year of the cueing event, the result is Figure 2(a). Because of the large number of events, and the fact that many

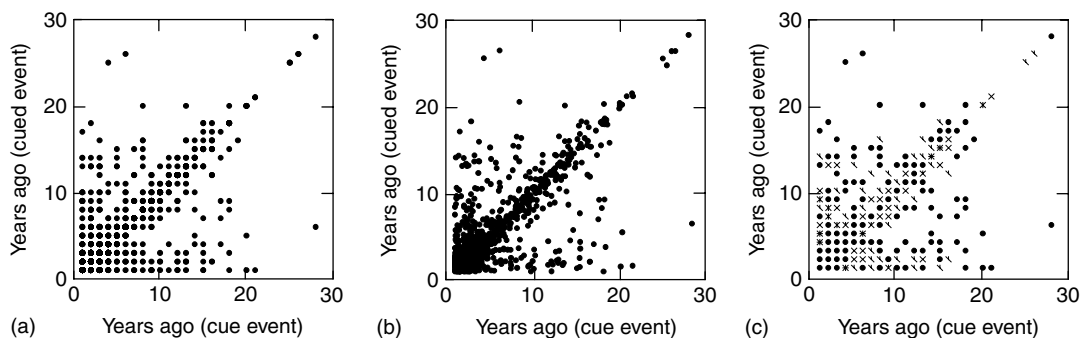


Figure 2 Scatterplots comparing the year of a remembered event with the year of the event that was used to cue the memory. Adapted from Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489, with permission from John Wiley & Sons Ltd

2 Scatterplots

overlap, this graph does not allow the reader to determine how many events are represented by each point.

In Figure 2(b), each data point has had a random number (uniformly distributed between -0.45 and +0.45) added to both its horizontal and vertical component. The result is that coordinates with more data points have more dots around them. Jittering is particularly useful with large data sets, like this one. Figure 2(c) shows the sunflower option. Here, individual coordinates are represented with sunflowers. The number of petals represents the number of data points. This option is more useful with smaller data sets. More information on jittering and sunflowers can be found in, for example, [1].

It is important to realize that a scatterplot does not summarize the data; it shows each case in terms

of its value on variable x and its value on variable y . Therefore, the choice of regression line and the addition of other summary information can be vital for communicating the main features in your data.

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.
- [2] Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489.

SIÂN E. WILLIAMS AND DANIEL B. WRIGHT

Scheffé, Henry

BRIAN S. EVERITT

Volume 4, pp. 1799–1800

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Scheffé, Henry

Born: April 11, 1907, in New York, USA.

Died: July 5, 1977, in California, USA.

Born in New York City, Scheffé attended elementary school in New York and graduated from high school in Islip, Long Island, in 1924. In 1928, he went to study mathematics at the University of Wisconsin receiving his B.A. in 1931. Four years later, he was awarded a Ph.D. for his thesis entitled 'Asymptotic solutions of certain linear differential equations in which the coefficient of the parameter may have a zero'. Immediately after completing his doctorate, Scheffé began a career as a university teacher in pure mathematics. It was not until 1941 that Scheffé's interests moved to statistics and he joined Samuel Wilks and his statistics team at Princeton. From here he moved to the University of California, and then to Columbia University where he became chair of the statistics department. In 1953, he left Columbia for Berkeley where he remained until his retirement in 1974.

Much of Scheffé's research was concerned with particular aspects of the **linear model** (e.g., [2]), particularly, of course, the **analysis of variance**,

resulting in 1959 in the publication of his classic work, *The Analysis of Variance*, described in [1] in the following glowing terms:

Its careful exposition of the different principal models, their analyses, and the performance of the procedures when the model assumptions do not hold is exemplary, and the book continues to be a standard list and reference.

Scheffé's book has had a major impact on generations of statisticians and, indirectly, on many psychologists. In 1999, the book was reprinted in the Wiley Classics series [3].

During his career, Scheffé became vice president of the American Statistical Association and president of the International Statistical Institute and was elected to many other statistical societies.

References

- [1] Lehmann, E.L. (1990). Biography of Scheffé, *Dictionary of Scientific Biography*, Supplement II Vol. 17–18, Scribner, New York.
- [2] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.
- [3] Scheffé, H. (1999). *The Analysis of Variance*, Wiley, New York.

BRIAN S. EVERITT

Second Order Factor Analysis: Confirmatory

JOHN TISAK AND MARIE S. TISAK

Volume 4, pp. 1800–1803

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Second Order Factor Analysis: Confirmatory

To motivate the discussion of confirmatory second-order factor analysis, a basic illustration will be provided to highlight the salient features of the topic. For pedagogical reason, the example will be used when all the variables are measured and then it will be repeated when some of the variables are not measured or are latent. This contrast emphasizes the functional similarities between these two approaches and associates the commonly known regression analysis (*see Multiple Linear Regression*) with the more complex and less familiar **factor analysis**.

MEASURED VARIABLES. Assume that a random sample of individuals from a specified population has been assessed on Thurstone's Primary Mental Abilities (PMA) scales. Concretely, these abilities are as follows:

- Verbal Meaning (VRBM) – a vocabulary recall ability test;
- Word Grouping (WORG) – a test of vocabulary recall ability;
- Number Facility (NUMF) – a measure of arithmetic reasoning ability;
- Letter Series (LETS) – tests reasoning ability by letter series; and
- Number Series (NUMS) – a reasoning ability test that uses number series.

Second, assume that the same individuals are measured on Horn and Cattell's Crystallized and Fluid Intelligences. Specifically, these two intelligences are as follows:

- Crystallized Intelligence (GC) – measured learned or stored knowledge; and
- Fluid Intelligence (GF) – evaluates abstract reasoning capabilities.

Third, consider that once again these individuals are tested on Spearman's General Intelligence (G), which is a global construct of general ability or intelligence. Notice that in moving from the Primary Mental Abilities to Crystallized and Fluid Intelligences to General Intelligence, there is a movement from more specific to more general constructs, which could be considered nested, that is, the more specific

variables are a subset of the more general variables. Finally, it is implicit that these variables are infallible or are assessed without measurement error.

First-order regression analysis. If these eight variables were all assessed, then one could evaluate how well the more general Crystallized (GC) and Fluid (GF) Intelligences predict the Primary Mental Abilities using multivariate multiple regression analysis. Precisely, the regression of the PMA scales onto Crystallized and Fluid Intelligences become

$$\text{VRBM} = \beta_0^1 + \beta_1^1 \text{GC} + \beta_2^1 \text{GF} + e^1,$$

$$\text{WORG} = \beta_0^2 + \beta_1^2 \text{GC} + \beta_2^2 \text{GF} + e^2,$$

$$\text{NUMF} = \beta_0^3 + \beta_1^3 \text{GC} + \beta_2^3 \text{GF} + e^3,$$

$$\text{LETS} = \beta_0^4 + \beta_1^4 \text{GC} + \beta_2^4 \text{GF} + e^4,$$

$$\text{and NUMS} = \beta_0^5 + \beta_1^5 \text{GC} + \beta_2^5 \text{GF} + e^5, \quad (1)$$

where β_0^j are the intercepts for predicting the j th outcome Primary Mental Abilities variable, and where β_1^j and β_2^j are the partial regression coefficients or slopes for predicting the j th outcome variable from Crystallized and Fluid Intelligences, respectively. Lastly, e^1 to e^5 are errors of prediction, that is, what is not explained by the prediction equation for each outcome variable.

Given knowledge of these variables, one could speculate that Crystallized Intelligence would be related to the Verbal Meaning and Word Grouping abilities, whereas Fluid Intelligence would predict the Number Facility, Letter Series, and Word Grouping abilities. Hence, the regression coefficients, $\beta_1^1, \beta_1^2, \beta_2^3, \beta_2^4$, and β_2^5 , would be substantial, while, $\beta_2^1, \beta_2^2, \beta_1^3, \beta_1^4$, and β_1^5 , would be relatively much smaller.

Second-order regression analysis. Along this line of development, Crystallized (GC) and Fluid (GF) Intelligences could be predicted by General Intelligence (G) by multivariate simple regression analysis. Concretely, the regression equations are

$$\text{GC} = \beta_0^6 + \beta_1^6 \text{G} + e^6$$

$$\text{and GF} = \beta_0^7 + \beta_1^7 \text{G} + e^7, \quad (2)$$

where β_0^6 and β_0^7 are the intercepts for predicting each of the Crystallized and Fluid Intelligence outcome variables, and where β_1^6 and β_1^7 are the partial regression coefficients or slopes for predicting the two outcome variables from General Intelligence

2 Second Order Factor Analysis: Confirmatory

respectively. Additionally, e^6 and e^7 again are errors of prediction. Substantively, one might conjecture that General Intelligence predicts Fluid Intelligence more than it does Crystallized Intelligence, hence β_1^6 would be less than β_1^7 .

LATENT VARIABLES. Suppose that Fluid (gf) and Crystallized (gc) Intelligences and General Intelligence (g) are not observed or measured, but instead they are **latent variables**. (Note that the labels on these three variables have been changed from upper to lower case to indicate that they are unobserved.) Thus, the impact of these variables must be determined by latent variable or factor analysis. As before, this analysis might be viewed in a two-step process: A first-order and a second-order factor analysis.

First-order factor analysis. If the five PMA variables were assessed, then one could evaluate how well the latent Crystallized (gc) and Fluid (gf) Intelligences predict the Primary Mental Abilities using a first-order factor analysis. (For details *see* **Factor Analysis: Confirmatory**) Precisely, the regression equations are

$$\begin{aligned} \text{VRBM} &= \tau_0^1 + \lambda_1^1 \text{gc} + \lambda_2^1 \text{gf} + \varepsilon^1, \\ \text{WORG} &= \tau_0^2 + \lambda_1^2 \text{gc} + \lambda_2^2 \text{gf} + \varepsilon^2, \\ \text{NUMF} &= \tau_0^3 + \lambda_1^3 \text{gc} + \lambda_2^3 \text{gf} + \varepsilon^3, \\ \text{LETS} &= \tau_0^4 + \lambda_1^4 \text{gc} + \lambda_2^4 \text{gf} + \varepsilon^4, \\ \text{and NUMS} &= \tau_0^5 + \lambda_1^5 \text{gc} + \lambda_2^5 \text{gf} + \varepsilon^5, \quad (3) \end{aligned}$$

where τ_0^j are the intercepts for predicting the j th observed outcome Primary Mental Abilities variable, and where λ_1^j and λ_2^j are the partial regression coefficients or slopes for predicting the j th outcome variable from unobserved Crystallized and Fluid Intelligences, respectively. Unlike multiple regression analysis, in factor analysis, the errors, ε^1 to ε^5 , are now called *unique factors* and each contain two entities, a specific factor and a measurement error. For example, the unique factor, ε^1 , consists of a specific factor that contains what is not predicted in Verbal Meaning by Crystallized and Fluid Intelligences and a measurement error induced by imprecision in the assessment of Verbal Meaning. Notice that these regression coefficients and errors have been relabeled to emphasize that the predictors are now latent variables and to be consistent with the nomenclature used in LISREL, a commonly used computer package for these analyses (*see* **Structural Equation Modeling:**

Software), but that their interpretation is analogous to those in the measured variable section.

As before, one could speculate that Crystallized Intelligence would be related to the Verbal Meaning and Word Grouping abilities, whereas Fluid Intelligence would predict the Number Facility, Letter Series, and Word Grouping abilities. Unfortunately, unlike before, the (latent) variables or factors, gc and gf, are unknown, which creates an indeterminacy in the previous equations, that is, the regression coefficients cannot be uniquely determined. A standard solution to this problem is the use of marker or reference variables. Specifically for each factor, an observed variable is selected that embodies the factor. For example, since Crystallized Intelligence could be considered learned knowledge, one might select Verbal Meaning, accumulated knowledge, to represent it. For this case, the intercept is equated to zero ($\tau_0^1 \equiv 0$); the slope associated with gc equated to one ($\lambda_1^1 \equiv 1$) and the slope associated with gf equated to zero ($\lambda_2^1 \equiv 0$). Hence for Verbal Meaning, the regression equation becomes $\text{VRBM} = \text{gc} + \varepsilon^1$. By similar reasoning, Number Facility could serve as a reference variable for Fluid Intelligence, because they are linked by reasoning ability. Hence, for Number Facility, the regression equation is $\text{NUMF} = \text{gf} + \varepsilon^3$. Implicit is that the associated intercept and slope for gc are zero ($\tau_0^3 \equiv 0$ and $\lambda_1^3 \equiv 0$) and that the slope for gf is one ($\lambda_2^3 \equiv 1$).

Again, in keeping with our theoretical speculation (i.e., that VRBM and WORG are highly predicted from gc, but not gf, and vice versa for NUMF, LETS, and NUMS), the regression coefficients or loadings, λ_1^2 , λ_2^4 , and λ_2^5 , would be substantial, while λ_2^2 , λ_1^4 , and λ_1^5 , would be relatively much smaller.

Second-order factor analysis. Similar to second-order regression analysis, the latent variables, Crystallized (gc) and Fluid (gf) Intelligences, could be predicted by General Intelligence (g) by a second-order factor analysis. Concretely,

$$\begin{aligned} \text{gc} &= \alpha_0^1 + \gamma_1^1 \text{g} + \zeta^1 \\ \text{and gf} &= \alpha_0^2 + \gamma_1^2 \text{g} + \zeta^2, \quad (4) \end{aligned}$$

where α_0^1 and α_0^2 are the intercepts for predicting each of the Crystallized and Fluid Intelligence outcome variables, and where γ_1^1 and γ_1^2 are the regression coefficients or slopes for predicting the two outcome variables from General Intelligence, respectively. Further, ζ^1 and ζ^2 are second-order

specific factors or what has not been predicted by General Intelligence in Crystallized and Fluid Intelligences, respectively. Lastly, recognize that measurement error is not present in these second-order specific factors, because it was removed in the first order equations.

As before, one might conjecture that General Intelligence relates more to Fluid Intelligence than it predicts Crystallized Intelligence. Furthermore, as with the first-order factor analysis, there is indeterminacy in that the regression coefficients are not unique. Hence, again a reference variable is required for each latent variable or second-order factor. For example, Fluid Intelligence could be selected as a reference variable. Thus, the intercept and slope for it will be set to zero and one respectively ($\alpha_0^2 \equiv 0$ and $\gamma_1^2 \equiv 1$). So, the regression equation for Fluid Intelligence becomes $gf = g + \zeta^2$.

TECHNICAL DETAILS. In the foregoing discussion, similarities between multiple regression and factor analysis were developed by noting the linearity of the function form or prediction equation. Additionally, first- and second-order models were developed separately. Parenthetically, when originally developed, a first-order factor analysis would be performed initially and then a second-order factor would be undertaken using these initial results (or, more accurately, the variances and covariances between the first-order factors). It is the current practice to perform both the first- and second-order factor analysis in the same model or at the same time. Moreover, remember that for the confirmatory approach, in either the first- or second-order factor analytic

model, *a priori* specifications are required to ensure the establishment of reference variables.

In the previous developments, the regression coefficients or parameters were emphasized because they denote the linear relationships among the variables. It is important to note that there are also means, variances, and covariances associated with the second-order factor analytic model. Specifically, the second-order factor – General Intelligence in the example – has a mean and a variance parameter. Note that if there was more than one second-order factor, there would be additional mean, variance, and covariance parameters associated with it. Also, the second-order specific factors typically have variance and covariance parameters (the mean of these specific factors are assumed to be zero). Finally, each of the unique factors from the first-order model has a variance parameter (by assumption, their means and covariances are zero).

If normality of the observed variables is tenable, then all of these parameters may be determined by a statistical technique called **maximum likelihood estimation**. Further, a variety of theoretical hypotheses may be evaluated or tested by chi-square statistics.

For example, one could test if Verbal Meaning and Word Grouping are predicted only by Crystallized Intelligence by hypothesizing that the regression coefficients for Fluid Intelligence are zero, that is, $\lambda_2^1 = 0$ and $\lambda_2^2 = 0$.

JOHN TISAK AND MARIE S. TISAK

Selection Study (Mouse Genetics)

NORMAN HENDERSON

Volume 4, pp. 1803–1804

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Selection Study (Mouse Genetics)

Some of the earliest systematic studies on the inheritance of behavioral traits in animals involved artificial selection. These include Tolman's selection for 'bright' and 'dull' maze learners in 1924, Rundquist's selection for 'active' and 'inactive' rats in a running wheel in 1933, and Hall's 1938 selection for 'emotional' and 'nonemotional' rats in an open field. These pioneering studies triggered an interest in selective breeding for a large variety of behavioral and neurobiological traits that has persisted into the twenty-first century. For a description of early studies and some of the subsequent large-scale behavioral selection studies that followed, see [2] and [5].

Selection experiments appear simple to carry out and most are successful in altering the levels of expression of the selected behavior in only a few generations. Indeed, for centuries preceding the first genetic experiments of Mendel, animal breeders successfully bred for a variety of behavioral and related characters in many species. The simplicity of artificial selection experiments is deceptive, however, and the consummate study requires a considerable effort to avoid problems that can undermine the reliability of the genetic information sought. Frequently, this genetic information includes the realized narrow **heritability** (h^2), the proportion of the observed phenotypic variance that is explained by additive genetic variance, but more often the genotypic correlations (*see* **Gene-Environment Correlation**) between the selected trait and other biobehavioral measures. A lucid account of the quantitative genetic theory related to selection and related issues involving small populations can be found in [1]. Some key issues are summarized here.

Realized heritability of the selected trait is estimated from the ratio of the response to selection to the selection differential:

$$\begin{aligned} h^2 &= \frac{\text{Response}}{\text{Selection differential}} \\ &= \frac{\text{Offspring mean} - \text{Base population mean}}{\text{Selected parent mean} - \text{Base population mean}} \end{aligned} \quad (1)$$

Normally, selection is bidirectional, with extreme scoring animals chosen to be parents for high and low lines, and the heritability estimates averaged. In the rare case where the number of animals tested in the base population is so large that the effective N of each of the selected parent groups exceeds 100, (1) will work well, even in a single generation, although h^2 is usually estimated from several generations by regressing the generation means on the cumulative selection differential. Also, when parent N s are very large and the offspring of high and low lines differ significantly on any other nonselected trait, one can conclude that the new trait is both heritable and genetically correlated with the selected trait, although the correlation cannot be estimated unless h^2 of the new trait is known [3].

Since few studies involve such large parent groups in each breeding generation, most selection experiments are complicated by the related effects of inbreeding, random genetic drift, and genetic differentiation among subpopulations at all genetic loci. Over succeeding generations, high-trait and low-trait lines begin to differ on many genetically influenced characters that are unrelated to the trait being selected for. The magnitude of these effects is a function of the effective breeding size (N_e), which influences the increment in the coefficient of inbreeding (ΔF) that occurs from one generation to the next. When selected parents are randomly mated, $\Delta F = 1/(2N_e)$. When sib matings are excluded, $\Delta F = 1/(2N_e + 4)$. N_e is a function of the number of male and female parents in a selected line that successfully breed in a generation:

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad (\text{approx.}) \quad (2)$$

N_e is maximized when the number of male and female parents is equal. Thus, when the parents of a selected line consist of 10 males and 10 females, $N_e = 20$ for that generation, whereas $N_e = 15$ when the breeding parents consist of 5 males and 15 females.

When selection is carried out over many generations, alleles that increase and decrease expression of the selected trait continue to segregate into high- and low-scoring lines, but inbreeding and consequently random drift are also progressing at all loci. If we define the base population as having an inbreeding coefficient of zero, then the inbreeding

2 Selection Study (Mouse Genetics)

coefficient in any subsequent selected generation, t , is approximately:

$$F_t = 1 - [(1 - \Delta F_1) \times (1 - \Delta F_2) \times (1 - \Delta F_3) \times \cdots \times (1 - \Delta F_t)] \quad (3)$$

If N_e , and thus ΔF , are constant over generations, (3) simplifies to $F_t = 1 - (1 - \Delta F)^t$. Typically, however, over many generations of selection, N_e fluctuates and may even include some 'genetic bottleneck' generations, where N_e is quite small and ΔF large. It can be seen from (3) that bottlenecks can substantially increase cumulative inbreeding. For example, maintaining 10 male and 10 female parents in each generation of a line for 12 generations will result in an F_t of approximately $1 - (1 - 0.025)^{12} = 0.26$, but, if in just one of those 12 generations only a single male successfully breeds with the 10 selected females, N_e drops from 20 to 3.6 for that generation, causing F_t to increase from 0.26 to 0.35.

As inbreeding continues over successive generations, within-line genetic variance decreases by $1 - F$ and between-line genetic variance increases by $2F$ at all loci. Selected lines continue to diverge on many genetically influenced traits due to random drift, which is unrelated to the trait being selected for. Genetic variance contributed by drift can also exaggerate or suppress the response to selection and is partly responsible for the variability in generation means as selection progresses. Without genotyping subjects, the effects of random drift can only be

assessed by having replicated selected lines. One cannot obtain empirical estimates of sampling variation of realized heritabilities in experiments that do not involve replicates. Since the lines can diverge on unrelated traits by drift alone, the lack of replicate lines also poses problems for the common practice of comparing high and low lines on new traits thought to be genetically related to the selected trait. Unless inbreeding is extreme or heritability low, the size of a high-line versus low-line mean difference in phenotypic SD units can help determine if the difference is too large to be reasonably due to genetic drift [4].

References

- [1] Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th Edition, Longman, New York.
- [2] Fuller, J.L. & Thompson, W.R. (1978). *Foundations of Behavior Genetics*, Mosby, New York.
- [3] Henderson, N.D. (1989). Interpreting studies that compare high- and low-selected lines on new characters, *Behavior Genetics* **19**, 473–502.
- [4] Henderson, N.D. (1997). Spurious associations in unreplicated selected lines, *Behavior Genetics* **27**, 145–154.
- [5] Plomin, R., DeFries, J.C., McClearn, G.E. & McGuffin, P. (2000). *Behavioral Genetics: A Primer*, 4th Edition, W. H. Freeman, New York.

(See also **Inbred Strain Study**)

NORMAN HENDERSON

Sensitivity Analysis

JON WAKEFIELD

Volume 4, pp. 1805–1809

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sensitivity Analysis

Introduction

Consider an experiment in which varying dosage levels of a drug are randomly assigned to groups of individuals. If the **randomization** is successfully implemented, the groups of subjects will be balanced with respect to all variables other than the dosage levels of interest, at least on average (and if the groups are sufficiently large, effectively in practice also). The beauty of randomization is that the groups are balanced not only with respect to measured variables, but also with respect to unmeasured variables. In a nonrandomized situation, although one may control for *observed* explanatory variables, one can never guarantee that observed associations are not due to unmeasured variables. Selection bias, in which the chances of observing a particular individual depend on the values of their responses and explanatory variables, is another potential source of bias. A further source of bias is due to measurement error in the explanatory variable(s) (in the randomized example this could correspond to inaccurate measurement of the dosage received, though it could also correspond to other explanatory variable). This problem is sometimes referred to as *errors-in-variables* and is discussed in detail in [1]. Many other types of sensitivity analysis are possible (for example, with respect to prior distributions in a Bayesian analysis) (see **Bayesian Statistics**), but we consider confounding, measurement error, and selection bias only. For more discussion of these topics in an epidemiological context, see [4, Chapter 19].

A general approach to sensitivity analyses is to first write down a plausible model for the response in terms of accurately measured explanatory variables (some of which may be unobserved), and with respect to a particular selection model. One may then derive the induced form of the model, in terms of observed variables and the selection mechanism assumed in the analysis. The parameters of the derived model can then be compared with the parameters of interest in the ‘true’ model, to reveal the extent of bias. We follow this approach, but note that it should be pursued only when the sample size in the original study is large, so that sampling variability is negligible; references in the discussion consider more general situations.

In the following, we assume that data are not available to control for bias. So in the next section, we consider the potential effects of *unmeasured* confounding. In the errors-in-variables context, we assume that we do observe ‘gold standard’ data in which a subset of individuals provides an accurate measure of the explanatory variable, along with the inaccurate measure. Similarly, with respect to selection bias, we assume that the sampling probabilities for study individuals are unknown and cannot be controlled for (as can be done in matched case-control studies, see [4, Chapter 16] for example), or that supplementary data on the selection probabilities of individuals are not available, as in two-phase methods (e.g., [6]); in both of these examples, the selection mechanism is known from the design (and would lead to bias if ignored, since the analysis must respect the sampling scheme).

Sensitivity to Unmeasured Confounding

Let Y denote a univariate response and X a univariate explanatory variable, and suppose that we are interested in the association between Y and X , but Y also potentially depends on U , an unmeasured variable. The discussion in [2] provided an early and clear account of the sensitivity of an observed association to unmeasured confounding, in the context of lung cancer and smoking. For simplicity, we assume that the ‘true’ model is linear and given by

$$E[Y|X, U] = \alpha^* + X\beta^* + U\gamma^*. \quad (1)$$

Further assume that the linear association between U and X is $E[X|U] = a + bX$. Roughly speaking, a variable U is a *confounder* if it is associated with both the response, Y , and the explanatory variable, X , but is not caused by Y or on the causal pathway between X and Y . For a more precise definition of confounding, and an extended discussion, see [4, Chapter 8]. We wish to derive the implied linear association between Y and X , since these are the variables that are observed. We use iterated expectation to average over the unmeasured variable, given X :

$$\begin{aligned} E[Y|X] &= E_{U|X} \{E[Y|X, U]\} \\ &= E_{U|X} \{\alpha^* + X\beta^* + U\gamma^*\} \end{aligned}$$

2 Sensitivity Analysis

$$\begin{aligned} &= \alpha^* + X\beta^* + E[U|X]\gamma^* \\ &= \alpha^* + X\beta^* + (a + bX)\gamma^* = \alpha + X\beta, \end{aligned}$$

where $\alpha = \alpha^* + \gamma^*a$ and, of more interest,

$$\beta = \beta^* + \gamma^*b. \quad (2)$$

Here the ‘true’ association parameter, β^* , that we would like to estimate, is represented with a * superscript, while the association parameter that we can estimate, β , does not have a superscript. Equation (2) shows that the bias $\beta - \beta^*$ is a function of the level of association between X and U (via the parameter b), and the association between Y and U (via γ^*). Equation (2) can be used to assess the effects of an unmeasured confounding using plausible values of b and γ^* , as we now demonstrate through a simple example.

Example Consider a study in which we wish to estimate the association between the rate of oral cancer and alcohol intake in men over 60 years of age. Let Y represent the natural logarithm of the rate of oral cancer, and let us suppose that we have a two-level alcohol variable X with $X = 0$ corresponding to zero intake and $X = 1$ to nonzero intake. A regression of Y on X gives an estimate $\hat{\beta} = 1.20$, so that the rate of oral cancer is $e^{1.20} = 3.32$ higher in the $X = 1$ population when compared to the $X = 0$ population.

The rate of oral cancer also increases with tobacco consumption (which we suppose is unmeasured in our study), however, and the latter is also positively associated with alcohol intake. We let $U = 0/1$ represent no tobacco/tobacco consumption. Since U is a binary variable, $E[U|X] = P(U = 1|X)$. Suppose that the probability of tobacco consumption is 0.05 and 0.45 in those with zero and nonzero alcohol consumption respectively; that is $P(U = 1|X = 1) = 0.05$ and $P(U = 1|X = 0) = 0.45$, so that $a = 0.05$ and $a + b = 0.45$, to give $b = 0.40$. Suppose further, that the log rate of oral cancer increases by $\gamma^* = \log 2.0 = 0.693$ for those who use tobacco (in both alcohol groups). Under these circumstances, from (2), the true association is

$$\hat{\beta}^* = \hat{\beta} - \gamma^*b = 1.20 - 0.693 \times 0.40 = 0.92, \quad (3)$$

so that the increase in the rate associated with alcohol intake is reduced from 3.32 to $\exp(0.92) = 2.51$.

In a real application, the sensitivity of the association would be explored with respect to a range of values of b and γ^* .

Sensitivity to Measurement Errors

In a similar way, we may examine the potential effects of measurement errors in the regressor X . As an example, consider a simple linear regression and suppose the true model is

$$Y = E[Y|X] + \epsilon^* = \alpha^* + \beta^*X + \epsilon^* \quad (4)$$

where $E[\epsilon^*] = 0$, $\text{var}(\epsilon^*) = \sigma_{\epsilon^*}^2$. Rather than measure X , we measure a surrogate W where

$$W = X + \delta \quad (5)$$

with $E[\delta] = 0$, $\text{var}(\delta) = \sigma_{\delta}^2$, and $\text{cov}(\delta, \epsilon^*) = 0$. The least squares estimator of β^* in model (4), from a sample (X_i, Y_i) , $i = 1, \dots, n$, has the form

$$\hat{\beta}^* = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}. \quad (6)$$

In the measurement error situation we fit the model

$$Y = E[Y|W] + \epsilon = \alpha + \beta W + \epsilon \quad (7)$$

where $E[\epsilon] = 0$, $\text{var}(\epsilon) = \sigma_{\epsilon}^2$. The least squares estimator of β in model (7), from a sample (W_i, Y_i) , $i = 1, \dots, n$, has the form

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W})^2} = \frac{\text{cov}(W, Y)}{\text{var}(W)}, \quad (8)$$

and to assess the extent of bias, we need to compare $E[\hat{\beta}]$ with β^* . From (8) we have

$$\begin{aligned}
\hat{\beta} &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i + \delta_i - \bar{X} - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i + \delta_i - \bar{X} - \bar{\delta})^2} \\
&= \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + \frac{1}{n} \sum_{i=1}^n (\delta_i - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{2}{n} \sum_{i=1}^n (X_i - \bar{X})(\delta_i - \bar{\delta}) + \frac{1}{n} \sum_{i=1}^n (\delta_i - \bar{\delta})^2} \\
&= \frac{\text{cov}(X, Y) + \text{cov}(\delta, Y)}{\text{var}(X) + 2\text{cov}(X, \delta) + \text{var}(\delta)}. \tag{9}
\end{aligned}$$

Under the assumptions of our model, $\text{cov}(\delta, Y) = 0$ and $\text{cov}(X, \delta) = 0$ and so, from (6)

$$\begin{aligned}
E[\hat{\beta}] &\approx \frac{\text{cov}(X, Y)}{\text{var}(X) + \text{var}(\delta)} \\
&= \frac{\text{cov}(X, Y)/\text{var}(X)}{1 + \text{var}(\delta)/\text{var}(X)} = \beta^* r
\end{aligned}$$

where the *attenuation factor*

$$r = \frac{\text{var}(X)}{\text{var}(X) + \sigma_\delta^2} \tag{10}$$

describes the amount of bias by which the estimate is attenuated toward zero. Note that with no measurement error ($\sigma_\delta^2 = 0$), $r = 1$, and no bias results, and also that the attenuation will be smaller in a well-designed study in which a large range of X is available. Hence, to carry out a sensitivity analysis, we can examine, for an observed estimate $\hat{\beta}$, the increase in the true coefficient for different values of σ_δ^2 via

$$\hat{\beta}^* = \frac{\hat{\beta}}{r}. \tag{11}$$

It is important to emphasize that the above derivation was based on a number of strong assumptions such as independence between errors in Y and in W , and constant variance for errors in both Y and W . Care is required in more complex situations, including those in which we have more than one explanatory variable. For example, if we regress Y on both X (which is measured without error), and a second explanatory variable is measured with error, then we will see bias in our estimator of the coefficient

associated with X , if there is a nonzero correlation between X and the second variable (see [1] for more details).

Example Let Y represent systolic blood pressure (in mmHg) and X sodium intake (in mmol/day), and suppose that a linear regression of Y on W produces an estimate of $\hat{\beta} = 0.1$ mmHg, so that an increase in daily sodium of 100 mmol/day is associated with an increase in blood pressure of 10 mmHg. Suppose also that $\text{var}(X) = 4$ mmHg. Table 1 shows the sensitivity of the coefficient associated with X , β^* , to different levels of measurement error; as expected, the estimate increases with increasing measurement error.

Sensitivity to Selection Bias

This section concerns the assessment of the bias that is induced when the probability of observing the data of a particular individual depends on the data of that individual. We consider a slightly different scenario to those considered in the last two sections, and assume we have a binary outcome variable, Y , and a

Table 1 The effect of measurement error when $\text{var}(X) = 4$

Measurement error σ_δ^2	Attenuation factor r	True estimate $\hat{\beta}^*$
0	0	0.1
1	0.8	0.125
2	0.67	0.15
4	0.5	0.2

4 Sensitivity Analysis

binary exposure, X , and let $p_x^* = P(Y = 1|X = x)$, $x = 0, 1$, be the ‘true’ probability of a $Y = 1$ outcome given exposure x , $x = 0, 1$. We take as parameter of interest the **odds ratio**:

$$\begin{aligned} \text{OR}^* &= \frac{P(Y = 1|X = 1)/P(Y = 0|X = 1)}{P(Y = 1|X = 0)/P(Y = 0|X = 0)} \\ &= \frac{p_1^*/(1 - p_1^*)}{p_0^*/(1 - p_0^*)}, \end{aligned} \quad (12)$$

which is the ratio of the odds of a $Y = 1$ outcome given exposed ($X = 1$), to the odds of such an outcome given unexposed ($X = 0$).

We now consider the situation in which we do not have constant probabilities of responding (being observed) across the population of individuals under study, and let $R = 0/1$ correspond to the event nonresponse/response, with response probabilities:

$$P(R = 1|X = x, Y = y) = q_{xy}, \quad (13)$$

for $x = 0, 1$, $y = 0, 1$; and we assume that we do not know these response rates. We do observe estimates

$$s = \frac{P(R = 1|X = 1, Y = 1)/P(R = 1|X = 0, Y = 1)}{P(R = 1|X = 1, Y = 0)/P(R = 1|X = 0, Y = 0)} = \frac{q_{11}/q_{01}}{q_{10}/q_{00}}. \quad (16)$$

of $p_x = P(Y = 1|X = x, R = 1)$, the probability of a $Y = 1$ outcome given both values of x and *given* response. The estimate of the odds ratio for the *observed* responders is then given by:

$$\text{OR} = \frac{P(Y = 1|X = 1, R = 1)/P(Y = 0|X = 1, R = 1)}{P(Y = 1|X = 0, R = 1)/P(Y = 0|X = 0, R = 1)} = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)}. \quad (14)$$

To link the two odds ratios we use Bayes theorem on each of the terms in (14) to give:

where the selection factor s is determined by the probabilities of response in each of the exposure-outcome groups. It is of interest to examine situations in which $s = 1$ and there is no bias. One such situation is when $q_{xy} = u_x \times v_y$, $x = 0, 1$; $y = 0, 1$, so that there is ‘no multiplicative interaction’ between exposure and outcome in the response model. Note that u_x and v_y are *not* the marginal response probabilities for, respectively, exposure, and outcome.

Example Consider a study carried out to examine the association between childhood asthma and maternal smoking. Let $Y = 0/1$ represent absence/presence of asthma in a child and $X = 0/1$ represent nonexposure/exposure to maternal smoking. Suppose a questionnaire is sent to parents to determine whether their child has asthma and whether the mother smokes. An odds ratio of $\widehat{\text{OR}} = 2$ is observed from the data of the responders, indicating that the odds of asthma is doubled if the mother smokes.

To carry out a sensitivity analysis, there are a number of ways to proceed. We write

Suppose that amongst noncases, the response rate in the exposed group is q times that in the unexposed group (that is $q_{10}/q_{00} = q$), while amongst the cases,

the response rate in the exposed group is $0.8q$ times that in the unexposed group (i.e., $q_{11}/q_{01} = 0.8q$). In

$$\begin{aligned} \text{OR} &= \frac{\frac{P(R = 1|X = 1, Y = 1)P(Y = 1|X = 1)}{P(R = 1|X = 1)}}{\frac{P(R = 1|X = 0, Y = 1)P(Y = 1|X = 0)}{P(R = 1|X = 0)}} \bigg/ \frac{\frac{P(R = 1|X = 1, Y = 0)P(Y = 0|X = 1)}{P(R = 1|X = 1)}}{\frac{P(R = 1|X = 0, Y = 0)P(Y = 0|X = 0)}{P(R = 1|X = 0)}} \\ &= \frac{p_1^*/(1 - p_1^*)}{p_0^*/(1 - p_0^*)} \times \frac{q_{11}q_{00}}{q_{10}q_{01}} = \text{OR}^* \times s, \end{aligned} \quad (15)$$

this scenario, $s = 0.8$ and

$$\widehat{\text{OR}}^* = \frac{\widehat{\text{OR}}}{0.8} = \frac{2}{0.8} = 2.5, \quad (17)$$

and we have underestimation because exposed cases were underrepresented in the original sample.

Discussion

In this article we have considered sensitivity analyses in a number of very simple scenarios. An extension would be to simultaneously consider the combined sensitivity to multiple sources of bias. We have also considered the sensitivity of point estimates only, and have not considered hypothesis testing or interval estimation. A comprehensive treatment of observational studies and, in particular, the sensitivity to various forms of bias may be found in [3]. The above derivations can be extended to various different modeling scenarios, for example, [5] examines sensitivity to unmeasured confounding in the context of Poisson regression in spatial epidemiology.

References

- [1] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement in Nonlinear Models*, Chapman & Hall/CRC Press, London.
- [2] Cornfield, J., Haenszel, W., Hammond, E.C., Lillienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.
- [3] Rosenbaum, P.R. (2002). *Observational Studies, Second Studies*, Springer-Verlag, New York.
- [4] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott-Raven, Philadelphia.
- [5] Wakefield, J.C. (2003). Sensitivity analyses for ecological regression, *Biometrics* **59**, 9–17.
- [6] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

(See also **Clinical Trials and Intervention Studies**)

JON WAKEFIELD

Sensitivity Analysis in Observational Studies

PAUL R. ROSENBAUM

Volume 4, pp. 1809–1814

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sensitivity Analysis in Observational Studies

Randomization Inference and Sensitivity Analysis

Randomized Experiments and Observational Studies

In a randomized experiment (*see* **Randomization**), subjects are assigned to treatment or control groups at random, perhaps by the flip of a coin or perhaps using random numbers generated by a computer [7]. Random assignment is the norm in **clinical trials** of treatments intended to benefit human subjects [21, 22]. Intuitively, randomization is an equitable way to construct treated and control groups, conferring no advantage to either group. At baseline before treatment in a randomized experiment, the groups differ only by chance, by the flip of the coin that assigned one subject to treatment, another to control. Therefore, comparing treated and control groups after treatment in a randomized experiment, if a common statistical test rejects the hypothesis that the difference in outcomes is due to chance, then a treatment effect is demonstrated. In contrast, in an observational study, subjects are not randomly assigned to groups, and outcomes may differ in treated and control groups for reasons other than effects caused by the treatment. Observational studies are the norm when treatments are harmful, unwanted, or simply beyond control by the investigator.

In the absence of random assignment, treated and control groups may not be comparable at baseline before treatment. Baseline differences that have been accurately measured in observed covariates can often be removed by **matching**, **stratification** or model based adjustments [2, 28, 29]. However, there is usually the concern that some important baseline differences were not measured, so that individuals who appear comparable may not be. A sensitivity analysis in an observational study addresses this possibility: it asks what the unmeasured covariate would have to be like to alter the conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias: some are sensitive to very small biases, while others are insensitive to quit large biases.

The First Sensitivity Analysis

The first **sensitivity analysis** in an observational study was conducted by Cornfield, et al. [6] for certain observational studies of cigarette smoking as a cause of lung cancer; see also [10]. Although the tobacco industry and others had often suggested that cigarettes might not be the cause of high rates of lung cancer among smokers, that some other difference between smokers and nonsmokers might be the cause, Cornfield, et al. found that such an unobserved characteristic would need to be a near perfect predictor of lung cancer and about nine times more common among smokers than among nonsmokers. While this sensitivity analysis does not rule out the possibility that such a characteristic might exist, it does clarify what a scientist must logically be prepared to assert in order to defend such a claim.

Methods of Sensitivity Analysis

Various methods of sensitivity analysis exist. The method of Cornfield, et al. [6] is perhaps the best known of these, but it is confined to binary responses; moreover, it ignores sampling variability, which is hazardous except in very large studies. A method of sensitivity analysis that is similar in spirit to the method of Cornfield et al. will be described here; however, this alternative method takes account of sampling variability and is applicable to any kind of response; see, for instance, [25, 26, 29], and Section 4 of [28] for detailed discussion. Alternative methods of sensitivity analysis are described in [1, 5, 8, 9, 14, 18, 19, 23, 24], and [33].

The sensitivity analysis imagines that in the population before matching or stratification, subjects are assigned to treatment or control independently with unknown probabilities. Specifically, two subjects who look the same at baseline before treatment – that is, two subjects with the same observed covariates – may nonetheless differ in terms of unobserved covariates, so that one subject has an odds of treatment that is up to $\Gamma \geq 1$ times greater than the odds for another subject. In the simplest randomized experiment, everyone has the same chance of receiving the treatment, so $\Gamma = 1$. If $\Gamma = 2$ in an observational study, one subject might be twice as likely as another to receive the treatment because of unobserved pre-treatment differences. The sensitivity analysis asks how much hidden bias can be present – that is, how

large can Γ be – before the qualitative conclusions of the study begin to change. A study is highly sensitive to hidden bias if the conclusions change for Γ just barely larger than 1, and it is insensitive if the conclusions change only for quite large values of Γ .

If $\Gamma > 1$, the treatment assignments probabilities are unknown, but unknown only to a finite degree measured by Γ . For each fixed $\Gamma \geq 1$, the sensitivity analysis computes bounds on inference quantities, such as P values or **confidence intervals**. For $\Gamma = 1$, one obtains a single P value, namely the P value for a randomized experiment [7, 16, 17]. For each $\Gamma > 1$, one obtains not a single P value, but rather an interval of P values reflecting uncertainty due to hidden bias. As Γ increases, this interval becomes longer, and eventually it becomes uninformative, including both large and small P values. The point, Γ , at which the interval becomes uninformative is a measure of sensitivity to hidden bias. Computations are briefly described in Section titled ‘Sensitivity Analysis Computations’ and an example is discussed in detail in Section titled ‘Sensitivity Analysis: Example’.

Sensitivity Analysis Computations

The straightforward computations involved in a sensitivity analysis will be indicated briefly in the case of one standard test, namely Wilcoxon’s signed rank test for matched pairs (*see Distribution-free Inference, an Overview*) [17]. For details in this case [25] and many others, see Section 4 of [28]. The null hypothesis asserts that the treatment is without effect, that each subject would have the same response under the alternative treatment. There are S pairs, $s = 1, \dots, S$ of two subjects, one treated, one control, matched for observed covariates. The distribution of treatment assignments within pairs is simply the conditional distribution for the model in Section titled ‘Methods of Sensitivity Analysis’ given that each pair includes one treated subject and one control. Each pair yields a treated-minus-control difference in outcomes, say D_s . For brevity in the discussion here, the D_s will be assumed to be untied, but ties are not a problem, requiring only slight change to formulas. The absolute differences, $|D_s|$, are ranked from 1 to S , and Wilcoxon’s signed rank statistic, W , is the sum of the ranks of the positive differences, $D_s > 0$.

For the signed rank statistic, the elementary computations for a sensitivity analysis closely parallel the elementary computations for a conventional

analysis. This paragraph illustrates the computations and may be skipped. In a moderately large randomized experiment, under the null hypothesis of no effect, W is approximately normally distributed with expectation $S(S+1)/4$ and variance $S(S+1)(2S+1)/24$; see Chapter 3 of [17]. If one observed $W = 300$ with $S = 25$ pairs in a randomized experiment, one would compute $S(S+1)/4 = 162.5$ and $S(S+1)(2S+1)/24 = 1381.25$, and the deviate $Z = (300 - 162.5)/\sqrt{1381.25} = 3.70$ would be compared to a Normal distribution to yield a one-sided P value of 0.0001. In a moderately large observational study, under the null hypothesis of no effect, the distribution of W is approximately bounded between two Normal distributions, with expectations $\mu_{\max} = \lambda S(S+1)/2$ and $\mu_{\min} = (1-\lambda) S(S+1)/2$, and the same variance $\sigma^2 = \lambda(1-\lambda) S(S+1)(2S+1)/6$, where $\lambda = \Gamma/(1+\Gamma)$. Notice that if $\Gamma = 1$, these expressions are the same as in the randomized experiment. For $\Gamma = 2$ and $W = 300$ with $S = 25$ pairs, one computes $\lambda = 2/(1+2) = 2/3$, $\mu_{\max} = (2/3) 25(25+1)/2 = 216.67$, $\mu_{\min} = (1/3) 25(25+1)/2 = 108.33$, and $\sigma^2 = (2/3)(1/3) 25(25+1)(2 \times 25+1)/6 = 1227.78$; then two deviates are calculated, $Z_1 = (300 - 108.33)/\sqrt{1227.78} = 5.47$ and $Z_2 = (300 - 216.67)/\sqrt{1227.78} = 2.38$, which are compared to a Normal distribution, yielding a range of P values from 0.00000002 to 0.009. In other words, a bias of magnitude $\Gamma = 2$ creates some uncertainty about the correct P value, but it would leave no doubt that the difference is significant at the conventional 0.05 level.

Just as W has an exact randomization distribution useful for small S , so too there are exact sensitivity bounds. See [31] for details including S-Plus code.

Sensitivity Analysis: Example

A Matched Observational Study of an Occupational Hazard

Studies of occupational health usually focus on workers, but Morton, Saah, Silberg, Owens, Roberts and Saah [20] were worried about the workers’ children. Specifically, they were concerned that workers exposed to lead might bring lead home in clothes and hair, thereby exposing their children as well. They matched 33 children whose fathers worked in a battery factory to 33 unexposed control children of the

Table 1 Blood lead levels, in micrograms of lead per decaliter of blood, of exposed children whose fathers worked in a battery factory and age-matched control children from the neighborhood. Exposed father's lead exposure at work (high, medium, low) and hygiene upon leaving the factory (poor, moderate, good) are also given. Adapted for illustration from Tables 1, 2 and 3 of Morton, et al. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555. [20]

Pair s	Exposure	Hygiene	Exposed child's Lead level $\mu\text{g/dl}$	Control child's Lead level $\mu\text{g/dl}$	Dose Score
1	high	good	14	13	1.0
2	high	moderate	41	18	1.5
3	high	poor	43	11	2.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
33	low	poor	10	13	1.0
Median			34	16	

same age and neighborhood, and used Wilcoxon's signed rank test to compare the level of lead found in the children's blood, measured in μg of lead per decaliter (dl) of blood. They also measured the father's level of exposure to lead at the factory, classified as high, medium, or low, and the father's hygiene upon leaving the factory, classified as poor, moderate, or good. Table 1 is adapted for illustration from Tables 1, 2, and 3 of Morton, et al. (1982) [20]. The median lead level for children of exposed fathers was more than twice that of control children, $34 \mu\text{g/dl}$ versus $16 \mu\text{g/dl}$.

If Wilcoxon's signed rank test W is applied to the exposed-minus-control differences in Table 1, then the difference is highly significant in a one-sided test, $P < 0.0001$. This significance level would be appropriate in a randomized experiment, in which children were picked at random for lead exposure. Table 2 presents the sensitivity analysis, computed as in Section titled 'Sensitivity Analysis Computations'. Table 2 gives the range of possible one-sided significance levels for several possible magnitudes of hidden bias, measured by Γ . Even if the matching of exposed and control children had failed to control an unobserved characteristic strongly related to blood lead levels and $\Gamma = 4.25$ times more common among exposed children, this would still not explain the higher lead levels found among exposed children.

Where Table 2 focused on significance levels, Table 3 considers the one sided 95% confidence interval, $[\hat{\tau}_{\text{low}}, \infty)$, for an additive effect obtained by inverting the signed rank test [28]. If the data in Table 1 had come from a randomized experiment

Table 2 Sensitivity analysis for one-sided significance levels in the lead data. For unobserved biases of various magnitudes, the table gives the range of possible significance levels

Γ	min	max
1	<0.0001	<0.0001
2	<0.0001	0.0018
3	<0.0001	0.0136
4	<0.0001	0.0388
4.25	<0.0001	0.0468
5	<0.0001	0.0740

Table 3 Sensitivity analysis for one-sided confidence intervals for an additive effect in the lead data. For unobserved biases of various magnitudes, the table gives smallest possible endpoint for the one-sided confidence interval

Γ	min $\hat{\tau}_{\text{low}}$
1	10.5
2	5.5
3	2.5
4	0.5
4.25	0.0
5	-1.0

($\Gamma = 1$) with an additive treatment effect τ , then we would be 95% confident that father's lead exposure had increased his child's lead level by $\hat{\tau}_{\text{low}} = 10.5 \mu\text{g/dl}$ [17]. In an observational study with $\Gamma > 1$, there is a range of possible endpoints for the 95% confidence interval, and Table 3 reports the smallest value in the range. Even if $\Gamma = 3$, we would

4 Sensitivity Analysis in Observational Studies

Table 4 Sensitivity to hidden bias in four observational studies. The randomization test assuming no hidden bias is highly significant in all four studies, but the magnitude of hidden bias that could alter this conclusion varies markedly between the four studies

Treatment	$\Gamma = 1$	(Γ , max P value)
Smoking/Lung Cancer Hammond [11]	<0.0001	(5, 0.03)
Diethylstilbestrol/ vaginal cancer Herbst, et al. [12]	<0.0001	(7, 0.054)
Lead/Blood lead Morton, et al. [20]	<0.0001	(4.25, 0.047)
Coffee/MI Jick, et al. [15]	0.0038	(1.3, 0.056)

be 95% confident exposure increased lead levels by 2.5 $\mu\text{g}/\text{dl}$.

Studies Vary in Their Sensitivity to Hidden Bias

Studies vary markedly in their sensitivity to hidden bias. As an illustration, Table 4 compares the sensitivity of four studies, a study of smoking as a cause of lung cancer [11], a study of prenatal exposure to diethylstilbestrol as a cause of vaginal cancer [12], the lead exposure study [20], and a study of coffee as a cause of myocardial infarction [15].

If no effect is tested using a conventional test appropriate for a randomized experiment ($\Gamma = 1$), the results are highly significant in all four studies. The last column of Table 4 indicates sensitivity to hidden bias, quoting the magnitude of hidden bias $\Gamma \geq 1$ needed to produce an upper bound on the P value close to the conventional 0.05 level. The study [12] of the effects of diethylstilbestrol becomes sensitive at about $\Gamma = 7$, while the study [15] of the effects of coffee becomes sensitive at $\Gamma = 1.3$. A small bias could explain away the effects of coffee, but only an enormous bias could explain away the effects of diethylstilbestrol. The lead exposure study, although quite insensitive to hidden bias, is about halfway between these two other studies, and is slightly more sensitive to hidden bias than the study of the effects of smoking.

Reducing Sensitivity to Hidden Bias

Accurately predicting a highly specific pattern of associations between treatment and response is often

said to strengthen the evidence that the effects of the treatment caused the association. For instance, Cook, Campbell, and Peracchio [3] write: ‘Conclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.’ Hill [13] and Weiss [34] emphasized the role of a dose response relationship. Cook and Shadish [4] write: ‘Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations.’

Does successful prediction of a complex pattern of associations affect sensitivity to hidden bias? It may, or it may not, and the degree to which it has done so can be appraised using methods similar to those in Section titled ‘Sensitivity Analysis Computations’. See [27] and [30] for methods of analysis, and [32] for issues in research design. The issues will be illustrated using the example in Table 1.

Recall that exposed fathers were classified by their degree of exposure and their hygiene upon leaving the factory. If the fathers’ exposure to lead at work were the cause of the higher lead levels among exposed children, then one would expect more lead in the blood of children whose fathers had higher exposure and poorer hygiene. Here, exposed children are divided into three groups of roughly similar size. The 13 exposed children in the category (high exposure, poor hygiene) were assigned a score of 2.0. Low exposure with any hygiene was assigned a score of 1, as was good hygiene with any exposure, and there were 12 such exposed children. The remaining 8 exposed children in intermediate situations were assigned a score of 1.5; they had either high exposure with moderate hygiene or medium exposure with poor hygiene. (None of the 33 matched children had medium exposure with moderate hygiene, although one unmatched child not used here fell into this category.)

The coherent or dose signed rank statistic D gives greater weight to matched pairs with higher doses [27, 30]. Table 5 compares the sensitivity to hidden bias of the usual Wilcoxon signed rank test W , which ignores doses, to the sensitivity of the coherent signed rank statistic. In particular, Table 5 gives the upper bound on the one-sided significance level for testing no effect. For W , this is the same computation as in Table 2. In fact, the coherent pattern of associations is somewhat less sensitive to hidden bias in this example: the upper bound on the

Table 5 Coherent patterns of associations can reduce sensitivity to hidden bias. Upper bounds on one-sided significance levels in the lead data, ignoring and using dose information

Γ	Wilcoxon W	Coherent D
1	<0.0001	<0.0001
3	0.0136	0.0119
4.35	0.0502	0.0398
4.75	0.0645	0.0503

P value for W ignoring doses is just above 0.05 at $\Gamma = 4.35$, but using doses with D the corresponding value is $\Gamma = 4.75$.

Exposed children had higher lead levels than unexposed controls, and also exposed children with higher exposures had higher lead levels than exposed children with lower lead levels. A larger hidden bias is required to explain this pattern of associations than is required to explain the difference between exposed and control children. In short, accurate prediction of a pattern of associations may reduce sensitivity to hidden bias, and whether this has happened, and the degree to which it has happened, may be appraised by a sensitivity analysis.

Acknowledgment

This work was supported by grant SES-0345113 from the US National Science Foundation.

References

- [1] Berk, R.A. & De Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design, *Journal of the American Statistical Association* **94**, 1045–1052.
- [2] Cochran, W.G. (1965). The planning of observational studies of human populations (with Discussion), *Journal of the Royal Statistical Society, Series A* **128**, 134–155.
- [3] Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi-experimentation, in *Handbook of Industrial and Organizational Psychology*, M. Dunnette & L. Hough, eds, Consulting Psychologists Press, Palo Alto, pp. 491–576.
- [4] Cook, T.D. & Shadish, W.R. (1994). Social experiments: some developments over the past fifteen years, *Annual Review of Psychology* **45**, 545–580.
- [5] Copas, J.B. & Li, H.G. (1997). Inference for non-random samples (with discussion), *Journal of the Royal Statistical Society B* **59**, 55–96.
- [6] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.
- [7] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [8] Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.
- [9] Gastwirth, J.L., Krieger, A.M. & Rosenbaum, P.R. (1998b). Dual and simultaneous sensitivity analysis for matched pairs, *Biometrika* **85**, 907–920.
- [10] Greenhouse, S. (1982). Jerome Cornfield's contributions to epidemiology, *Supplement of Biometrics* **38**, 33–45.
- [11] Hammond, E.C. (1964). Smoking in relation to mortality and morbidity: findings in first thirty-four months of follow-up in a prospective study started in 1959, *Journal of the National Cancer Institute* **32**, 1161–1188.
- [12] Herbst, A., Ulfelder, H. & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women, *New England Journal of Medicine* **284**, 878–881.
- [13] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.
- [14] Imbens, G.W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* **93**, 126–132.
- [15] Jick, H., Miettinen, O., Neff, R., Jick, H., Miettinen, O.S., Neff, R.K., Shapiro, S., Heinonen, O.P., Slone, D. (1973). Coffee and myocardial infarction, *New England Journal of Medicine*, **289**, 63–77.
- [16] Kempthorne, O. (1952). *Design and Analysis of Experiments*, John Wiley & Sons, New York.
- [17] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, Prentice Hall, Upper Saddle River.
- [18] Lin, D.Y., Psaty, B.M. & Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics* **54**, 948–963.
- [19] Manski, C. (1995). *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge.
- [20] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555.
- [21] Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I, *British Journal of Cancer* **34**, 585–612.
- [22] Piantadosi, S. (1997). *Clinical Trials*, Wiley, New York.
- [23] Robins, J.M., Rotnitzky, A. & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured

- confounding in missing data and causal inference models, in *Statistical Models in Epidemiology*, E. Halloran & D. Berry, eds, Springer, New York, pp. 1–94.
- [24] Rosenbaum, P.R. (1986). Dropping out of high school in the United States: an observational study, *Journal of Educational Statistics* **11**, 207–224.
- [25] Rosenbaum, P.R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies, *Biometrika* **74**, 13–26.
- [26] Rosenbaum, P.R. (1995). Quantiles in nonrandom samples and observational studies, *Journal of the American Statistical Association* **90**, 1424–1431.
- [27] Rosenbaum, P.R. (1997). Signed rank statistics for coherent predictions, *Biometrics* **53**, 556–566.
- [28] Rosenbaum, P.R. (2002a). *Observational Studies*, Springer, New York.
- [29] Rosenbaum, P.R. (2002b). Covariance adjustment in randomized experiments and observational studies (with Discussion), *Statistical Science* **17**, 286–327.
- [30] Rosenbaum, P.R. (2003a). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1–10.
- [31] Rosenbaum, P.R. (2003b). Exact confidence intervals for nonconstant effects by inverting the signed rank test, *American Statistician* **57**, 132–138.
- [32] Rosenbaum, P.R. (2004). Design sensitivity in observational studies, *Biometrika* **91**, 153–164.
- [33] Rosenbaum, P.R. & Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society Series B* **45**, 212–218.
- [34] Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of ‘dose-response.’, *American Journal of Epidemiology* **113**, 487–90.

PAUL R. ROSENBAUM

Sequential Decision Making

HANS J. VOS

Volume 4, pp. 1815–1819

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sequential Decision Making

Well-known examples of fixed-length mastery tests in the behavioral sciences include pass/fail decisions in education, certification, and successfulness of therapies. The fixed-length mastery problem has been studied extensively in the literature within the framework of (empirical) Bayesian decision theory [9] (*see Bayesian Item Response Theory Estimation*). In this approach, the following two basic elements are distinguished: A measurement model relating the probability of a correct response to student's (unknown) true level of functioning, and a loss structure evaluating the total costs and benefits for each possible combination of decision outcome and true level of functioning. Within the framework of Bayesian decision theory [2, 5], optimal rules (i.e., Bayes rules) are obtained by minimizing the posterior expected losses associated with all possible decision rules. Decision rules are hereby prescriptions specifying for each possible observed response pattern what action has to be taken. The Bayes principle assumes that prior knowledge about student's true level of functioning is available and can be characterized by a probability distribution called *the prior*. This prior probability represents our best prior beliefs concerning student's true level of functioning; that is, before any item yet has been administered.

The test at the end of the treatment does not necessarily have to be a fixed-length mastery test but might also be a sequential mastery test (SMT). In this case, in addition to the actions declaring mastery or nonmastery, also the action of to continue testing and administering another random item is available. Sequential mastery tests are designed with the goal of maximizing the probability of making correct classification decisions (i.e., mastery and nonmastery) while at the same time minimizing test length [6]. For instance, Ferguson [3] showed that average test lengths could be reduced by half without sacrificing classification accuracy.

The purpose of this entry is to derive optimal rules for SMT in the context of education using the framework of Bayesian sequential decision theory [2, 5]. The main advantage of this approach is that costs of testing (i.e., administering another random item) can be taken explicitly into account.

Bayesian Sequential Principle Applied to SMT

It is indicated in this section how the framework of Bayesian sequential decision theory, in combination with the binomial distribution for modeling response behavior (i.e., the measurement model) and adopting threshold loss for the loss function involved, is applied to SMT.

Framework of Bayesian Sequential Decision Theory

Three basic elements can be identified in Bayesian sequential decision theory. In addition to a measurement model and a loss function, costs of administering one additional item must be explicitly specified in this approach. Doing so, posterior expected losses corresponding to the mastery and nonmastery decisions can now be calculated straightforward at each stage of testing. As far as the posterior expected loss corresponding to continue testing concerns, this quantity is determined by averaging the posterior expected losses corresponding to each of the possible future classification outcomes relative to observing those outcomes (i.e., the posterior predictive distributions). Optimal rules (i.e., Bayesian sequential rules) are now obtained by minimizing the posterior expected losses associated with all possible decision rules at each stage of testing using techniques of backward induction (i.e., dynamic programming). This technique starts by considering the final stage of testing (where the option to continue testing is not available) and then works backward to the first stage of testing.

Notation

In order to classify within a reasonable period of time those students for whom the decision of declaring mastery or nonmastery is not as clear-cut, a sequential mastery test is supposed to have a maximum length of n ($n \geq 1$). Let the observed item response at each stage of testing k ($1 \leq k \leq n$) for a randomly sampled student be denoted by a discrete random variable X_k , with realization x_k . The observed response variables $X_1, \dots, X_k$ are assumed to be independent and identically distributed for each value of k , and take the values 0 and 1 for respectively incorrect and correct responses to item k . Furthermore, let the observed number-correct score after

2 Sequential Decision Making

k items have been administered be denoted by a discrete random variable $S_k = X_1 + \dots + X_k$, with realization $s_k = x_1 + \dots + x_k$ ($0 \leq s_k \leq k$).

Student's true level of functioning is unknown due to measurement and sampling error. All that is known is his/her observed number-correct score s_k . In other words, the mastery test is not a perfect indicator of student's true performance. Therefore, let student's (unknown) true level of functioning be denoted by a continuous random variable T on the latent proportion-correct metric, with realization t ($0 \leq t \leq 1$).

Finally, a criterion level t_c ($0 \leq t_c \leq 1$) on T must be specified in advance by the decision-maker using methods of standard-setting (e.g., [1]). A student is considered a true nonmaster and true master if his/her true level of functioning t is smaller or larger than t_c , respectively.

Threshold Loss and Costs of Testing

Generally speaking, as noted before, a loss function evaluates the total costs and benefits of all possible decision outcomes for a student whose true level of functioning is t . These costs may concern all relevant psychological, social, and economic consequences which the decision brings along. As in [6], here the well-known threshold loss function is adopted as the loss structure involved. The choice of this loss function implies that the 'seriousness' of all possible consequences of the decisions can be summarized by possibly different constants, one for each of the possible classification outcomes.

For the sequential mastery problem, following Lewis and Sheehan [6], a threshold loss function can be formulated as a natural extension of the one for the fixed-length mastery problem at each stage of testing k as shown in Table 1.

The value e represents the costs of administering one random item. For the sake of simplicity, these

costs are assumed to be equal for each classification outcome as well as for each testing occasion. Applying an admissible positive linear transformation [7], and assuming the losses l_{00} and l_{11} associated with the correct classification outcomes are equal and take the smallest values, the threshold loss function in Table 1 was rescaled in such a way that l_{00} and l_{11} were equal to zero. Hence, the losses l_{01} and l_{10} must take positive values.

The ratio l_{10}/l_{01} is denoted as the loss ratio R , and refers to the relative losses for declaring mastery to a student whose true level of functioning is below t_c (i.e., false positive) and declaring nonmastery to a student whose true level of functioning exceeds t_c (i.e., false negative).

The loss parameters l_{ij} ($i, j = 0, 1; i \neq j$) associated with the incorrect decisions have to be empirically assessed, for which several methods have been proposed in the literature. Most texts on decision theory, however, propose lottery methods [7] for assessing loss functions empirically. In general, the consequences of each pair of actions and true level of functioning are scaled in these methods by looking at the most and least preferred outcomes. But, in principle, any psychological scaling method can be used.

Measurement Model

Following Ferguson [3], in the present entry the well-known binomial model will be adopted for the probability that after k items have been administered, s_k of them have been answered correctly (see **Catalogue of Probability Density Functions**). Its distribution at stage k of testing for given student's true level of functioning t , $P(s_k|t)$, can be written as follows:

$$P(s_k|t) = \binom{k}{s_k} t^{s_k} (1-t)^{k-s_k}. \quad (1)$$

If each response is independent of the other, and if student's probability of a correct answer remains constant, the distribution function of s_k , given true level of functioning t , is given by (1). The binomial model assumes that the test given to each student is a random sample of items drawn from a large (real or imaginary) item pool [10]. Therefore, for each student a new random sample of items must be drawn in practical applications of the sequential mastery problem.

Table 1 Table for threshold loss function at stage k ($1 \leq k \leq n$) of testing

Decision	True level of functioning	
	$T \leq t_c$	$T > t_c$
Declaring nonmastery	ke	$l_{01} + ke$
Declaring mastery	$l_{10} + ke$	ke

Optimizing Rules for the Sequential Mastery Problem

In this section, it will be shown how optimal rules for SMT can be derived using the framework of Bayesian sequential decision theory. Doing so, given an observed item response vector $(x_1, \dots, x_k)$, first the Bayesian principle will be applied to the fixed-length mastery problem by determining which of the posterior expected losses associated with the two classification decisions is the smallest. Next, applying the Bayesian principle again, optimal rules for the sequential mastery problem are derived at each stage of testing k by comparing this quantity with the posterior expected loss associated with the option to continue testing.

Applying the Bayesian Principle to the Fixed-length Mastery Problem

As noted before, the Bayesian decision rule for the fixed-length mastery problem can be found by minimizing the posterior expected losses associated with the two classification decisions of declaring mastery or nonmastery. In doing so, the posterior expected loss is taken with respect to the posterior distribution of T . It can easily be verified from Table 1 and (1) that mastery is declared when the posterior expected loss corresponding to declaring mastery is smaller than the posterior expected loss corresponding to declaring nonmastery, or, equivalently, when s_k is such that

$$\begin{aligned} (l_{10} + ke)P(T \leq t_c | s_k) + (ke)P(T > t_c | s_k) \\ < (ke)P(T \leq t_c | s_k) + (l_{01} + ke)P(T > t_c | s_k), \end{aligned} \quad (2)$$

and that nonmastery is declared otherwise. Rearranging terms, it can easily be verified from (2) that mastery is declared when s_k is such that

$$P(T \leq t_c | s_k) < \frac{1}{1 + R}, \quad (3)$$

and that nonmastery is declared otherwise.

Assuming a beta prior for T , it follows from an application of Bayes' theorem (see **Bayesian Belief Networks**) that under the assumed binomial model from (1), the posterior distribution of T will be a member of the beta family again (the conjugacy

property, see, e.g., [5]). In fact, if the beta function $B(\alpha, \beta)$ with parameters α and β ($\alpha, \beta > 0$) is chosen as prior distribution (i.e., the natural conjugate of the binomial distribution) and student's observed number-correct score is s_k from a test of length k , then the posterior distribution of T is $B(\alpha + s_k, k - s_k + \beta)$. Hence, assuming a beta prior for T , it follows from (3) that mastery is declared when s_k is such that

$$B(\alpha + s_k, k - s_k + \beta) < \frac{1}{1 + R}, \quad (4)$$

and that nonmastery is declared otherwise.

The uniform distribution on the standard interval $[0,1]$ as a noninformative prior will be assumed in this entry, which results as a special case of the beta distribution $B(\alpha, \beta)$ for $\alpha = \beta = 1$ (see **Catalogue of Probability Density Functions**). In other words, prior true level of functioning can take on all values between 0 and 1 with equal probability. It then follows immediately from (4) that mastery is declared when s_k is such that

$$B(1 + s_k, k - s_k + 1) < \frac{1}{1 + R}, \quad (5)$$

and that nonmastery is declared otherwise. The beta distribution has been extensively tabulated [8]. Normal approximations are also available [4].

Derivation of Bayesian Sequential Rules

Let $d_k(x_1, \dots, x_k)$ denote the decision rule yielding the minimum of the posterior expected losses associated with the two classification decisions at stage k of testing, and let the posterior expected loss corresponding to this minimum be denoted as $V_k(x_1, \dots, x_k)$. Bayesian sequential rules can now be found by using the following backward induction computational scheme: First, the Bayesian sequential rule at the final stage n of testing is computed. Since the option to continue testing is not available at this stage of testing, it follows immediately that the Bayesian sequential rule is given by $d_n(x_1, \dots, x_n)$, and its corresponding posterior expected loss by $V_n(x_1, \dots, x_n)$.

To compute the posterior expected loss associated with the option to continue testing at stage $(n - 1)$ until stage 0, the risk $R_k(x_1, \dots, x_k)$ will be introduced at each stage k ($1 \leq k \leq n$) of testing. Let the

4 Sequential Decision Making

risk at stage n be defined as $V_n(x_1, \dots, x_n)$. Generally, given response pattern $(x_1, \dots, x_k)$, the risk at stage $(k-1)$ is then computed inductively as a function of the risk at stage k as:

$$R_{k-1}(x_1, \dots, x_{k-1}) = \min\{V_{k-1}(x_1, \dots, x_{k-1}), E[R_k(x_1, \dots, x_{k-1}, X_k)|x_1, \dots, x_{k-1}]\}. \quad (6)$$

The posterior expected loss corresponding to administering one more random item after $(k-1)$ items have been administered, $E[R_k(x_1, \dots, x_{k-1}, X_k)|x_1, \dots, x_{k-1}]$, can then be computed as the expected risk at stage k of testing as

$$E[R_k(x_1, \dots, x_{k-1}, X_k)|x_1, \dots, x_{k-1}] = \sum_{x_k=0}^{x_k=1} R_k(x_1, \dots, x_k) P(X_k = x_k|x_1, \dots, x_{k-1}), \quad (7)$$

where $P(X_k = x_k|x_1, \dots, x_{k-1})$ denotes the posterior predictive distribution of X_k at stage $(k-1)$ of testing. Computation of this conditional distribution is deferred until the next section. Note that (7) averages the posterior expected losses associated with each of the possible future classification outcomes with weights corresponding to the probabilities of observing those outcomes.

The Bayesian sequential rule at stage $(k-1)$ is now given by: Administer one more random item if $E[R_k(x_1, \dots, x_{k-1}, X_k)|x_1, \dots, x_{k-1}]$ is smaller than $V_{k-1}(x_1, \dots, x_{k-1})$; otherwise, decision $d_{k-1}(x_1, \dots, x_{k-1})$ is taken. The Bayesian sequential rule at stage 0 denotes the decision whether or not to administer at least one random item.

Computation of Posterior Predictive Distributions

As is clear from (7), the posterior predictive distribution $P(X_k = x_k|x_1, \dots, x_{k-1})$ is needed for computing the posterior expected loss corresponding to administering one more random item at stage $(k-1)$ of testing. Assuming the binomial distribution as measurement model and the uniform distribution $B(1,1)$ as prior, it was shown (e.g., [5]) that $P(X_k = 1|x_1, \dots, x_{k-1}) = (1 + s_{k-1})/(k+1)$, and, thus, that $P(X_k = 0|x_1, \dots, x_{k-1}) = [1 - (1 + s_{k-1})/(k+1)] = (k - s_{k-1})/(k+1)$.

Determination of Appropriate Action for Different Number-correct Score

Using the general backward induction scheme discussed earlier, for a given maximum number n ($n \geq 1$) of items to be administered, a program BAYES was developed to determine the appropriate action (i.e., nonmastery, continuation, or mastery) at each stage k of testing for different number-correct score s_k .

As an example, the appropriate action is depicted in Table 2 as a closed interval for a maximum of 20 items (i.e., $n = 20$). Students were considered as true masters if they knew at least 55% of the subject matter. Therefore t_c was fixed at 0.55. Furthermore, the loss corresponding to the false mastery decision was perceived twice as large as the loss corresponding to the false nonmastery decision (i.e., $R = 2$). On a scale in which one unit corresponded to the constant costs of administering one random item (i.e., $e = 1$), therefore, l_{10} and l_{01} were set equal to 200 and 100, respectively. These numerical values reflected the assumption that the losses corresponding to taking incorrect classification decisions were rather

Table 2 Appropriate action calculated by stage of testing and number-correct score

Stage of testing	Appropriate action by number-correct score		
	Nonmastery	Continuation	Mastery
0		0	
1		[0,1]	
2	0	[1,2]	
3	0	[1,3]	
4	[0,1]	[2,4]	
5	[0,1]	[2,4]	5
6	[0,2]	[3,5]	6
7	[0,2]	[3,5]	[6,7]
8	[0,3]	[4,6]	[7,8]
9	[0,4]	[5,7]	[8,9]
10	[0,4]	[5,7]	[8,10]
11	[0,5]	[6,8]	[9,11]
12	[0,5]	[6,8]	[9,12]
13	[0,6]	[7,9]	[10,13]
14	[0,7]	[8,9]	[10,14]
15	[0,7]	[8,10]	[11,15]
16	[0,8]	[9,10]	[11,16]
17	[0,9]	[10,11]	[12,17]
18	[0,10]	11	[12,18]
19	[0,11]		[12,19]
20	[0,12]		[13,20]

large relative to the costs of administering one random item.

As can be seen from Table 2, at least five random items need to be administered before mastery can be declared. However, in principle, nonmastery can be declared already after administering two random items. Also, generally a rather large number of items have to be answered correctly before mastery can be declared. This can be accounted for the relatively large losses corresponding to false positive decisions (i.e., 200) relative to the losses corresponding to false negative decisions (i.e., 100). In this way, relatively large posterior expected losses from taking false positive decisions can be avoided.

Discussion and Conclusions

In this entry, using the framework of Bayesian sequential decision theory, optimal rules for the sequential mastery problem (nonmastery, mastery, or to continue testing) in the context of education were derived. It should be emphasized, however, that the Bayes sequential principle is especially appropriate when costs of testing can be assumed to be quite large. For instance, when testlets (i.e., blocks of parallel items) rather than single items are considered. Also, the proposed strategy might be appropriate in the context of sequential testing problems in psychodiagnostics. Suppose that a new treatment (e.g., cognitive-analytic therapy) must be tested on patients suffering from some mental health problem (e.g., anorexia nervosa). Each time after having exposed a patient to the new treatment, it is desired to make a decision concerning the effectiveness/ineffectiveness of the new treatment or to continue testing and exposing the new treatment to another random patient suffering from the same

mental health problem. In such clinical situations, costs of testing generally are quite large and the Bayesian sequential approach might be considered as an alternative to fixed-length mastery tests.

References

- [1] Angoff, W.H. (1971). Scales, norms and equivalent scores, in *Educational Measurement*, 2nd Edition, R.L. Thorndike, ed., American Council on Education, Washington, pp. 508–600.
- [2] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.
- [3] Ferguson, R.L. (1969). *The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction*, Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.
- [4] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin, Boston.
- [5] Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, 3rd Edition, Macmillan Publishers, New York.
- [6] Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement* **14**, 367–386.
- [7] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions*, John Wiley & Sons, New York.
- [8] Pearson, K. (1930). *Tables for Statisticians and Biometricians*, Cambridge University Press, London.
- [9] van der Linden, W.J. (1990). Applications of decision theory to test-based decision making, in *New Developments in Testing: Theory and Applications*, R.K. Hambleton & J.N. Zaal, eds, Kluwer Academic Publishers, Boston, pp. 129–155.
- [10] Wilcox, R.R. (1981). A review of the beta-binomial model and its extensions, *Journal of Educational Statistics* **6**, 3–32.

HANS J. VOS

Sequential Testing

D.S. COAD

Volume 4, pp. 1819–1820

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sequential Testing

When a **clinical trial** is carried out, it is often desirable to stop the trial early if there is convincing evidence that one treatment is superior to the others or there is clearly no difference between the treatments. The use of sequential methods can lead to substantial reductions in the numbers of patients required compared with a fixed-sample design in order to achieve the same **power**. The methods involve the monitoring of a test statistic, and the trial is stopped as soon as this statistic crosses some stopping boundary. The stopping boundary is chosen in order that the trial has a given type I error probability and power for detecting a prespecified treatment difference. Since the development of sequential methods in the 1940s, there is now a wide range of methods available. The method to be used depends on such considerations as the type of patient response being observed, the testing problem under consideration, and how many treatments there are.

Most of the work on sequential testing in the 1950s and 1960s focused on the fully-sequential case, that is, the data are inspected after each new patient's response. Many of the early sequential methods developed for clinical trials are described in the classic book [1]. The often impractical nature of fully-sequential methods led to the development of group-sequential methods in the 1970s. With these methods, the data are inspected after each group of patient responses, and approaches are available for both equal and unequal group sizes. For a given type I error probability and power, the values of the stopping boundary are computed numerically for each of the planned *interim analyses* [4]. Two statistical packages are available for the design of sequential clinical trials—PEST 4 [2] and EaSt-2000 [3]. Both packages also incorporate methods for drawing statistical inferences upon termination, such as **confidence intervals**.

As an example, suppose that we wish to compare two treatments A and B when patient response is binary. Then, the usual measure of treatment

difference is the *log odds ratio* (see **Odds and Odds Ratios**) $\theta = \log\{p_A q_B / (p_B q_A)\}$, where p_A and p_B are the success probabilities for the two treatments, and $q_A = 1 - p_A$ and $q_B = 1 - p_B$. Following [5], one approach is to use the statistics

$$Z = \frac{n_B S_A - n_A S_B}{n_A + n_B} \text{ and} \\ V = \frac{n_A n_B (S_A + S_B)(n_A + n_B - S_A - S_B)}{(n_A + n_B)^3}, \quad (1)$$

where n_A and n_B are the numbers of patients on the two treatments, and S_A and S_B are the numbers of successes. For example, if the probability of success for treatment B is expected to be about $p_B = 0.6$ and we wish to test whether treatment A leads to an improvement of $p_A = 0.8$, then, for a given type I error probability of $\alpha = 0.05$ and power of $1 - \beta = 0.9$, we can use PEST 4 to obtain the stopping boundary for the test. After each group of patients, the value of Z is plotted against the value of V until a point falls outside of the stopping boundaries. Depending on which boundary is crossed first, we either conclude that treatment A leads to an improvement or that there is insufficient evidence of a treatment difference.

References

- [1] Armitage, P. (1975). *Sequential Clinical Trials*, 2nd Edition, Blackwell, Oxford.
- [2] Brunier, H.C. & Whitehead, J. (2000). *Pest 4.0 Operating Manual*, University of Reading.
- [3] Cytel Software Corporation (2001). *EaSt-2001: Software for the Design and Interim Monitoring of Group Sequential Clinical Trials*, Cytel Software Corporation, Cambridge.
- [4] Jennison, C. & Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*, Chapman and Hall, London.
- [5] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, Revised 2nd Edition, Wiley, Chichester.

D.S. COAD

Setting Performance Standards: Issues, Methods

BARBARA S. PLAKE

Volume 4, pp. 1820–1823

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Setting Performance Standards: Issues, Methods

Standardized tests are used for many purposes. Such purposes include which students are eligible for high school graduation, which applicants pass a licensure test, and who, among qualified applicants, will receive scholarships or other awards and prizes. In order to make these decisions, typically one or more score values is identified from the possible test scores to be the 'passing score' or 'cutscore'.

Setting performance standards, or cutscores, is often referred to as '*standard setting*'. This is because, by determining the passing score, the 'standard' is set for the score needed to pass the test. Standard setting methods are often differentiated by whether they focus on the test takers (the examinee population) or on the test questions. Methods are presented that illustrate both of these approaches.

Examinee Focused Methods

Two different examinee focused methods are discussed. The first method includes strategies whereby examinees are assigned to performance categories by someone who is qualified to judge their performance. The second method uses the score distribution on the test to make examinee classifications.

When using the first of these methods, people who know the examinees well (for example, their teachers) are asked to classify them into performance categories. These performance categories could be simply 'qualified to pass' or 'not qualified to pass', or more complex, such as 'Basic', 'Proficient' and 'Advanced'. When people make these classifications, they do not know how the examinees did on the test. After examinees are classified into the performance categories, the test score that best separates the classification categories is determined.

The second general approach to setting cutscores, using examinee performance data, is through 'norm-based methods'. When using norm-based methods, the scores from the current examinee group are summarized, calculating the average (or mean) of the set of scores and some measure of how spread out

across the score range the test scores fall (variability). In some applications, the cutscore is set at the mean of the score distribution or the mean plus or minus a measure of score variability (standard deviation). Setting the passing score above the mean (say one standard deviation above the mean) would, in a bell-shaped score distribution, pass about 15% of the examinees; likewise, setting the passing score one standard deviation below the mean would fail about 15% of the examinees.

Test-based Methods

Test-based methods for setting passing scores consider the questions that comprise the test. Before discussing test-based methods, it is necessary to know more about the kinds of questions that comprise the test. Many tests are composed of multiple-choice test items. These items have a question and then several (often four) answer choices from which the examinee selects the right or best answer. Multiple-choice test items are often favored because they are quick and easy to score and, with careful test-construction efforts, can cover a broad range of content in a reasonable length of testing time.

Other tests have questions that ask the examinee to write an answer, not select an answer from a set list (such as with multiple-choice items). Sometimes these types of questions are called *constructed-response* questions because the examinee is required to construct a response. Some agencies find these kinds of questions appealing because they are seen as more directly related, in some situations, to the actual work that is required.

The methods used for setting passing scores will vary based on the type of questions and tasks that comprise the test. In every case, however, a panel of experts is convened (called *subject matter experts*, or SMEs). Their task is to work with the test content in determining the recommended minimum passing score for the test, or in some situation, the multiple cutscores for making performance classifications for examinees (such as Basic, Proficient, and Advanced).

Multiple-choice Questions

Until recently, most of the tests that were used in standard setting contained only multiple-choice questions. The reasons for this were mostly because

2 Setting Performance Standards

of the ability to obtain a lot of information about the examinees' skill levels in a limited amount of time. The ease and accuracy of scoring were also a consideration in the popularity of the multiple-choice test question in high-stakes testing. Two frequently used standard setting methods for multiple-choice questions are described.

Angoff Method. The most prevalent method for setting cutscores on multiple-choice tests for making pass/fail decisions is the Angoff method [1]. Using this method, SMEs are asked to conceptualize an examinee who is just barely qualified to function at the designated performance level (called the *minimally competent candidate*, or MCC). Then, for each item in the test, the SMEs are asked to estimate the probability that a randomly selected MCC would be able to correctly answer the item. The SMEs work independently when making these predictions. Once these predictions have been completed, the probabilities assigned to the items in the test are added together for each SME. These estimates are then averaged across the SMEs to determine the overall minimum passing score. Because the SMEs will vary in their individual estimates of the minimum passing score, it is possible to also compute the variability in their minimum passing score estimates. The smaller the variability, the more cohesive the SMEs are in their estimates of the minimum passing score. Sometimes this variability is used to adjust the minimum passing score to be either higher or lower than the average value.

In most application of the Angoff standard setting method, more than one set of estimates is obtained from the SME. The first set (described above) is often called *Round 1*. After Round 1, SMEs are typically given additional information. For example, they may be told the groups' Round 1 minimum passing score value and the range, so they can learn about the level of cohesion of the panel at this point. In addition, it is common for the SMEs to be told how recent examinees performed on the test. The sharing of examinee performance information is somewhat controversial. However, it is often the case that SMEs need performance information as a reality check. If data are given to the SMEs, then it is customary to conduct a second round of ratings, called *Round 2*. The minimum passing scores are then calculated using the Round 2 data in a manner identical to that for the Round 1 estimates. Again,

panels' variability may be used to adjust the final recommended minimum passing score.

There have been many modifications to the Angoff method, so many in fact that there is not a single, standard set of procedures that define the Angoff standard setting method. Variations include whether or not performance data is provided between Rounds, whether or not there is more than one round, whether or not SMEs may discuss their ratings, and how the performance estimates are made by the SMEs. Another difference in the application of the Angoff method is the definitions for the skill levels for the MCC.

Bookmark Method. Another standard setting method that is used with multiple-choice questions (and also with mixed formats that include multiple-choice and constructed-response question) is called the *Bookmark Method* [2]. In order to conduct this method, the test questions have to be assembled into a special booklet with one item per page and organized in ascending order of difficulty (easiest to hardest). SMEs are given a booklet and asked to page through the booklet until they encounter the first item that they believe that the MCC would have less than a 67% chance of answering correctly. They place their bookmark on the page prior to that item in the booklet. The number of items that precedes the location of the bookmark represents an individual SME's estimate of the minimum passing score. The percent level (here identified as 67%) is called the *response probability* (RP); RP values other than 67% are sometimes used. Individual SME estimates of the minimum passing score are shared with the group, usually graphically. After discussion, SMEs reconsider their initial bookmark location. This usually continues through multiple rounds, with SME's minimum passing scores estimates shared after each round. Typically there is large diversity in minimum passing score estimates after round 1, but the group often reaches a small variability in minimum passing score estimates following the second or third round.

Constructed-response Questions

As stated previously, constructed-response questions ask the examinee to prepare a response to a question. This could be a problem-solving task for mathematics, preparing a cognitive map for a reading passage,

presenting a critical reasoning essay on a contemporary problem. What is common across these tasks is that the examinee cannot select an answer but rather must construct one. Another distinguishing feature of constructed-response questions is that they typically are worth more than one point. There is often a scoring rubric or scoring system used to determining the number of points an examinee's answer will receive for each test question.

There are several methods for setting standards on constructed-response tests. One method, called *Angoff extension* [3] asks the SMEs to estimate the total number of points that the MCC is likely to earn out of the total available for that test question. The calculation of the minimum passing score is similar to that for the Angoff method, except that the total number of points awarded to each question is added to calculate the minimum passing score estimate for each SME. As with the Angoff method, multiple rounds are usually conducted with some performance data shared with the SMEs between rounds.

Another method used with constructed-response questions is called the *Analytical Judgment* (AJ) method [5]. For this method, SME read examples of examinee responses and assign them into multiple performance categories. For example, if the purpose were to set one cutscore (say for Passing), the categories would be 'Clearly Failing', 'Failing', 'Nearly Passing', 'Just Barely Passing', 'Passing', and 'Clearly Passing'. Usually a total of 50 examinee responses are collected for each constructed-response questions, containing examples of low, middle, and high performance. Scores on the examples are not shown to the SMEs. After the SMEs have made their initial categorizations of the example papers into these multiple performance categories, the examples that are assigned to the 'Nearly Passing' and 'Just Barely Passing' categories are identified. The scores on these examples are averaged and the average score is used as the recommended minimum passing score. Again, the variability of scores that were assigned to these two performance categories may be used to adjust the minimum passing score. An advantage of this method is that actual examinee responses are used in the standard setting effort. A disadvantage of the method is that it considers the examinees' test responses question by question, without an overall consideration of the examinees' test performance. A

variation of this method, called the *Integrated Judgment Method* [4] has SMEs consider the questions individually and then collectively in making only one overall test classification decision into the above multiple categories.

Other methods exist for use with constructed-response questions but they are generally consistent with the two approaches identified above. More information about these and other standard setting methods can be found in Cizek's 2001 book titled *Setting Performance Standards: Concepts, Methods, and Perspectives*.

Conclusion

Tests are used for a variety of purposes. Because decisions are based on test performance, the score value for passing the test must be determined. Typically, a standard setting procedure is used to identify recommended values for these cutscores. The final decision about the value of the cutscore is a policy decision that should be made by the governing agency.

Tests serve an important purpose. It is desired to certify that the examinee does have the requisite skills and competencies needed to graduate from school programs, practice in an occupation or profession, or receive elevated status within a profession. If the passing scores on these tests are not set appropriately, there is no assurance that these outcomes will be achieved. Therefore, it is critically important that sound methods are used when determining these cutscores.

References

- [1] Angoff, W.H. (1971). Scales, norms, and equivalent scores, 2nd Edition, in *Educational Measurement*, R.L. Thorndike, ed., American Council on Education, Washington.
- [2] Cizek, G.J., ed. (2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*, Lawrence Erlbaum, Mahwah.
- [3] Hambleton, R.K. & Plake, B.S. (1995). Extended Angoff procedure to set standards on complex performance assessments, *Applied Measurement in Education* 8, 41–56.
- [4] Jaeger, R.M. & Mills, C.N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments, in *Setting Performance Standards:*

4 Setting Performance Standards

Concepts, Methods, and Perspectives, G.J. Cizek, ed.,
Lawrence Erlbaum, Mahwah.

- [5] Plake, B.S. & Hambleton, R.K. (2000). A standard setting
method for use with complex performance assessments:

categorical assignments of student work, *Educational
Assessment* **6**, 197–215.

BARBARA S. PLAKE

Sex-Limitation Models

THALIA C. ELEY

Volume 4, pp. 1823–1826

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sex-Limitation Models

Sex differences are one of the most marked and frequently reported features of behavioral phenotypes. Many aspects of emotional difficulties within the internalizing arena are more prevalent in females than males. For example, depression is roughly twice as common in women than in men [3]. In contrast, males show higher levels of externalizing or behavioral problems, with conduct problems two to three times as common in males than females [4]. Another area in which males show higher rates than females is in learning and communication difficulties. For example, language delays are more common in boys than girls, and autism is also more prevalent in males than females. As a result of these widespread sex differences, behavioral scientists have focused considerable attention on trying to identify methods by which to explore their origins. Sex limitation models use a **structural equation model**-fitting approach with twin or adoption data to test for both *quantitative* and *qualitative* sex differences in the genetic and environmental etiology of the phenotype.

A first step in any examination of sex differences is to establish whether there are mean or variance differences between the two sexes. Or, alternatively, if the phenotype is a disorder, whether there are prevalence differences between the sexes. All these features can be left free to differ between the two sexes in structural equation models, providing a more accurate fit to the data. Having established these core differences, the next step is to examine whether there are either *quantitative* and/or *qualitative* sex differences in the genetic and environmental etiology. In twin studies (see **Twin Designs**) and **adoption studies**, the simple **ACE model** divides the sources of variance into additive genetic influence (A), common or shared environment (C: environmental influences that make family members similar to one another) and nonshared environment (E: environmental influences that make family members different from one another).

Quantitative Sex Differences

Quantitative sex differences refer to there being different relative proportions of genetic, shared, and

nonshared environmental influence on the phenotype for the two sexes. Thus, with twin data, if there is a greater difference in resemblance between monozygotic (MZ) females and dizygotic (DZ) females than there is between the two groups in males, this indicates greater **heritability** for females. In order to test for differences of this kind, two parallel models must be run. First, a model in which A, C, and E are free to differ for males and females is run (heterogeneity or free model). Next, a model is run in which these parameters are fixed to be of the same size (homogeneity or constrained model). As the constrained model is 'nested' within the free model, the difference in fit between the two can be examined by looking at the change in chi-square, which is itself distributed as a chi-square with an associated degrees of freedom (difference between the degrees of freedom for the two models) and *P* value. A significant difference in the fit of these two models indicates a significant difference in the relative influence of A, C, and E on males and females for the trait.

In order to illustrate this type of effect, we take as an example some data published on antisocial behavior (ASB) in a sample of Swedish adolescents [2]. As can be seen in Figure 1, the intra-class correlations are given for 5 groups: MZ male, DZ male, MZ female, DZ female, DZ opposite-sex. One clear feature of these data is that while the DZ male and female correlations are similar, the MZ female correlation is significantly higher than the male MZ correlation. This indicates a greater genetic influence on females as compared

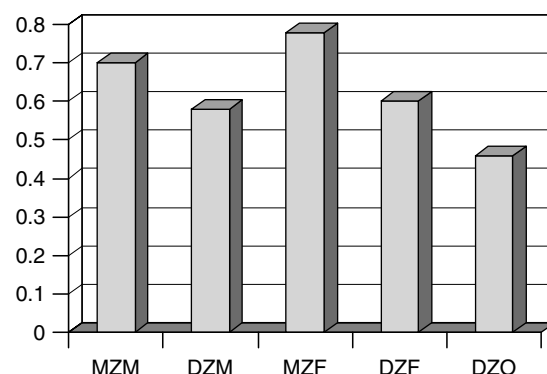


Figure 1 Within-pair correlations for nonaggressive antisocial behavior in young Swedish twins. Adapted from Eley, T.C., Lichtenstein, P. & Stevenson, J. (1999) [2]

2 Sex-Limitation Models

Table 1 Quantitative and qualitative sex-limitation models for nonaggressive antisocial behavior in young Swedish Twins

Model	Males			Females			DZOS		AIC	χ^2	df	p
	a ²	c ²	e ²	a ²	c ²	e ²	r _A	r _C				
1 ^b	0.47 ^a	0.29 ^a	0.24 ^a	0.47 ^a	0.29 ^a	0.24 ^a	0.5	1.0	4.83	28.83	12	0.01
2 ^c	0.26 ^a	0.47 ^a	0.27 ^a	0.60 ^a	0.19 ^a	0.21 ^a	0.5	1.0	-2.82	15.18	9	0.09
3	0.30 ^a	0.44 ^a	0.26 ^a	0.41 ^a	0.37 ^a	0.22 ^a	0.14	1.0	-6.10	9.91	8	0.27
4	0.30 ^a	0.44 ^a	0.26 ^a	0.41 ^a	0.37 ^a	0.22 ^a	0.5	0.69	-6.10	9.91	8	0.27

^aDenotes a parameter that could not be dropped from the model without significant deterioration in the fit.

^bModel 1 fits significantly worse than models 2 ($p < .01$), 3 & 4 ($p < .001$).

^cModel 2 fits significantly worse than models 3 & 4 ($p < .05$).

to males. In the constrained model, A, C, and E were estimated at 47, 29, and 24%, respectively. In contrast, for the sexes-free model, these parameters were 26, 47, and 27%, respectively, for males and 60, 19, and 21% for females. The difference in fit (chi-square) between the two models was 13.65 for 3 degrees of freedom, which is significant at the 1% level (see Table 1). Thus, there is a clear quantitative sex difference in the etiology of nonaggressive antisocial behavior in this sample of Swedish young people. However, there is also another interesting feature of the intraclass correlations in Figure 1, which is that the DZ opposite-sex correlation is significantly lower than that for the same-sex pairs. This is indicative of a *qualitative* sex difference.

Qualitative Sex Differences

While it is clear that genes and environment could influence the two sexes to a different extent, what is also of interest is that there may be differences in the specific aspects of the genetic or environmental influence for males and females. In other words, there may be different genes that make antisocial behavior heritable in girls from those that influence this phenotype in boys. Qualitative sex differences take this model-fitting approach one step further and allow for the examination of whether the genetic or environmental influences on the phenotype are the *same* in both boys and girls. DZ opposite-sex pairs allow us to examine this issue because any decrease in resemblance between them relative to same-sex DZ pairs is indicative of a lower level of sharing of either genes or environment. This is tested for either by allowing the genetic correlation between members of DZ opposite-sex pairs to be free rather

than fixed to 0.5 as is the case for same-sex pairs, or by leaving the shared environment correlation for DZO pairs free to be estimated rather than fixed at 1.0. Unfortunately, because there is only one more unit of information in this model (the DZO covariance), only one of these effects can be tested at any one time.

As noted above, in the data from the young Swedish twins, the DZO correlation was rather lower than that for the male or female DZ pairs (0.46 as compared to 0.58 and 0.60 respectively). The fit for the two models where either the genetic correlation or the shared environment correlation between members of the DZO pairs were left free to vary are given in the final two rows of Table 1. As can be seen, these models have identical fit, which is often the case, as it is difficult to distinguish between these two models. However, both of the qualitative sex difference models fit significantly better than the model with quantitative sex differences only, and the genetic and shared environment correlations are estimated at 0.14 (as compared to 0.5) and 0.69 (as compared to 1.0), respectively. This indicates that there are somewhat different genetic and/or shared environmental influences on nonaggressive antisocial behavior in girls and boys. Furthermore, in this model, there are still significant differences between the relative impact of genes and environment on males and females, with estimates of 30, 44, and 26%, respectively, for genes, shared, and nonshared environment in boys, and 41, 37, and 22%, respectively, in girls.

Sex Limitation and Extreme Scores

The models described above refer to sex differences in the origins of individual differences in the normal

range. There are two approaches that can be taken when examining sex differences in more extreme phenotypes. First, with disorders, a threshold model can be undertaken. This can allow for sex differences in the threshold on the estimated liability (above which individuals express the disorder), and also both quantitative and qualitative sex differences on the causes of variation in this underlying liability distribution. This approach can also be used for data that are very skewed, and that are best modeled as a series of categories with thresholds in between each on an underlying liability. However, if the phenotype is a continuous measure (e.g., depression symptoms), in addition to examining sex effects on the distribution of the full range of scores it may be of interest to examine sex effects on extreme scores (i.e., those with high depressive symptoms, thus at high risk for disorder). To examine quantitative sex differences in extreme scores, a regression approach is used [1] (*see DeFries–Fulker Analysis*) in which the co-twins' scores are predicted from the probands' scores (those in the extreme group), the degree of genetic relatedness (1.0 for MZ pairs, 0.5 for DZ pairs), sex, and the interactions between sex and both the proband score and the degree of genetic relatedness. The interaction between sex and proband score provides an overall indication of whether there are male–female differences in twin resemblance, and the interaction between sex and genetic relatedness indicates the difference in heritability between the two sexes. As noted in *DeFries–Fulker Analysis*, the core feature of this approach is that following transformation, mean scores for the co-twin groups fall between 0 and 1 and can be interpreted in a similar way to correlations. A model-fitting approach to this method also allows for the testing of qualitative sex differences [5]. The likelihood of both quantitative and qualitative sex differences is indicated by the relative size of MZ versus DZ co-twin means in males versus females, and in the comparison of DZO co-twin means with the same-sex DZ co-twins means. Figure 2 illustrates a pattern of transformed co-twin means indicative of both quantitative and qualitative sex differences. Examination of the data implies a **heritability** for males of around 80%, as compared to 40% for the females. Furthermore, the DZO co-twin mean is much lower than either the male or female DZ mean, indicating a qualitative sex difference in the influence of genes and

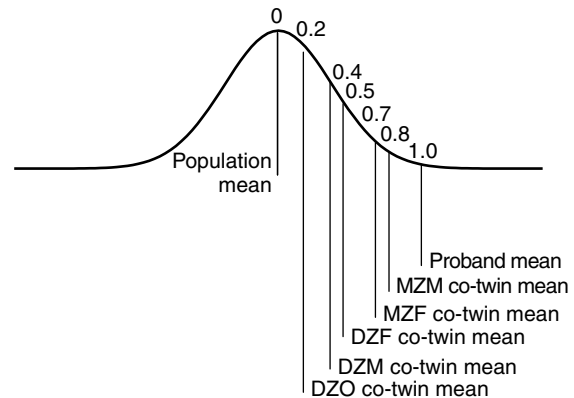


Figure 2 Hypothetical distribution of transformed population, proband, and co-twin means indicating both quantitative and qualitative sex differences

environment on extreme group membership for this trait.

Summary

In addition to the basic sex effects (mean, prevalence, and variance difference), a model-fitting approach can be used to test for two main types of sex difference: quantitative sex differences (level of genetic and environmental influence on males versus females) and qualitative sex differences (different genes impact on males versus females). These models can be tested for variation in the full range of scores, for the liability underlying a disorder, or for the membership of an extreme group defined as being at one or other end of a normally distributed trait. Such findings can be informative particularly with regard to molecular genetics. If there are different genes impacting on a trait in males versus females, then molecular genetic work needs to be undertaken on the two sexes independently. Similarly, for such finding with regard to environmental influence, social researchers need to examine the two sexes separately.

References

- [1] DeFries, J.C., Gillis, J.J. & Wadsworth, S.J. (1998). Genes and genders: a twin study of reading disability, in *Dyslexia and Development: Neurobiological Aspects of Extra-Ordinary Brains*, A.M., Galaburda, ed., Harvard University Press, Cambridge, pp. 186–344.

- [2] Eley, T.C., Lichtenstein, P. & Stevenson, J. (1999). Sex differences in the aetiology of aggressive and non-aggressive antisocial behavior: results from two twin studies, *Child Development* **70**, 155–168.
- [3] Hankin, B.L., Abramson, L.Y., Moffitt, T.E., Silva, A., McGee, R. & Angell, K.E. (1998). Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study, *Journal of Abnormal Psychology* **107**, 128–140.
- [4] Moffitt, T.E., Caspi, A., Rutter, M. & Silva, P.A. (2001). *Sex Differences in Antisocial Behaviour: Conduct Disorder, Delinquency and Violence in the Dunedin Longitudinal Study*, Cambridge University Press, Cambridge.
- [5] Purcell, S. & Sham, P.C. (2003). A model-fitting implementation of the DeFries-Fulker model for selected twin data, *Behavior Genetics* **33**, 271–278.

THALIA C. ELEY

Shannon, Claude E

SANDY LOVIE

Volume 4, pp. 1827–1827

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Shannon, Claude E

Born: April 30, 1916, in Michigan, USA.

Died: February 24, 2001, in Massachusetts, USA.

For most psychologists, Shannon is known and admired for only one piece of work, his 1948 paper on information theory and the mathematics of communication systems (see the accessible version in Shannon and Weaver, [5]). Although information theory approaches to cognition flourished in the 1950s and 1960s, culminating in Garner's textbook on information and structure dated 1962 [2], it is now only of historical interest, although one could speculate that the recent interest in complexity theory might once again have psychologist's dusting off their copies of 'A *Mathematical Theory of Communication*'. However, at that time, information theory appeared to offer a universal way of defining and assessing information and processing capacity, leading people like George Miller to argue in 1956 for the existence of the 'magical number 7' as the capacity limit to the human information processing system [3]. None of this would have been possible without the somewhat eccentric and reclusive electrical engineer Shannon.

Shannon's early education was at the University of Michigan, where he obtained BSc degrees in mathematics and electrical engineering in 1936, while he had obtained an MSc and a PhD by 1940 from the Massachusetts Institute of Technology (MIT), with a pioneering thesis on the Boolean analysis of logical switching circuits for the former degree, and one on population genetics for the latter. He had also worked at MIT with Vannevar Bush (who was later to invent *hypertext* as a way of structuring and accessing knowledge) on an early form of computer, the *differential analyzer*. Shannon's next move was to the prestigious Bell Laboratories of AT&T Bell Telephones in New Jersey, where he stayed until 1972, latterly as a consultant. In the interim, he became a visiting member of the faculty at MIT in 1956, and then Donner Professor of Science there from 1958 onward until his retirement two decades later.

Earlier on in 1948 he had published his key paper on the mathematics of information and communication where he defined the information in any

message as its predictability, thus turning it into a *probabilistic* measure. He also noted that all signals could be represented digitally to any degree of precision, that is, by binary digits or 'bits', an abbreviation that was also claimed by **John Tukey**. Thus, information could be represented as the sum of the ($\log_2$) of the probabilities of the events in an array of signals. The difference in stimulus and response information also defined the channel capacity of the system, and much work was done by psychologists in the 1950s and 1960s to measure this for many types of stimuli. Shannon also showed that the addition of extra bits in a message that was subject to noise or interference improved the reception of that message, leading to the concept of *redundancy*, an idea exploited by Attneave [1] in perception and by Miller [4] for the recall of simple strings.

Shannon was a somewhat retiring scientist who did most of his work behind closed doors. He also alarmed his colleagues at both Bell and MIT by riding a unicycle from his small collection of such machines down the corridors of these august institutions. It is even rumored that one unicycle was equipped with an asymmetric hub, which attracted crowds to see him progress in an up and down, sinusoidal fashion along the corridor! His other early contributions were to computer encryption, and the creation of unbreakable codes for military use, again drawing on the ideas first set out so eloquently in 1948. Later efforts were concentrated on AI and computing, in particular the formal outline of a practical Turing machine, ideas that had to await the age of the solid state device before seeing their implementation.

References

- [1] Attneave, F. (1959). *Applications of Information Theory to Psychology*, Holt, Rinehart & Winston, New York.
- [2] Garner, W.R. (1962). *Uncertainty and Structure as Psychological Concepts*, Wiley, New York.
- [3] Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review* **63**, 81–97.
- [4] Miller, G.A. (1958). Free recall of redundant strings of letters, *Journal of Experimental Psychology* **56**, 485–491.
- [5] Shannon, C.E. & Weaver, W. (1964). *A Mathematical Theory of Communication*, University of Illinois Press, Urbana.

SANDY LOVIE

Shared Environment

DANIELLE M. DICK

Volume 4, pp. 1828–1830

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Shared Environment

Behavior genetic twin studies partition variance in a behavior into genetic and environmental influences. Environmental influences can be further divided into shared (also called common) environment and **non-shared** (or unique) **environment**. When behavior geneticists refer to shared environment, they are referring to environmental influences that make sibling more similar to one another. Examples of shared environmental influences could include growing up in the same home, shared parental rules and upbringing, shared family experiences, sharing the same school and community, and peers that are common to both siblings.

In general, genetically informative studies of behavioral variation have not found a strong role for shared environment. In fact, behavior genetic research consistently demonstrates little or no effect of shared environment for many outcomes, including personality, psychopathology, and adult IQ [4]. Studies of children and adolescents have found more consistent evidence of shared environmental effects, as one might expect when children are still living together at home. Shared environment is widely accepted as an important influence on conduct problems and juvenile delinquency, and IQ in childhood [2, 4, 7, 8, 10]. Additionally, for substance use behaviors, the initiation of substance use appears to show evidence of shared environmental effects, but once an individual begins to use the substance, genetic influences gradually assume a greater role in impacting the behavior. In other words, parents and peers may have a big effect on whether (or when) one starts to drink or smoke, but once they initiate, an individual's own specific genetic predispositions assume greater influence on their subsequent use. This is evident in a longitudinal study of alcohol use among Finnish adolescents, whose frequency of alcohol use was measured at ages 16, 17, and 18.5. At age 16, more than 50% of the variation in frequency of alcohol use was attributed to shared environmental effects, but by age 18.5, less than a third was, as genetic influences had assumed a much greater role [5]. This is a common finding in the literature: shared environmental influences tend to decrease across the life span. As siblings increasingly gain independence from their families, make their own decisions, grow into adults, and start their

own careers and families, the impact of shared environment is largely displaced by genetic influences and unique environmental experiences.

So how can it be that there is so little evidence for strong shared environmental influences on behavior? Part of the confusion stems from the fact that shared environment – as behavior geneticists use the term – is not necessarily environmental effects that siblings objectively share, but rather, environmental events that make them more similar. Therefore, it is possible that an environmental event, such as parental divorce, could be objectively shared by both siblings, and yet have different effects on each child. For example, one sibling might react to the divorce by trying to please the parents by being on his/her best behavior in an attempt to reunite the parents. However, another sibling might react to the divorce by acting out, engaging in substance use and delinquent behavior, in a rebellious attempt to 'get back at' the parents. In this case, the environmental event was shared by the siblings, but it had the effect of making them different from one another rather than more similar. In this case, parental divorce would be classified as a 'nonshared environment' because it made the siblings' behavior more diverse. Thus, shared environment is not synonymous with family environment, and a lack of evidence supporting shared environmental effects should not be interpreted as evidence that family influences are not important. Familial influences may be nonshared, in the respect that they serve to make siblings different, rather than similar. Furthermore, nonfamilial influences, such as peer groups, can be shared and serve to make siblings more alike. The distinction between objectively shared environments and the term 'shared environment' as used by behavior geneticists is important when interpreting the literature [9].

Another reason that shared environmental influences might not be widespread in the literature is that **classic twin studies** are not particularly powerful for detecting shared environmental influences. The problem is that the 'a' and 'c' parameters, representing additive genetic, and common (shared) environmental effects, respectively, are highly correlated. **Power** analyses have indicated that in order to detect a common environmental effect of even 50%, more than 100 pairs of twins of each zygosity are needed. As the influence of common environment decreases, the number of pairs needed to

2 Shared Environment

detect the effect increases exponentially. Power is further decreased when the outcome is binary, such as affected or unaffected status. **Gene-environment correlation** can also contribute to inflated estimates of genetic influence at the expense of shared environment.

Despite these limitations, behavior genetic studies are advancing to better characterize shared environmental influences. While traditional behavior genetic designs modeled shared environment latently, twin researchers are increasingly *measuring* aspects of the shared environment and incorporating specific environmental measures into genetically informative designs. These models have more power to detect environmental effects, even when these environments only have small effects. In a study by Kendler and colleagues [3], parental loss was included in the classic biometrical twin model and contributed significantly to the variance in major depression – despite the fact that shared environmental influences were not significant when modeled latently. We have studied effects of parental monitoring and home atmosphere on behavior problems in 11- to 12-year-old Finnish twins; both parental monitoring and home atmosphere contributed significantly to the development of children's behavior problems, accounting for 2 to 5% of the total variation, and as much as 15% of the total common environmental effect. Recent research in the United Kingdom found neighborhood deprivation influenced behavior problems, too, accounting for about 5% of the effect of shared environment [1]. Incorporation of specific, measured environments into genetically informative designs offers a powerful technique to study and specify environmental effects.

Another new development in studying the shared environment has been to partition the shared environment into more distinct components. As mentioned previously, shared environmental effects can include everything from parents and peers, to school and community influences. In a unique design used by our research group in studying Finnish twins, a classmate control of the same gender and age was included for each twin. All members of each dyad shared the same neighborhood, school, and classroom, but only the co-twins shared common household experience. In this way, the classic shared environment component could be separated into family environment and school/neighborhood environment. Studying a large sample of 11- to 12-year-old same-sex Finnish twins,

sampled from more than 500 classrooms throughout Finland, we found that for some behaviors, including early onset of smoking and drinking, there were significant correlations for both control-twin and control-control dyads. This demonstrates that for these behaviors, the shared environment includes significant contributions from nonfamilial environments – schools, neighborhoods, and communities [6].

In conclusion, shared environment refers to any environmental event that makes siblings more similar to each other. It can include family, peer, school, and neighborhood effects; however, each of these potential environmental effects can also be nonshared if they have the effect of making siblings different from one another. New developments in behavior genetics are making it possible to specify shared environments of importance and to tease apart familial and nonfamilial effects.

References

- [1] Caspi, A., Taylor, A., Moffitt, T.E. & Plomin, R. (2000). Neighborhood deprivation affects children's mental health: environmental risks identified in a genetic design, *Psychological Science* **11**, 338–342.
- [2] Goldsmith, H.H. & Bihun, J.T. (1997). Conceptualizing genetic influences on early behavioral development, *Acta Paediatrica* **422**, p. 54–59.
- [3] Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1992). Childhood parental loss and adult psychopathology in women: a twin study perspective, *Archives of General Psychiatry* **49**, 109–116.
- [4] McGue, M. & Bouchard, T.J. (1998). Genetic and environmental influences on human behavioral differences, *Annual Review of Neuroscience* **21**, p. 1–24.
- [5] Rose, R.J., Dick, D.M., Viken, R.J. & Kaprio, J. (2001). Gene-environment interaction in patterns of adolescent drinking: regional residency moderates longitudinal influences on alcohol use, *Alcoholism: Clinical and Experimental Research* **25**, 637–643.
- [6] Rose, R.J., Viken, R.J., Dick, D.M., Bates, J., Pulkkinen, L. & Kaprio, J. (2003). It does take a village: nonfamilial environments and children's behavior, *Psychological Science* **14**, 273–277.
- [7] Rose, R.J., Dick, D.M., Viken, R.J., Pulkkinen, L., Nurnberger Jr, J.I. & Kaprio, J. (2004). Genetic and environmental effects on conduct disorder, alcohol dependence symptoms, and their covariation at age 14, *Alcoholism: Clinical and Experimental Research* **28**, 1541–1548.
- [8] Rutter, M., Silberg, J., O'Connor, T. & Simonoff, E. (1999). Genetics and child psychiatry: II. Empirical research findings, *Journal of Child Psychology and Psychiatry* **40**, 19–55.

- [9] Turkheimer, E. & Waldron, M. (2000). Nonshared environment: a theoretical, methodological, and quantitative review, *Psychological Bulletin* **126**, 78–108.
- [10] Waldman, I.D., DeFries, J.C. & Fulker, D.W. (1992). Quantitative genetic analysis of IQ development in young children: multivariate multiple regression with orthogonal polynomials, *Behavior Genetics* **22**, 229–238.

DANIELLE M. DICK

Shepard Diagram

JAN DE LEEUW

Volume 4, pp. 1830–1830

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Shepard Diagram

In a general nonmetric scaling situation (see **Multidimensional Scaling**), using the Shepard–Kruskal approach, we have data $y_i, \dots, y_n$ and a model $f_i(\theta)$ with a number of free parameters θ . Often this is a nonmetric multidimensional scaling model, in which the model values are distances, but linear models and inner product models can be and have been treated in the same way. We want to choose the parameters in such a way that the rank order of the model approximates the rank order of the data as well as possible.

In order to do this, we construct a loss function of the form

$$\sigma(\theta, \hat{y}) = \sum_{i=1}^n w_i (\hat{y}_i - f_i(\theta))^2,$$

where the w_i are known weights. We then minimize σ over all $\hat{y}$ that are monotone with the data y and over the parameters θ (see **Monotonic Regression**).

After we have found the minimum, we can make a **scatterplot** with the data y on the horizontal axis and the model values f on the vertical axis. This is what we would also do in linear or nonlinear regression analysis. In nonmetric scaling, however, we also have the $\hat{y}$, which are computed by **monotone regression**. We can add the $\hat{y}$ to vertical axis and use them to draw the best-fitting monotone step function through the scatterplot. This shows the **optimal scaling** of the data, in this case the monotone transformation of the data, which best fits the fitted model values. The scatterplot with y and f , and $\hat{y}$ drawn in, is called the Shepard diagram. In Figure 1, we show an example from a nonmetric analysis of the classical Rothkopf Morse code confusion data [2]. The stimuli are 36 Morse code signals. The raw data are the proportions p_{ij} , which signals i and j were judged to be the same by over 500 subjects. Dissimilarities were computed using the transformation

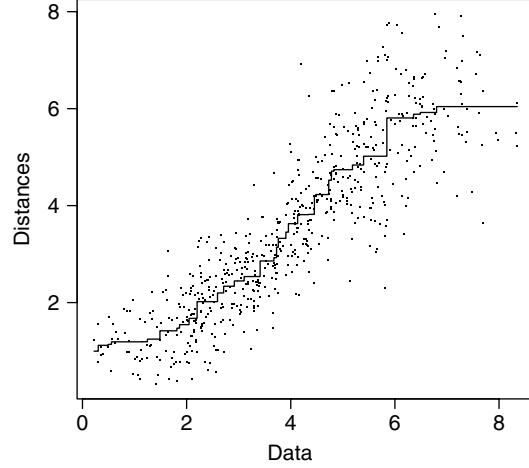


Figure 1 Shepard diagram Morse code data

$$\delta_{ij} = -\frac{1}{2} \log \frac{p_{ij} p_{ji}}{p_{ii} p_{jj}},$$

which is suggested by both Shepard's theory of stimulus generalization and by Luce's choice model for discrimination (see [1] for details). A nonmetric scaling analysis in two dimensions of these dissimilarities gives the Shepard diagram in Figure 1.

References

- [1] Luce, R.D. (1963). Detection and recognition, in *Handbook of Mathematical Psychology*, Vol. 1, Chap. 3, R.D. Luce, R.R. Bush & E. Galanter, eds, Wiley, pp. 103–189.
- [2] Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired associate learning, *Journal of Experimental Psychology* **53**, 94–101.

JAN DE LEEUW

Sibling Interaction Effects

DORRET I. BOOMSMA

Volume 4, pp. 1831–1832

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sibling Interaction Effects

The effects of social interaction (*see* **Social Interaction Models**) among siblings to individual differences in behavior were first discussed by Eaves [3] and later by Carey [1] and others. In the context of behavior genetic research, social interaction effects reflect that alleles may cause variation in one or more traits of individuals carrying these alleles, but may also, through social interaction, influence the phenotypes of individuals who do not carry them [4]. Social interactions between siblings thus create an additional source of variance and generate genotype-environment covariance if the genes causing the social interaction overlap with the genes that influence the phenotype under study.

Social interaction effects between siblings can either be cooperative (imitation) or competitive (contrast), depending on whether the presence in the family of, for example, a high-scoring sibling inhibits or facilitates the behavior of the other siblings. Cooperation implies that behavior in one sibling leads to similar behavior in the other siblings. In the case of competition, the behavior in one child leads to the opposite behavior in the other child.

In the classical **twin design**, cooperation or positive interaction leads to increased twin correlations for both monozygotic (MZ) and dizygotic (DZ) twins. The relative increase is larger for DZ than for MZ correlations, and the pattern of correlations thus resembles the pattern that is seen if a trait is influenced by common environmental factors. Positive interactions have been observed for traits such as antisocial tendencies [2]. Negative sibling interaction, or competition, will result in MZ correlations, which are more than twice as high as DZ correlations, a pattern also seen in the presence of genetic **dominance**. Such a pattern of correlations has been reported in genetic studies of Attention Deficit Hyperactivity Disorder (ADHD) and related phenotypes in children (e.g., [6]). In adults, a pattern of high MZ and low DZ correlations has been observed for anger [7].

In data obtained from parental ratings of the behavior of their children, the effects of cooperation and competition may be mimicked (e.g., [8]). When parents are asked to evaluate and report upon their children's phenotype, they may compare the behavior. In this way, the behavior of one child becomes the standard against which the behavior of the other

child is rated. Parents may stress either similarities or differences between children, resulting in an apparent cooperation or competition effect.

The presence of an interaction effect, either true sibling interaction or rater bias, is indicated by differences in MZ and DZ variances. If the interaction effect is cooperative the variances of MZ and DZ twins are both inflated, and this effect is greatest on the MZ variance. The opposite is observed if the effect is competitive; MZ and DZ variances are both deflated and again this effect is greatest on the MZ variance. In addition to heterogeneity in MZ and DZ variances, the presence of interaction affects MZ and DZ correlations. Under competition MZ correlations are much larger than DZ correlations. Under cooperation MZ correlations are less than twice the DZ correlation. These patterns of correlations are not only consistent with contrast and cooperation effects, but also with genetic nonadditivity (e.g., genetic dominance) and common environmental influences, respectively. In order to distinguish between these alternatives, it thus is crucial to consider MZ and DZ variance-covariance structures in addition to MZ and DZ correlations.

Rietveld et al. [5] carried out a simulation study to determine the statistical **power** to distinguish between the two alternatives of genetic dominance and social interaction. The results showed that when both genetic dominance and contrast effects are present, genetic dominance is more likely to go undetected in the classical twin design than the interaction effect. Failure to detect the presence of genetic dominance leads to slightly biased estimates of additive genetic, unique environmental, and interaction effects (*see* **ACE Model**). Competition is more easily detected in the absence of genetic dominance. If the significance of the interaction effect is evaluated while also modeling genetic dominance, small contrast effects are likely to go undetected, resulting in a relatively large bias in estimates of the other parameters. Alternative designs, such as including pairs of unrelated siblings reared together to the classical twin study, or including the nontwin siblings of twins, increases the statistical power to detect contrast effects as well as the power to distinguish between genetic dominance and contrast effects.

Sibling interaction will go undetected in the classical twin design if the genes responsible for the interaction differ from the genes which influence the trait. In such cases, a comparison with data from singletons

2 Sibling Interaction Effects

may permit further investigation. In parental ratings, the question whether an interaction effect represents true sibling interaction or rater bias also must be resolved through the collection of additional data, for example, from teachers.

References

- [1] Carey, G. (1986). Sibling imitation and contrast effects, *Behavior Genetics* **16**, 319–341.
- [2] Carey, G. (1992). Twin imitation for antisocial behavior: implications for genetic and family environment research, *Journal of Abnormal Psychology* **101**, 18–25.
- [3] Eaves, L.J. (1976). A model for sibling effects in man, *Heredity* **36**, 205–214.
- [4] Eaves, L.J., Last, K.A., Young, P.A. & Martin, N.G. (1978). Model-fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.
- [5] Rietveld, M.J.H., Posthuma, D., Dolan, C.V. & Boomsma, D.I. (2003). ADHD: sibling interaction or dominance: an evaluation of statistical power, *Behavior Genetics* **33**, 247–255.
- [6] Rietveld, M.J.H., Hudziak, J.J., Bartels, M., van Beijsterveldt, C.E.M. & Boomsma, D.I. (2004). Heritability of attention problems in children. Longitudinal results from a study of twins age 3 to 12, *Journal of Child Psychology and Psychiatry* **45**, 577–588.
- [7] Sims, J., Boomsma, D.I., Carrol, D., Hewitt, J.K. & Turner, J.R. (1991). Genetics of type A behavior in two European countries: evidence for sibling interaction, *Behavior Genetics* **21**, 513–528.
- [8] Simonoff, E., Pickles, A., Hervas, A., Silberg, J.L., Rutter, M. & Eaves, L. (1998). Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction, *Psychological Medicine* **28**, 825–837.

DORRET I. BOOMSMA

Sign Test

SHLOMO SAWILOWSKY

Volume 4, pp. 1832–1833

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sign Test

The logic of the nonparametric Sign test is ‘almost certainly the oldest of all formal statistical tests as there is published evidence of its use long ago by **J. Arbuthnott** (1710)!’ [6, p. 65]. The modern version of this procedure is generally credited to **Sir Ronald A. Fisher** in 1925. Early theoretical development with applications appeared in [1].

The Sign test is often used to test a population median hypothesis, or with matched data as a test for the equality of two population medians. It is based upon the number of values above or below the hypothesized median. Gibbons [4] positioned it as a counterpart of the parametric one-sample t Test. However, Hollander and Wolfe [5] stated ‘generally, (but not always) the sign test is less efficient than the **signed rank test**’.

Assumptions

Data may be discrete or continuous variables. It is assumed that the data are symmetric about the median, with no values equal to the hypothesized population median. There are a variety of procedures to handle values located at the median, including (a) deleting them from the analysis and (b) alternating the assignment for or against the null hypothesis.

Hypotheses

The null hypothesis being tested is $H_0 : M = M_0$, where M is the population median and M_0 is a hypothesized value for that parameter. The nondirectional alternative hypothesis is $H_1 : M \neq M_0$. Directional alternative hypotheses are of the form $H_1 : M < M_0$ and $H_1 : M > M_0$.

Procedure

Each x_i is compared with M_0 . If $x_i > M_0$, then a plus sign ‘+’ is recorded. If $x_i < M_0$, then a minus sign ‘-’ is recorded. In this way, all data are reduced to ‘+’ and ‘-’ signs. If the alternative hypothesis is $H_1 : M > M_0$, the logic of the test would indicate

that there will be more plus signs than minus signs. If there are values equal to the median, count half of them as plus and half as minus.

Test Statistic

The test statistic is the number of ‘+’ signs or the number of ‘-’ signs. If the expectation under the alternative hypothesis is that there will be a preponderance of ‘+’ signs, the test statistic is the number of ‘-’ signs. Similarly, if the expectation is a preponderance of ‘-’ signs, the test statistic is the number of ‘+’ signs. If the test is two-tailed, use the smaller of the two counts. Thus,

S = the number of ‘+’ or ‘-’ signs
(depending upon the context).

Large Sample Size

The large sample approximation is given by

$$S^* = \frac{S - (n/2)}{\sqrt{n/4}},$$

where S is the test statistic and n is the sample size. S^* is compared to the standard normal z scores for the appropriate α level. Monte Carlo simulations conducted by Fahoom and Sawilowsky [3] and Fahoom [2] indicated that n should not be less than 50.

Example

Consider the following sample data ($n = 15$). The null hypothesis is $H_0 : M = 18$, which is to be tested against the alternative hypothesis $H_0 : M \neq 18$.

20	33	4	34	13	6	29	17	39	26
+	+	-	+	-	-	+	-	+	+
13	9	33	16	36					
-	-	+	-	+					

Each of the scores is assigned a plus or a minus, depending on whether the score is above or below the median. The number of minuses is 7 and the number of pluses is 8. Thus, choose $S = 7$. Using a table of critical values, S is not significant and the

2 Sign Test

null hypothesis cannot be rejected on the basis of the evidence provided by the sample.

References

- [1] Dixon, W.J. & Mood, A.M. (1946). The statistical sign test, *Journal of the American Statistical Association* **41**, 557–566.
- [2] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.
- [3] Fahoome, G. & Sawilowsky, S. (2000). Twenty nonparametric statistics, in *Annual Meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.
- [4] Gibbons, J.D. (1971). *Nonparametric Methods for Quantitative Analysis*, Holt, Rinehart, & Winston, New York.
- [5] Hollander, M. & Wolfe, D. (1973). *Nonparametric Statistical Methods*, John Wiley & Sons, New York.
- [6] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY

Signal Detection Theory

JOHN T. WIXTED

Volume 4, pp. 1833–1836

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Signal Detection Theory

Signal-detection theory is one of psychology's most notable achievements, but it is not a theory about typical psychological phenomena such as memory, attention, vision or psychopathology (even though it applies to all of those areas and more). Instead, it is a theory about how we use evidence to make decisions. The evidence could be almost anything (e.g., the intensity of an auditory perception, the strength of a retrieved memory, or the number of symptoms suggestive of schizophrenia), and the task of the decision-maker is to decide whether or not enough evidence exists to declare the presence of the condition in question (e.g., the presence of a tone or a remembered item or a disease).

What makes decisions like these difficult is that we sometimes think we hear sounds that did not, in fact, occur. Similarly, faces sometimes seem familiar even upon first encounter, and symptoms of schizophrenia can be exhibited even by people who do not suffer from the disease. That being the case, we cannot make a positive decision merely because there is the slightest evidence pointing in the positive direction. Instead, there must be *enough* evidence, and signal-detection theory is all about understanding the process of deciding that there is, indeed, sufficient evidence to make a positive decision.

Signal-detection theory was initially introduced to the field of psychology in the area of psychophysics (Green & Swets, 1966), where the prototypical task was to decide whether a tone was presented on a particular trial or not. Since then, the basic detection framework has been much more broadly applied, such that it is now the major decision theory in tasks as diverse as recognition memory and diagnostic radiology. To illustrate the basic principles of signal-detection theory, a standard recognition memory task will be considered, though many other domains of application would do just as well. In a typical recognition task, participants are presented with a list of stimuli (e.g., a series of faces) followed by a recognition test involving items that appeared on the list (the targets) randomly intermixed with items that did not appear on the list (the lures, which are also known as distractors). The recognition test is the signal-detection task in this example. In the simplest case, the targets and lures are presented one at a time for a yes/no recognition decision ('yes' means that

the item is judged to have appeared on the list), and there is an equal number of each. The proportion of targets that receive a correct 'yes' response is the hit rate (HR), and the proportion of lures that receive an *incorrect* 'yes' response is the false alarm rate (FAR).

Although a test item falls into one category or the other (i.e., the item is either a target or a lure), a signal-detection analysis of the recognition task begins with the assumption that the test items vary *continuously* along a psychologically meaningful dimension. That dimension need not be named, but leaving it unspecified often seems somehow wrong to those who are new to the theory. Thus, for the sake of illustration only, we might name that dimension 'familiarity' (at least for the specific case of recognition memory). When presented with a test item on a recognition test, the participant will experience some sense of familiarity, and this holds true even for the lures. The familiarity of the lures might be low, on average, but some sense of familiarity will occur for each one (perhaps because the face somewhat resembles a face that appeared on the list or because it resembles an acquaintance, etc.). Moreover, the lures will not all generate the exact same low level of familiarity. Instead, they will generate a range of familiarity values (i.e., a distribution of values). Some of that variability arises because items presumably differ in inherent familiarity (e.g., an average-looking face might seem more familiar than a strange-looking face). Variability can also arise from internal sources. That is, even two faces with the same inherent levels of familiarity may generate different internal responses due to moment-to-moment processing differences. Thus, variability in subjective evidence always exists, even in the simplest auditory detection experiment in which the same physical stimulus is presented on many trials.

The left distribution in Figure 1 represents the hypothetical distribution of familiarity values associated with the lures on a recognition test. The mean of that distribution occurs at a relatively low point on the familiarity scale (labeled μ_{Lure} , although its true value is unknown), and the distribution itself simply reflects the fact that some of the lures have a higher familiarity value than the mean whereas others have a lower familiarity value than the mean. This distribution is analogous to the 'noise' distribution in an

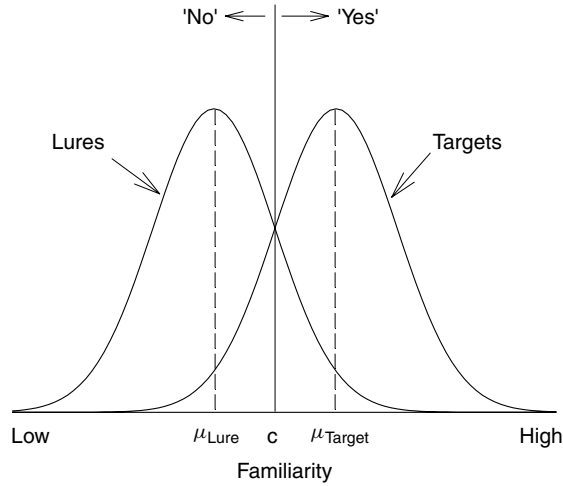


Figure 1 Prototypical signal-detection model of a Yes/No recognition memory test

auditory perception experiment (i.e., trials in which a to-be-detected tone is not presented).

The targets that are presented during the recognition test also have some mean familiarity value (labeled μ_{Target}), but that mean value is relatively high because the faces were recently seen on a list. Once again, though, all of the faces do not have the same high familiarity value. Instead, there is a distribution of familiarity values about the mean. In the hypothetical example shown in Figure 1, the distribution of familiarity values associated with the targets is also assumed to be Gaussian (i.e., normal), and it is assumed to have the same standard deviation as the lure distribution. This distribution is analogous to the 'signal' (i.e., white noise plus tone) distribution in an auditory detection experiment. Note that the equal-variance assumption is not required, and it is often not true in practice, but Figure 1 depicts the simplest version of detection theory.

The crux of the decision problem is this: how familiar must a test item be before it is declared to have appeared on the list? It is somewhat frustrating to realize that there is no obvious answer to this question, and it is up to the participant to pick a criterion familiarity value above which items receive a 'yes' response. In the hypothetical example illustrated in Figure 1, the participant has placed the decision criterion at the point labeled 'c' on the familiarity scale, which happens to be halfway between the means of the target and lure distributions. In a case like

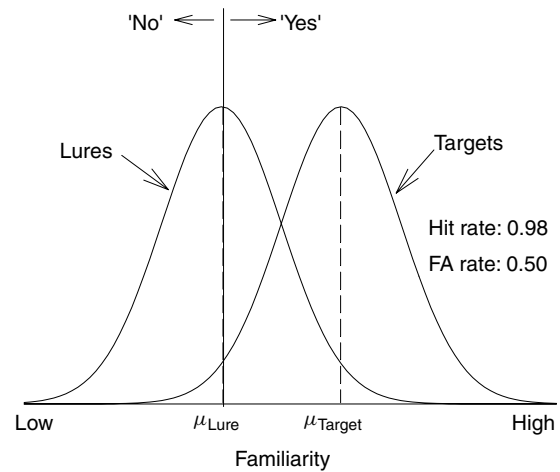
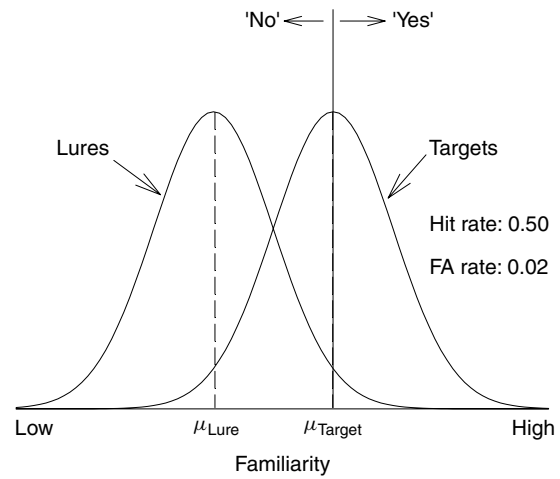


Figure 2 An illustration of conservative (upper panel) and liberal (lower panel) placements of the decision criterion

that, the HR would be 0.84 (i.e., 84% of the targets yield a familiarity greater than c) and the FAR would be 0.16 (i.e., 16% of the lures yield a familiarity greater than c), where the HR is the proportion of targets that receives a correct 'yes' response and the FAR is the proportion of lures that receives an incorrect 'yes' response. But a more conservative participant, illustrated in the upper panel of Figure 2, might place the criterion at a higher point on the familiarity scale, in which case both the HR and the FAR would both be lower. This conservative participant requires that a test item generate a familiarity value greater than μ_{Target} before saying 'yes'. Only 50% of targets and approximately 2% of the lures

exceed that familiarity value, so the HR for this participant would be 0.50 and the FAR would be 0.02. A liberal participant, illustrated in the lower panel of Figure 2, might instead only require that a test item generate a familiarity value greater than μ_{Lure} before saying ‘yes,’ in which case the HR would be 0.98 and the FAR would be 0.50. In practice, participants exhibit variability just like this (i.e., some have high hit and FARs whereas others have low hit and FARs).

What is the appropriate measure of memory performance for these participants? A natural choice would be to use the percentage of correct responses, but the problem with that choice is that it does not remain constant as bias changes. The neutral, conservative and liberal observers in the example above all have the same memory (i.e., the distributions are the same distance apart for each) – they differ only in their willingness to say ‘yes’. That is, they differ in *bias*, not in their ability to discriminate targets from lures. If half the test items are targets and half are lures, then percent correct is equal to the average of the HR and the correct rejection rate, where the correct rejection rate is equal to 1 minus the FAR. For the neutral, conservative and liberal observers, percent correct is equal to 84%, 74% and 74%, respectively. The value should remain constant because memory remains constant, and the fact that it does not reveals a flaw with that measure.

A better approach would be to use the distance between the means of the target and lure distributions as a measure of discriminability and to use the location of the decision criterion as a measure of bias. An intuitively appealing measure like the percentage of correct responses conflates these two separable properties of discriminative performance. The measure of discriminability derived from detection theory is d' , which is the distance between the means of the target and lure distributions *in standard deviation units* (not in units of familiarity). In the example shown in Figure 1, d' is equal to 2.0. That is, the means are 2 standard deviations apart. Had the items on the list been given less study time, d' would be less than 2.0 (down to a minimum of 0, at which point chance responding would prevail). Had the items been given more study time, d' would be greater than 2.0 (up to a practical maximum of about 4, at which point very few mistakes would be made).

The formula for computing d' is $z(HR) - z(FAR)$, where $z(p)$ is the z-score associated with the cumulative normal probability of p . Thus, for the neutral participant, $d' = z(0.84) - z(0.16)$, which is approximately equal to 1 – 1, or 2. For the conservative participant, $d' = z(0.50) - z(0.02)$, which is also approximately equal to 2. And for the liberal participant, $d' = z(0.98) - z(0.50)$, which, again, is approximately 2.0. Thus, detection theory provides a means of computing discriminability *independent of bias*. In this hypothetical example, the hit and FARs vary across observers, but the ability to discriminate a target from a lure (i.e., the distance between the means of the target and lure distributions) remains constant.

A plot of the HR vs. the FAR over different levels of bias is called the *Receiver Operating Characteristic* (ROC; see **Receiver Operating Characteristics Curves**). The typical ROC traces out a curvilinear path (some sense of this can be obtained by plotting the 3 pairs of hit and FARs discussed above), and the entire path represents a single value of d' over a continuous range of bias. Note that d' is an excellent choice of dependent measure when the equal-variance model applies. When an unequal-variance model is applicable, other detection-related measures are more appropriate [1, 2].

Detection theory also provides various ways to specifically quantify the degree of bias. One common measure is C , which is the distance from the point of intersection between the two distributions (i.e., from the midpoint) to the location of the criterion (again, in standard deviation units). The computational formula for C is: $-0.5*[z(HR) + z(FAR)]/d'$. This value will be zero for the unbiased case (i.e., criterion midway between the distributions), but it will be positive for more conservative (i.e., higher) placements of the criterion and negative for more liberal (i.e., lower) placements of the criterion. For the neutral, conservative and liberal responders considered above, the d' is 2.0 for all three and the corresponding C values are 0, 0.5, and –0.5. Thus, for any pair of hit and FARs, one can compute a bias-free discriminability measure (d') and a measure of the participant’s degree of bias (C).

It is important to emphasize that d' and C are measured *in standard deviation units*. We do not really know anything about the underlying distributions (i.e., we do not know their means or their standard deviations), but we can compute d' and C

from the obtained hit and FARs nonetheless. It seems like magic until you realize, for example, that if a participant has a HR of 0.84 and a FAR of 0.16, there is no way to draw the equal-variance Gaussian signal-detection depiction of those values except as drawn in Figure 1. The distributions must be placed two standard deviations apart, and the criterion must be placed halfway between. No other arrangement would correspond to a HR of 0.84 and a FAR of 0.16 (and this remains true even if we do not know what to name the decision axis, which has been named ‘*familiarity*’ in Figure 1 for the sake of illustration only).

The great value of signal-detection theory lies not only in its ability to separate discriminability and bias measures (which a measure like percent correct cannot do) but also in its ability to conceptualize

the underlying decision processes associated with an extremely wide range of endeavors. The conceptual utility of graphical depictions such as those shown in Figures 1 and 2 is hard to overstate whether the topic in question is perception, memory, or psychiatric diagnosis (to name just a few areas of application).

References

- [1] Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, Wiley, New York.
- [2] Macmillan, N.A. & Creelman, C.D. (1991). *Detection Theory: A User's Guide*, Cambridge University Press, Cambridge.

JOHN T. WIXTED

Signed Ranks Test

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Volume 4, pp. 1837–1838

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Signed Ranks Test

The nonparametric [4] Signed Ranks procedure provides a test of a hypothesis about the magnitude of the location parameter for a sample. It can be employed with a single sample to test a hypothesis about the median of the population sampled or with the differences between paired samples to test a hypothesis about the median of the population of such differences.

Although the nonparametric Wilcoxon Rank Sum test (see **Wilcoxon–Mann–Whitney Test**) is considerably more powerful than the parametric independent samples t Test under departures from population normality, the Wilcoxon Signed Ranks test presents more modest power advantages over the paired samples t Test.

Assumptions

It is assumed that the paired differences are independent, and originate from a symmetric, continuous distribution.

Hypotheses

The null hypothesis is $H_0 : \theta = \theta_0$, which is tested either against the nondirectional alternative $H_1 : \theta \neq \theta_0$ or against one of the directional alternatives $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$.

Procedure

In the one-sample case, compute the differences, D_i , by the formula

$$D_i = x_i - \theta_0 \quad (1)$$

and in the paired samples case

$$D_i = (x_{i2} - x_{i1}) - \theta_0, \quad (2)$$

where, for example, x_{i1} could be a ‘before’ or ‘pretest’ score and x_{i2} an ‘after’ or ‘posttest’ score.

Next, assign ranks, R_i , to the absolute values of these differences in ascending order, keeping track of the individual signs.

Test Statistic

The test statistic is the sum of either the positive ranks or the negative ranks. If the alternative hypothesis suggests that the sum of the positive ranks should be larger, then

$$T^- = \text{the sum of ranks of the negative differences.} \quad (3)$$

If the alternative hypothesis suggests that the sum of the negative ranks should be larger, then

$$T^+ = \text{the sum of ranks of the positive differences.} \quad (4)$$

For a two-tailed test, T is the smaller of the two rank sums. The total sum of the ranks is

$$\frac{N(N+1)}{2},$$

which gives the following relationship

$$T^+ = \frac{N(N+1)}{2} - T^-. \quad (5)$$

Large Sample Sizes

The large sample approximation is

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}, \quad (6)$$

where T is the test statistic. The resulting z is compared to the standard normal z for the appropriate alpha level. Monte Carlo simulations conducted by [3] and [2] indicated that the large sample approximation requires a minimum sample size of 10 for $\alpha = 0.05$ and 22 for $\alpha = 0.01$. Among others, [1] provides tables to be used with smaller sample sizes.

Example

A two-sided Wilcoxon Signed Rank test, with $\alpha = 0.05$, is computed on the following data set.

Test Score	87	90	88	88	89	91	89
Retest Score	90	85	94	97	96	90	99
D_i	3	–5	6	9	7	–1	10
R_i	2	3	4	6	5	1	7
Negative Ranks		3				1	

2 Signed Ranks Test

The obtained value, T , is the sum of the two negative ranks: $3 + 1 = 4$. The critical value for $N = 7$ is 3. Because $4 > 3$, the null hypothesis is rejected in favor of the alternative hypothesis that the median change in scores, from test to retest, differs from zero.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [2] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.
- [3] Fahoome, G., & Sawilowsky, S. (2000). Twenty non-parametric statistics, in *Annual Meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.
- [4] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Simple Random Assignment

CHAU THACH AND VANCE W. BERGER

Volume 4, pp. 1838–1840

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Simple Random Assignment

There are two types of simple random assignment – unrestricted random assignment and the random allocation rule (*see* **Randomization**). Unrestricted random assignment occurs when each subject in a randomized study is assigned to one of K possible treatments with a fixed probability, such as $1/K$, independent of all previous assignments in the study. Neither the total sample size of the study nor the sample size for each treatment group needs to be known in advance to use this procedure. For a study with two treatments and equal allocation to each treatment, unrestricted random assignment is equivalent to using a simple toss of a fair coin to allocate each subject to a treatment.

In practice, allocation is usually conducted not with coins but rather using random numbers generated from a computer. In contrast to unrestricted random assignment, the random allocation rule does require the total sample size of the study and the sample sizes in each treatment group to be known in advance. A randomly chosen subset of the subjects in the study is assigned to one treatment group. If there are only two treatment groups, then the unselected subjects comprise the other treatment group. If there are more than two treatment groups, then another randomly chosen subset is assigned to another treatment group, and so on, with the remaining subjects assigned to the last treatment group. For example, if four subjects were to be randomized to two groups, with two subjects to be randomized to each group, then there would be six possible pairs of subjects to make up the first group: {1, 2}, {1, 3}, {1, 4}, {2, 3}, {2, 4}, and {3, 4}. By the random allocation rule, each of these six groups is selected with the same probability, specifically $1/6$. In general, if there are two treatment groups and equal allocation to each, then the random allocation rule has $(2n)!/(n!)^2$ possible allocation sequences and unrestricted randomization has 2^{2n} possible allocation sequences, where there are $2n$ subjects in total, and n are assigned to each group for the random allocation rule. With unrestricted randomization, n is the expected size of each treatment group.

For both types of simple random assignment, the marginal (unconditional) probability of assignment to

a treatment is the same for each subject. For two treatments with equal allocation, each subject has the same probability, $1/2$, of being assigned to a treatment, over the set of all possible permutations, with both types of simple random assignment. With unrestricted random assignment, each subject still has the same constant conditional probability of assignment to each treatment, even given all the previous assignments. However, with the random allocation rule, the conditional probability of assignment to a treatment is not constant for each subject. For example, if four subjects were to be randomized to two groups using the random allocation rule and the first two subjects have both been randomized to the first treatment group, then the next two subjects would have no chance of being in the first treatment group.

One benefit of unrestricted random assignment is that it is the only allocation method to completely eliminate the type of selection bias that results from the random allocation process being predictable, thereby allowing an investigator to enter a specific subject into the study based on the concordance between this subject's characteristics and the next treatment to be assigned [2]. Note that many randomized studies are unmasked so that the prior assignments are known. Even those that are masked in theory may not be fully masked, so that at least some of the prior assignments are known. In such a case, any patterns created by restrictions on the random allocation allow for prediction of the future assignments [1]. It is in this context that unrestricted random assignment can be appreciated for its lack of any restrictions on the random allocation, and hence its resistance to selection bias [3].

For the random allocation rule, there is a small chance for selection bias in an unmasked study. Certainly, the last allocation is predictable, and depending on the allocation sequence, more allocations may also be predictable. The maximum number of predictable allocations is the size of the largest treatment group. For example, if four subjects were to be randomized to two groups and the first two subjects have been randomized to the first treatment group, then the investigator would know that the next two subjects would be in the second treatment group. The third and fourth allocations would be predictable. But if the allocation sequence were instead ABAB, then only the fourth allocation would be predictable [3]. The opportunity for selection bias can be eliminated

2 Simple Random Assignment

if all the subjects are allocated simultaneously rather than sequentially as they enter the study [1].

Although complete random assignment is perhaps the easiest allocation method to understand and implement, it is not widely used in practice. This is due mainly to the possibility of treatment imbalance – imbalances in the number of subjects assigned to each treatment – at any stage in the study. The imbalance in the numbers allocated to each group can be substantial if the sample size is small, and even if the sample size is large, it is still possible that there would be a gross imbalance during the early stages of the study. Imbalances in the size of the treatment groups can reduce the **power** to detect any true differences between the two treatment groups. The reduction in power is small, except for extreme imbalances. For example, if a study has 90% power with equal balance between treatments (1 : 1 assignment), then the power is reduced to about 85% if the imbalance in treatments is as extreme as 7 : 3 or greater [5].

If 20 subjects are being randomized to two treatments with 1 : 1 assignment (10 to each group on average, but this perfect balance would not be forced), then the probability of obtaining a 3 : 2 imbalance (12 subjects being assigned to one treatment and eight subjects being assigned to the other treatment) or worse is about 50%. With 100 subjects, this probability reduces to 5% [4]. As the sample size increases even further, this probability approaches 0%. These imbalances can reduce the power of the study to detect any true difference, but this concern can be addressed by using random allocation. However, both types of simple random assignment can be subject to accidental and chronological bias [3]. Accidental bias occurs when the subjects enrolled at one point in time differ systematically from those subjects enrolled at a later time but in the same study.

Consider the myriad number of ways that ambient conditions can change during the course of a study – study personnel may leave and be replaced, economic factors may change, new legislation may take effect during the study, flu season or allergy season may occur during one part of the study but not during another part of the study, and so on. These systematic differences that occur over time may not be measurable, or the variables that do measure them may not be recorded. This becomes a concern if a disproportionate number of subjects at one time are enrolled

to one group, and a disproportionate number of subjects at another time are enrolled to the other group. This is chronological bias [3], and more restrictions on the random allocation than simply terminal balance tend to be needed to minimize or eliminate it. We see, then, that there is a trade-off, in that more restrictions are needed to control chronological bias, yet fewer restrictions are required to control selection bias that arises from prediction of future assignments. It is generally impossible to simultaneously eliminate both [3], although a few exceptions to this rule of thumb exist [1]. In most cases, there will be covariate imbalances across the treatment groups, whether simple random assignment is used or not. However, complete random assignment has the smallest chance for accidental bias, while random allocation has a slightly larger chance [6].

Treatment imbalances do not invalidate the statistical tests that are generally performed to compare the treatment groups as long as they are random. Chronological bias, while called a bias, is actually a random error that is just as likely to favor one treatment group as it is to favor another. If it is the cause of a covariate imbalance, then standard inference is still unbiased. However, selection bias is a true bias, and so if it is the cause of a covariate imbalance, then standard comparisons are not unbiased, and standard tests are not valid. This, and the fact that unrestricted random assignment eliminates selection bias, is one reason to consider using it.

References

- [1] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias, *Journal of Modern Applied Statistical Methods* **2**(1), 80–86.
- [2] Berger, V.W. & Exner, D.V. (1999). Detecting selection bias in randomized clinical trials, *Controlled Clinical Trials* **20**, 319–327.
- [3] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.
- [4] Friedman, L.M., Furberg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, Springer, New York.
- [5] Lachin, J.M. (1988). Statistical properties of randomization in clinical trials, *Controlled Clinical Trials* **9**, 289–311.
- [6] Lachin, J.M. (1988). Properties of simple randomization in clinical trials, *Controlled Clinical Trials* **9**, 312–326.

CHAU THACH AND VANCE W. BERGER

Simple Random Sampling

VANCE W. BERGER AND JIALU ZHANG

Volume 4, pp. 1840–1841

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Simple Random Sampling

When a census of the entire population of interest is difficult to obtain, a sample is often used. There are many sampling designs that can be used to obtain a sample that would hopefully be representative of the population, and among these sampling designs, several allocate the same inclusion probability to each unit in the population (*see Survey Sampling Procedures*). Of course, for this to be the case, one would need the sampling frame from which units are selected to match the population of interest. Otherwise, any unit not in the sampling frame cannot be selected, or, rather, is selected with probability zero.

If the sampling frame is equivalent to the population of interest, and if each unit in the population (or sampling frame) has the same inclusion probability, and if all the units are independent, then we have a simple random sample. The independence refers to the selection probability, so that knowledge that not only does each unit have a common selection probability but also each pair of units has a (different) common selection probability. Another way to formulate this is to state that each sample consisting of a given number of units has the same probability of becoming the selected sample.

Simple random sampling treats each element in the population as being equally important; if this is true, then it would make sense that each unit would be sampled with equal probability. Sometimes this is not the case. For example, the population may be heterogeneous, with small but important subgroups that need to be represented adequately. In such a case, one might want to over-sample from these small subgroups. Also, if it is known that the subgroups are relatively equally well represented in the population, then this fact might be exploited to ensure balance in the sample, as well as in the population. For example, if sampling is undertaken without regard to gender, then it is possible that a gross imbalance would occur in the sample, in which case one gender would have more precise estimation than the other. Specifying a common number of males and females, and possibly selecting simple random samples from each, would guard against this potential problem. Of course, such a compound simple random sample does not itself constitute a

simple random sample, because not all subgroups would have the same probability of being selected. For example, a subgroup with unequal numbers of males and females would have probability zero of being selected.

Simple random sampling may be conducted with replacement or without replacement. As the name would suggest, simple random sampling with replacement involves sampling with replacement from the population. Each element in the population may be selected into the sample more than once. Consider, for example, an urn with 20 numbered balls. To obtain a sample of five balls by simple random sampling with replacement, one needs to draw a ball randomly from the urn five times, but after each selection the ball is put back into the urn before the next draw. Therefore, the sampling population remains the same for each draw. If we denote the size of the total population by N , then each unit has a chance of $1/N$ to be selected into the sample at each selection. A specific sample of size n has a chance of $(1/N)^n$ of being selected. This design has been modified to create the play-the-winner rule of patient allocation to clinical trials. Specifically, after a ball is selected, one replaces not only the selected ball but also more balls of the same color or of a different color depending on the outcomes [2].

In simple random sampling without replacement, units are sampled from the population without replacement. Consider the same example as above. To obtain a sample of 5 balls from the urn with 20 numbered balls by simple random sampling without replacement, one draws a ball randomly from the urn 5 times. But now the ball is not replaced after the selection, so after each draw, the population is different from what it was during the previous draw. After 5 draws, only 15 balls are left in the urn. And unlike the simple random sampling with replacement, the 5 balls in the sample are always distinct. With a population of size N , the chance of obtaining a specific sample of size n is $1/\binom{N}{n}$. The probability of obtaining a specific sample is now different from what it was when sampling with replacement. Note that even a simple random sample, be it with replacement or without, cannot confer independence to multiple observations taken from the same unit that was selected [1].

2 Simple Random Sampling

References

- [1] Max, L. & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research, *Journal of Speech, Language, and Hearing Research* **42**, 261–270.
- [2] Rosenberger, W. & Lachin, J.M. (2002). *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons, New York.

VANCE W. BERGER AND JIALU ZHANG

Simple V Composite Tests

RANALD R. MACDONALD

Volume 4, pp. 1841–1842

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Simple V Composite Tests

In a statistical test to test if a statistic S is significantly bigger than some fixed value c , the chance or **sampling distributions** of S based on an appropriate null hypothesis has to be assumed. This usually takes the form of supposing that S is c plus a variable error with an expected value of 0. Thus, under the null hypothesis the expected value of S is c and all the variation in the observed value of S is due to sampling error. The P value is then the probability of getting an S as or more extreme than was observed from this distribution (in a two-tailed test) (see **Classical Statistical Inference: Practice versus Presentation**). A more usual way of presenting this is to suppose that c is the value of an unknown parameter θ that is involved in the specification of the sampling distribution and that the null hypothesis H_0 is $\theta = c$. This hypothesis is called a simple hypothesis if θ is the only unknown parameter in the sampling distribution. Examples of tests of simple hypotheses include testing that a penny is fair on the basis of the proportion of heads in 100 tosses or that the mean IQ of a class is equal to 100, where IQs are supposed to be normally distributed with a standard deviation of 15.

Where the sampling distribution has more than one unknown parameter, the null hypothesis $H_0 \theta = c$ is said to be composite. Thus, a test of whether the mean IQ of a class is equal to 100, where IQs are supposed to be normally distributed with a standard deviation to be estimated from the data, is a composite hypothesis. However, composite hypotheses can involve several parameters. A composite null hypothesis has the general form $\theta_1 = c_1, \theta_2 = c_2, \dots, \theta_k = c_k$ with k degrees of freedom, where the θ s are unknown parameters and the c s fixed values. The general form incorporates hypotheses like $\theta_1 = \theta_2 = \theta_3, \dots = \theta_k$ as these can be rewritten as $\theta_1 - \theta_2 = 0, \theta_2 - \theta_3 = 0 \dots \theta_{k-1} - \theta_k = 0$ with $k - 1$ degrees of freedom. This is because the sampling distribution can be reexpressed using the differences between adjacent θ s plus the average value of θ instead of the θ s themselves. Examples of tests of composite hypotheses include **analyses of variance**, which test the hypothesis that the means of several groups are the same, chi-square tests with several degrees of freedom, and **meta analyses**, which

test the hypothesis that there is an effect present in a number of studies.

Statistical tests are also used to test nonparametric hypotheses. This term is somewhat loose and can be used in at least two ways [2]. First, it can apply to hypotheses where the probability distribution of the data is not specified. Examples here include using a Mann Whitney test (see **Wilcoxon–Mann–Whitney Test**) to test for differences between two sets of data that have been assumed to have been randomly sampled from identically shaped but unknown distributions and tests based on **bootstrap** methods where the sampling distributions are generated by randomly sampling the observed data to generate empirical distributions (see [1] and **Bootstrap Inference**). A second usage is that a nonparametric hypothesis can take the form that a set of data is consistent with some probability model or law with an unknown parameter or parameters. For example, one might wish to test the hypothesis that a set of data is normally distributed where both the mean and standard deviation are unknown. Such hypotheses are too complex to be entirely evaluated by the results of a single statistical test (see **Model Evaluation; Model Selection**).

Finally, it should be noted that all statistical tests of hypotheses whether simple, composite, or nonparametric, ultimately reduce the data to a single statistic. Even where a statistical test is derived from multivariate sampling distributions, the test can be seen as assessing whether its P value, itself a univariate statistic, is too small to be attributed to chance. On the other hand, if a composite null hypothesis has several degrees of freedom, it may be decomposed in a number of ways into independent component null hypotheses each with 1 degree of freedom (see **Multiple Comparison Procedures**). By testing these component hypotheses, more information about how the overall hypotheses are violated can be obtained though this raises the possibility of artificially increasing the achieved levels of significance (see **Multiple Testing**). It should also be noted that rejection of any component hypothesis itself implies a rejection of the overall hypothesis. Thus, despite the problems of multiple testing, where one has good reason for expecting that an effect with several degrees of freedom will conform to a particular pattern, the probability of rejecting the overall null hypothesis can be increased by using a test that incorporates these expectations.

2 Simple V Composite Tests

References

(See also **Directed Alternatives in Testing**)

- [1] Seigel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, New York.
- [2] Stuart, A., Ord, J.K. & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics*, 6th Edition, Vol. 2a, *Classical Inference and the Linear Model*, Arnold, London.

RANALD R. MACDONALD

Simulation Methods for Categorical Variables

STEFAN VON WEBER AND ALEXANDER VON EYE

Volume 4, pp. 1843–1849

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Simulation Methods for Categorical Variables

The simulation of real processes is a strong and legitimate tool in the hand of the research worker [1]. The parameters and variables used in simulation models can be discretely or continuously distributed. Under many conditions, discretely distributed variables are called **categorical variables**. They are called *categorical random variables* if they possess random characteristics. The following list presents examples of tasks that can meaningfully be approached using categorical random variables

1. *Monte Carlo Sampling* [9, 10]: the computer draws random samples from a population; sampling can be uniform, stratified, or by way of **bootstrap** resampling.
2. *Statistical Design or Simulation Design* [8, 11]: the following is a typical question for which simulation methods are employed: which sample size minimizes the cost of an expensive series of experiments. Costs can arise from the number of trials or the sample size, but also from the loss that results from falsely rejecting true hypotheses.
3. *Numerical mathematics* [4, 13]: Calculation of distributions that are either algebraically not tractable or can be calculated only with great effort; examples of such distributions include mixtures of various discrete distributions (see **Finite Mixture Distributions**). Statistical research concerning the performance of tests or the power of tests heavily relies on the means of simulations (e.g., [14]).
4. *Stochastic processes of national or business economy*: Simulation studies are used to test models of changes in population parameters and their impact on a national economy. Categorical random variates in such a study are, for instance, the number and the gender distribution of children in families.
5. *Random sampling to circumvent problems with complete enumeration* [12, 15]: For reasons of time or computer capacity, complete enumeration often causes problems. Therefore, random combinations are often drawn. Examples of applications include the optimization of the order of

steps in a process, the optimization of breeding processes, or the minimization of costs. The categorical random variates in these examples are the order of objects, the genetic patterns of the parental population, and the costs that come with combinations of parameters.

Discrete Distributions

A random variable that can assume the discrete values $0, 1, 2, \dots, n$ has a discrete distribution. In applied research, the discrete distribution, that is, the set of probabilities of patterns of variable categories, is almost always found by analyzing the data of a sample. Examples of such data include responses to questionnaires, code patterns in observational studies, or measurements that are categorized. Other discrete variables describe, for example, the gender distribution at universities over the last 25 years, or the distribution of countries that visiting students come from.

Discrete distributions are often derived by operations on discrete variables, for example, the sum function of one or more categorical variables. For example, the sum of 5 dichotomous variables that are scored as 0 and 1 is a binomially distributed random variable (see **Catalogue of Probability Density Functions**) with the six possible outcomes $0, 1, 2, \dots, 5$. An example of such a distribution describes the number of girls in families with 5 children.

One of the best known ways to create a *multivariate discrete distribution* is to cross two or more discrete variables. This operation yields a **contingency table**. Suppose we study the relationship between customer gender (G ; female = 1; male = 2) and time of the day of shopping (T ; 0 – 8 A.M. = 0; 8 A.M. – 4 P.M. = 1; 4 P.M. – 12 P.M. = 2). Crossing G with T results in the two-dimensional $G \times T$ contingency table.

Elements of a Simulation Study

Most simulations involve using computers. Researchers write programs using general purpose software such as FORTRAN, software specialized for mathematical problems such as MAPLE [3], or the programming tools provided by the general purpose statistical software packages such as SAS or SPSS [13]. The integral parts of a simulation study with categorical random variables are:

2 Simulation Methods for Categorical Variables

1. description of the question that the simulation will answer; examples of such questions are the estimation of the **power** of a statistical test, the estimation of the β -error, or the examination of relationships in cross-classifications;
 2. a model with a comprehensive list of the factors under study, for example, all categorical variables, the α -error, the sample size, the smallest difference of interest between a cell probability and the probability under independence;
 3. a good random number generator that creates uniformly distributed random numbers; alternatively, for large samples, a generator that creates normally distributed random numbers;
 4. algorithms that generate the desired discrete distribution from the uniformly distributed random numbers, for instance, the random frequencies of a contingency table;
 5. algorithms for the analysis of individual trials, for example, the test of a particular hypothesis for a particular cross-classification; storing of the results of the tests for the individual trial;
 6. algorithms that summarize the results for the individual trials; for example, these algorithms describe the number of trials in which a hypothesis was rejected.
1. *Using existing programs.* This option implies using a developer environment that provides the needed random number generators. For example, the environment MAPLE provides the uniform generator *random*; SAS provides the binomial generator *RanBin*; NEWTRAN provides the binomial generator *Binomial*; or CenterSpace provides the binomial generator *RandGenBinomial*.
 2. *Transforming uniform random numbers.* Here, one typically starts from an existing uniform random number generator that a programming environment such as Turbo Pascal, C, FORTRAN, or C++ makes available. The random numbers from these generators are then transformed such that the desired distribution results. The probabilities of the categories of the targeted distribution must be determined *a priori*.
 3. *Implementing stochastic processes.* First, a drawing process is started using the uniform random number generator. Then, a stochastic process is used to select numbers such that the desired distribution results. Here again, the probabilities of the categories of the targeted distribution must be determined *a priori*.

Generation of Discrete Random Numbers for Small Samples

Many programming environments provide acceptable generators for uniformly distributed random numbers. However, new generators are constantly being developed [7]. Most generators allow one to create reproducible series of random numbers. These are series that are the same each time the generator is invoked. However, most generators also allow one to create series that are hard to reproduce. For this option, a random function is used that determines the seed or the beginning of a series. As a compromise, one can reject the first k numbers from a series, with k determined anew for each simulation run.

If random numbers are needed that reflect a particular distribution, for example, the binomial distribution with n classes and an *a priori* specified probability for the occurrence of the target element A of $\{A, B\}$, a number of options exists. The three most important ones are:

The transformation of an uniformly distributed random number from the $[0, 1]$ interval to any discretely distributed random variable with n classes (= categories) is a mapping of the $[0, 1]$ interval onto the probability axis of the cumulative sum distribution. If the discrete probabilities are $0 < p_1 < p_2 < \dots < p_n = 1.0$, the simplest (but by no means the fastest) transformation method involves a series of if-statements. The following example uses the language of Turbo Pascal. Any other programming language could be used. The function RANDOM that is used in this example yields uniformly distributed random numbers from the $[0, 1]$ interval, where the number 1 does not occur. The cumulative probabilities must be given in the vector $\mathbf{p}$. That is, the programmer must either initialize the vector with the correct probabilities, or write program code that results in these probabilities. The first element of the vector $\mathbf{p}$ is p_1 , the second is $p_1 + p_2$, and so on. The last element has the value $p_1 + p_2 + \dots + p_n$. This value is always 1.0, that is, the simulation is exhaustive. In the example, k is the running index, and the quantity U is a dummy variable. The number of categories is n . The resulting quantity x has the desired distribution.

```

U:=RANDOM
for k:=n downto 1 do if
  (U<p[k]) then x:=k;

```

The stochastic drawing process of the binomial distribution is, for example, the drawing of white or black balls from an urn with returning each ball to the urn after registering its color, that is, with replacement. The probability of drawing a white ball is p , that of a black ball is $1 - p$. One generates a uniformly distributed random number $0 \leq U \leq 1$, using the function RANDOM, and takes 'white' or sets $\text{event} = 1$ if $U < p$. Alternatively, one takes 'black' or $\text{event} = 0$ if $U \geq p$. The desired n -class distribution is found by adding the n drawing events with values of 0 or 1. The categorical random variable is binomially distributed with the $n + 1$ categories being $0, 1, 2, \dots, n$. The following code (in Turbo Pascal) illustrates this procedure:

```

x:=0;
for k:=1 to n do if (RANDOM<p)
  then x:=x+1;

```

Whereas the first example (the transformation procedure above) is applicable to *any* distribution with a finite number of categories (as was mentioned above, the probabilities of the categories must be known), this second procedure creates only binomial distributions.

Generation of Discrete Random Numbers for Large Samples

If the expected frequency of a category is greater than 9, one can save computing time by using an approximative solution. Naturally, approximative solutions imply compromises concerning the characteristics of the resulting distribution. To give an example, we consider the generation of $2 \times 2 \times 2$ contingency tables with eight cells and multinomial distribution (see **Catalogue of Probability Density Functions**). The sample size N is assumed to be greater than 100. The probability p_j of Cell j should be known in this case. Therefore, the expectancy $e_j = Np_j$ is known also. Each cell frequency n_j of the eight cells is binomially distributed with $N + 1$ categories $0, 1, 2, \dots, N$, according to the cell probability p_j . The variance of a binomially distributed frequency is $\sigma^2 = Np_j(1 - p_j)$. Neglecting the mostly insignificant skew of the binomial distribution, and also

neglecting the mostly small deviation of the value $1 - p_j$ from the value 1, we can use the normal distribution of the cell frequency with mean Np_j and variance $\sigma^2 = Np_j$ as a sufficiently good approximation of the binomial distribution of a cell frequency.

To illustrate this method of creating binomially distributed cell frequencies, we now describe the calculation of a single discrete random cell frequency n_j by the following Turbo Pascal code. The quantity r is a real **dummy variable**, the variable j indicates the cell number.

```

r:=N*p[j];
(* Expectation of cell j*)
r:=r+NORMRAND*sqrt(r);
(* Adding standard deviation *)
if (r<0) then r:=0; (* Check
  against negative number *)
n[j]:=round(r); (* make it discrete
  and store *)

```

The sum of all cell frequencies n_j , S , is the desired sample size, N .

If there is no generator available that creates $N(0; 1)$, that is, normally distributed random numbers, one can use the method of summing 12 uniformly distributed random numbers from the interval $[0, 1]$. The Turbo Pascal code for this procedure is

```

Function NORMRAND : real;
(* normally {N,0;1}
  distributed using the central
  limit theorem of Gauss *)
var sum: real;
    i: integer;
begin
  sum:=0;
  for i:=1 to 12 do sum:=sum+
    RANDOM; NORMRAND:=sum-6;
end;

```

The procedure proposed by Box and Muller (1957) works with nearly the same speed. It uses two random numbers, r_1 and r_2 , with uniform $[0, 1]$ distribution, and transforms them to be a normally distributed $N(0; 1)$ random number. The transformation uses the mathematical functions square root, natural logarithm, and sine. The generated values are not limited to the interval between -6 and $+6$, as are the values generated by the method of summing 12 uniformly distributed random numbers. The Turbo Pascal code for Box and Muller's method follows.

4 Simulation Methods for Categorical Variables

```

Function NORMRAND : real;
(* normally  $N\{0;1\}$ 
   distributed random numbers
   using a transformation by
   Box and Muller *)
var r1 : real;
(* Number pi is pre-defined
   in Turbo Pascal *)
begin
  repeat r1:=RANDOM
  until r1>0;
  NORMRAND:=sqrt(-2*ln(r1))*
    sin(2*pi*RANDOM);
end;

```

Generation of Other Discrete Distributions

In behavioral research, a small number of basic theoretical discrete distributions is used. Among these are the binomial, the multinomial, the product-multinomial, the hypergeometric, and the Poisson distributions [6] (*see Catalogue of Probability Density Functions*). If one draws samples from samples from a population, we find ourselves in a situation in which frequencies follow both the binomial and the hypergeometric distributions. This case is of practical interest, because questionnaires are often administered in small units of larger entities, for example, departments of a college, small subsidiary plants.

The Poisson distribution is a special case because the number of categories is not restricted *a priori*. If categories exist with numbers $k \gg \lambda$, where λ is the expectation, then these categories come with very small probabilities, one truncates the distribution by eliminating large values of k . The remaining probabilities are then distributed proportionally to the first n categories. When programming Poisson processes, the transformation method is preferred over programming as a stochastic process.

In the following paragraphs, we illustrate the creation of a binomial distribution in a two-dimensional cross-classification. In the first step, we use the transformation method to generate Poisson-distributed random numbers.

In the following example, we simulate the answers that exactly 100 women and 100 men gave to a question. The question concerned a symptom such as headaches. The answers were scaled as 1 = none to 5 = severe. We assume $\lambda_f = 2.5$ for the female respondents, and $\lambda_m = 3.5$ for the male respondents.

The following steps are performed to generate the code for the 2×5 contingency table $CT[i, k]$, with $i = 1, 2$ and $k = 1, \dots, 5$.

1. Set all cells of $CT[i, k]$ to zero.
2. Compute, using the Poisson formula $P_k = \lambda^k/k!e^{-\lambda}$ for $k = 1, 2, \dots, 5$, the 2×5 class probabilities, and store them into the vectors P_f for women and P_m for men.
3. Calculate the two sums of probabilities S_{Pf} and S_{Pm} from the two vectors P_f and P_m to truncate the distributions. Divide the probabilities in vector P_f by S_{Pf} , and P_m by S_{Pm} (renorming by prorating).
4. Calculate the cumulative probabilities in the vectors P_f and P_m using commands such as 'for $k := 2$ to 5 do $P[k] := P[k] + P[k - 1]$;'.
5. Repeat 100 times the algorithm 'U := RANDOM; for $k := 5$ downto 1 do if ($U < P[k]$) then $x := k$; $CT[1,x] := CT[1,x] + 1$;'.
6. Repeat 100 times the algorithm 'U := RANDOM; for $k := 5$ downto 1 do if ($U < P[k]$) then $x := k$; $CT[2,x] := CT[2,x] + 1$;'.
7. After completing Step 6, the contingency table is ready for further analysis, for example, for calculation of a test statistic that describes the association between Gender and headaches.

In the following sections, we illustrate the stochastic drawing process of creating random numbers for discrete distributions for the case in which the class probabilities are unknown. The drawing process for the *hypergeometric distribution* (*see Catalogue of Probability Density Functions*) uses the urn example, just as for the binomial distribution, but without replacing the ball. As a consequence, each drawing changes the probabilities of the balls that remain in the urn. Let N_p be the total number of balls (respondents, responses, observations) in a subpopulation, for example, the employees of a factory, and the number of white balls i_p , and the number of black balls $j_p = N_p - i_p$. The start probability p of white balls is $p = i_p/N_p$. The following code, in Turbo Pascal, illustrates the drawing process for the generation of hypergeometrically distributed random numbers x with value range $0, 1, \dots, \min(n, j_p)$. The parameters of this

process are the *a priori* probability p , the number n of categories, and the size N_p of the subpopulation.

```
x:=0;
for k:=1 to n do
begin if RANDOM<p then begin
  x:=x+1; jp=jp-1; end;
  p:=jp/(Np-k);
end;
```

An extension of the binomial distribution is the *multinomial distribution*. This is the distribution of variables with more than two categories. Examples of such variables are the die with six possible numbers and the $k = 5$ possible categories for a question in a questionnaire. In simulations, the researcher has to determine the probabilities $p_1, p_2, \dots, p_k$ of the alternatives. The simplest option is to specify a uniform distribution, as what one would expect of a good die. Generally, finding these probabilities is often the goal of the simulation process. The random variate X is, after k drawings, k -dimensional, that is, x_1^n is the frequency of Alternative A_1 , x_2^n is the frequency of Alternative A_2 , and so on. The generation of a multinomial distribution can be performed using the above transformation procedure. We compute the vector p of the cumulative probabilities, set all elements of vector x to zero, and apply the random generation n times. After the n drawings, we find the k multinomially distributed frequencies as elements of vector x , that is, $x_1, x_2, \dots, x_k$, each of which has the value range $0, 1, \dots, n$. This procedure is illustrated by the following Turbo Pascal code.

```
U:=RANDOM;
for k:=5 downto 1 do if
  (U<p[k]) then j:=k;
x[j]:=x[j]+1;
```

A *product-multinomial distribution* results from simultaneously analyzing two or more multinomial distributions, each of which having the same number of categories. The drawings of the different multinomial distributions are performed separately. The drawing process is the same as described above for the multinomial distribution. The result of a product-multinomial drawing can most conveniently be presented in the form of a cross-tabulation. For example, 50 men and 100 women respond to the same question using a 5-category response format. The resulting

table has 2×5 cells. Let the first row contain the response frequencies of the male respondents. These frequencies can assume the values $0, 1, 2, \dots, 50$. This applies accordingly to the response frequencies for the women, in the second row. The total sum of all frequencies is 150. The row totals are 50 and 100, respectively.

A combined *binomial-hypergeometric distribution* results from drawing $n < N_p$ respondents from a finite subpopulation of size N_p , without replacement. The Turbo Pascal program given below illustrates this procedure. For this program, we need two loop counters, i and j , the *a priori* probability p for the occurrence of Alternative A_1 from $\{A_1, A_2\}$ in the finite basic population, the size N_p of our subpopulation, the number of classes, n , a dummy vector b for storing the N_p scores of the alternatives A_1/A_2 . From the vector b , the algorithm draws randomly the n respondents. Variable x has the desired binomial-hypergeometric distribution with value range $0, 1, 2, \dots, n$.

```
For j:=1 to Np do (* Create
  binomially distributed 0 or 1 *)
  if RANDOM<p then b[j]:=1
  else b[j]:=0;
i:=0;
x:=0;
repeat
  j:=round(Np*RANDOM+0.5);
  (* Random access index *)
  if ((j>0) and (j<=Np))
  then (* Check of
    index range *)
  begin
    if b[j]>=0 then
      (* Check whether
        selected already *)
    begin
      x:=x+b[j];
      (* add 0 or 1 *)
      b[j]:=-1;
      (* mark proband
        as selected *)
      i:=i+1;
      (* count selected
        probands *)
    end;
  end;
until (i=n);
(* Stop with case n *)
```


Analysis and Registration of the Elementary Events and Statistical Summary

The analysis of a single drawing varies depending on the aims of a study [5]. In most studies, results are derived directly from the random numbers. In our example with the Poisson distribution, one could calculate a X^2 -statistic for the cross-classification with the random numbers. The same applies to the example in which a product-multinomial distribution was simulated. Alternatives include studies in which running processes are simulated, for example, the evolution of a population. Here, we have a starting state, and after a finite number of drawings, we reach the end state. In this case, the computation of statistical summaries is already part of the analysis of the individual drawing. However, the result of the individual drawing is not of interest per se. One needs the summary of large numbers of drawings for reliable results.

The statistical summary describes the result of the simulation. In our example with the two-dimensional cross-classifications, the summary presents the number N_A of times the null hypothesis was rejected in relation to the total number of all simulation trials, N_R . A second result can be the Type II error (β or the complement of the statistical power of a test), calculated as $\beta = 1 - N_A/N_R$. Typically, simulation studies are meaningful because of the variation of parameters. In our examples with the two-dimensional cross-classification, we can, for instance, increase the sample size from $N = 20$ to $N = 200$ in steps of $\Delta N = 20$. Based on the results of this variation in N , we can produce graphs that show the degree to which the β -error depends on the sample size, N , while holding all other parameters constant (for an example, see [14]). Using this example, one can also show that the individual table is not really of interest. For the individual table, the β -error can only be either 0% or 100%, because we accept the correct alternative hypothesis or we reject it.

Using Different Computers and Parallel Computing

Simulation experiments of statistical problems are prototypical tasks for multiple computers and parallel computing. Suppose researchers have written a computer program or have produced it using a developer

environment, they can run the program under different parameter combinations using different computers. This way, the time needed for a study can be considerably shortened, the number of parameters can be increased, or the range of parameter variation can be broadened.

If a computer with multiple processors is available, for example, a machine with 256 processor units, one can employ software for parallel computing. The task of the programmer here is to start 256 trials using 256 random number processes (using 256 different seeds). At the end, the 256 resulting matrices need to be summarized.

Presenting Discrete Distributions

For most observed discrete distributions, theoretical distributions are defined. (see **Catalogue of Probability Density Functions**) These include the binomial, the hypergeometric and the multinomial distributions. The graphical representation of the *density* distribution is, because of the categorical nature of the distributions, a bar diagram (see **Bar Chart**) rather than a smooth curve. The probability of each class or category is shown by the corresponding height of a bar. These can be compared with bars whose height is determined by the theoretical distribution, that is, the comparison bars function as expected values. This way, the results of a simulation study can be evaluated with reference to some theoretical distribution.

The graphical representation of contingency tables with many cells is more complex. Here, the mosaic methods proposed by Friendly [2] are most useful.

References

- [1] Drasgow, F. & Schmitt, N. eds, (2001). *Measuring and Analyzing Behavior in Organizations*, Wiley, New York.
- [2] Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association* **89**, 190–200.
- [3] Heck, A. (1993). *Introduction to MAPLE*, Springer Verlag, New York.
- [4] Indurkha, A. & von Eye, A. (2000). The power of tests in configural frequency analysis, *Psychologische Beiträge* **42**, 301–308.
- [5] Klein, K.J. & Koslowski, S.W.J., eds, (2000). *Multilevel Theory, Research, and Methods in Organizations*, Jossey Bass, New York.
- [6] Mann, P.D. (1995). *Statistics for Business and Economics*, Wiley, New York.

-
- [7] Marsaglia, G. (2000). *The Monster-A Random Number Generator with Period Over 10^{2857} times Longer as Long as the Previous Touted Longest-Period One*, Boeing Scientific Research Laboratories, Seattle. Unpublished paper.
- [8] Mason, R.L., Gunst, R.F. & Hess, J.L. (1989). *Statistical Design and Analysis of Experiments*, Wiley, New York.
- [9] Nardi, P.M. (2002). *Doing Survey Research-A guide to Quantitative Methods*, Allyn & Bacon, Boston.
- [10] Noreen, E.W. (1989). *Computer Intensive Methods for Hypothesis Testing*, Wiley, New York.
- [11] Rasch, D., Verdooren, L.R. & Gowers, J.I. (1999). *Fundamentals in the Design and Analysis of Experiments and Surveys*, Oldenbourg, München.
- [12] Russell, S. & Norvig, P. (2003). *Artificial Intelligence-A Modern Approach*, 2nd Edition, Prentice Hall, Englewood Cliffs.
- [13] Stelzl, I. (2000). What sample sizes are needed to get correct significance levels for log-linear models? A Monte Carlo study using the SPSS procedure "Hilog-linear", *Methods of Psychological Research-Online* **5**, 95–116.
- [14] von Weber, S., von Eye, A. & Lautsch, E. (2004). The type II error of measures for the analysis of 2×2 tables, *Understanding Statistics* **3**(4), 259–282.
- [15] Wardrop, R.L. (1995). *Statistics: Learning in the Presence of Variation*, William C. Brown, Dubuque.

STEFAN VON WEBER AND
ALEXANDER VON EYE

Simultaneous Confidence Interval

CHRIS DRACUP

Volume 4, pp. 1849–1850

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Simultaneous Confidence Interval

It is sometimes desirable to construct **confidence intervals** for a number of population parameters at the same time. The problems associated with this are demonstrated most simply by considering intervals that are independent (in the sense that they are constructed from independent data sources). If two independent 95% confidence intervals are to be computed, then the probability that both will contain the true value of their estimated parameter is $0.95 \times 0.95 = 0.9025$. The probability that three such intervals will all contain their true parameter value falls to $0.95 \times 0.95 \times 0.95 = 0.857$. It is apparent that the more such intervals are to be computed, the smaller will be the probability that they will all contain their true parameter value. With independent confidence intervals, it would be relatively easy to increase the level of confidence for the individual intervals so as to make the overall confidence equal to 0.95. With two intervals, each could be constructed to give 97.47% confidence, since $0.9747 \times 0.9747 = 0.950$. With three intervals, each could be constructed to give 98.30% confidence, since $0.9830 \times 0.9830 \times 0.9830 = 0.950$.

With truly independent confidence intervals, which are addressing very different questions, researchers are not likely to be overconcerned about having a simultaneous confidence statement for all their estimated parameters. However, in the behavioral sciences, the construction of simultaneous confidence intervals is most commonly associated with experiments containing k conditions ($k > 2$). In such situations, researchers often wish to estimate the difference in population means for each pair of conditions, and to express these as confidence intervals. The number of such pairwise comparisons is equal to $k(k-1)/2$, and this figure increases rapidly as the number of conditions increases. Furthermore, the desired intervals are not statistically independent as the data from each condition contribute to the construction of more than one interval (see **Multiple Testing**).

The simple multiplication rule used above with independent confidence intervals is no longer appropriate when the confidence intervals are not independent, but the approach is basically the same. The

individual intervals are constructed at a higher confidence level in such a way as to provide, say, 95% confidence that all the intervals will contain their true parameter value. One of the most commonly used approaches employs the studentized range statistic [1, 2], and is the confidence interval version of Tukey's Honestly Significant Difference test (see **Multiple Comparison Procedures**). It will serve to illustrate the general approach.

For any pair of conditions, i and j , in a simple k -group study, the simultaneous 95% confidence interval for the difference in their population means is given by

$$(\bar{X}_i - \bar{X}_j) - q_{0.05,k,N-k} \sqrt{\frac{MS_{\text{error}}}{n}} \leq \mu_i - \mu_j \leq (\bar{X}_i - \bar{X}_j) + q_{0.05,k,N-k} \sqrt{\frac{MS_{\text{error}}}{n}} \quad (1)$$

where MS_{error} is the average of the sample variances of the k conditions, n is the number of observations in each condition (assumed equal), and $q_{0.05,k,N-k}$ is the critical value of the studentized range statistic for k conditions, and $N-k$ degrees of freedom, at the 0.05 significance level. By way of illustration, consider an experiment with $k = 3$ conditions each with $n = 21$ randomly assigned participants. With $k = 3$ and $N - k = 60$, the critical value of the studentized range statistic at the 0.05 level is 3.40. If the means of the samples on the dependent variable were 77.42, 69.52 and 64.97, and the MS_{error} was 216.42, then the simultaneous 95% confidence intervals for the difference between the population means would be

$$\begin{aligned} -3.01 &\leq \mu_1 - \mu_2 \leq 18.81 \\ 1.54 &\leq \mu_1 - \mu_3 \leq 23.36 \\ -6.36 &\leq \mu_2 - \mu_3 \leq 15.46 \end{aligned} \quad (2)$$

If confidence intervals are constructed in this way for all the pairs of conditions in an experiment, then the 95% confidence statement applies to all of them simultaneously. That is to say, over repeated application of the method to the data from different experiments, for 95% of the experiments *all* the intervals constructed will include their own true population difference. The additional confidence offered by such techniques is bought at the expense of a rather wider interval for each pair of conditions than would be required if that pair were the sole interest of the researcher. In the example, only the confidence

interval for the difference in means between conditions 1 and 3 does not include zero. Thus, Tukey's Honestly Significant Difference test would show only these two conditions to differ significantly from one another at the 0.05 level.

Scheffé [2] provided a method for constructing simultaneous confidence intervals for all possible contrasts (actually an infinite number) in a k -group study, not just the simple comparisons between pairs of conditions. But, once again, this high level of protection is bought at the price of each interval being wider than it would have been if it had been the only interval of interest to the researcher. As an example of the particular flexibility of the approach, consider an experiment with five conditions. Following inspection of the data, it may occur to the researcher that the first two conditions actually have a feature in common, which is not shared by the other three conditions. Scheffé's method allows a confidence interval to be constructed for the difference between the population means of the two sets of conditions (i.e., for $\mu_{1\&2} - \mu_{3\&4\&5}$), and for any other contrast that might occur to the researcher. If Scheffé's method is used to construct a number (possibly a very large number)

of 95% confidence intervals from the data of a study, then the method ensures that in the long run for at most 5% of such studies would any of the calculated intervals fail to include its true population value.

Like Tukey's Honestly Significant Difference test, Scheffé's test (*see Multiple Comparison Procedures*) is usually applied in order to test nil null hypotheses (*see Confidence Intervals*). However, it can be applied to test whether two sets of conditions have means that differ by some value other than zero, for example, $H_0 : \mu_{1\&2} - \mu_{3\&4\&5} = 3$. Any hypothesized difference that is not included in the calculated simultaneous 95% confidence interval would be rejected at the 0.05 level by a Scheffé test.

References

- [1] Kendall, M.G., Stuart, A. & Ord, J.K. (1983). *The Advanced Theory of Statistics, Vol. 3, Design and Analysis, and Time-Series*, 4th Edition, Griffin, London.
- [2] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

CHRIS DRACUP

Single-Case Designs

PATRICK ONGHENA

Volume 4, pp. 1850–1854

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Single-Case Designs

Single-case designs refer to research designs that are applied to experiments in which one entity is observed repeatedly during a certain period of time, under different levels ('treatments') of at least one independent variable. The essential characteristics of such single-case experiments are (a) that only one entity is involved (single-case), and (b) that there is a manipulation of the independent variable(s) (experiment). These characteristics imply that (c) the entity is exposed to all levels of the independent variable (like in a within-subject design), and (d) that there are repeated measures or observations (like in a longitudinal or time series design). In this characterization of single-case designs, we use the generic term 'entity' to emphasize that, although 'single-case' is frequently equated with 'single-subject', also experiments on a single school, a single family or any other specified 'research unit' fit the description [1, 15, 18, 20].

A Long Tradition and a Growing Impact

Single-case designs have a long history in behavioral science. Ebbinghaus' pivotal memory research and Stratton's study on the effect of wearing inverting lenses are classical nineteenth century experiments involving a single participant that had a tremendous impact on psychology [6]. During the twentieth century, the influential work of Skinner [27] and Sidman [26] provided the major impetus for continued interest in these kind of designs, and it was their methodological approach that laid the foundations of the current popularity of single-case research in behavior modification and clinical psychology [1, 18], neuropsychology [3, 31], psychopharmacology [4, 5], and educational research [20, 22].

In these areas, single-case research is evidently one of the only viable options if rare or unique conditions are involved, but it also seems to be the research strategy of first choice in research settings where between-entity variability is considered negligible or if demonstration in a single case is sufficient to confirm the existence of a phenomenon or to refute a supposedly universal principle. Another potential motivation to embark upon single-case research is its (initial) focus on the single case, which mimics

the care for the individual patient that is needed in clinical work. In such applied settings, generalization is sometimes only a secondary purpose, which can be achieved by replication and aggregation of single-case results. In addition, the replicated single-case designs model is much more consistent with the way in which consecutive patients are entered into **clinical trials** than the random sampling model underlying many group designs and standard statistical techniques.

Single-case Designs, Case Studies, and Time Series

Single-case research using experimental designs should not be confused with case study research or observational time series research. In a traditional case study approach, a single phenomenon is also studied intensively, but there is not necessarily a purposive manipulation of an independent variable nor are there necessarily repeated measures (*see Case Studies*). Furthermore, most case studies are reported in a narrative way while results of single-case experiments usually are presented numerically or graphically, possibly accompanied by sophisticated statistical analyses. In observational time series research, there are also repeated measures and very often complex statistical analyses, but the main difference with single-case experiments lies in the absence of a designed intervention. Although time series **intervention analyses** can be applied to results from simply designed **interrupted time series** experiments (if the number of observations is large enough), a designed intervention is not crucial for time series research. Incidentally, the vast majority of applications of **time series analysis** concern mere observational series.

To Randomize or not to Randomize

An important consideration in designing a single-case experiment is whether or not to randomize (i.e., to randomly assign measurement occasions to treatments) [10]. This randomization provides statistical control over both known and unknown **confounding variables** that are time-related (e.g., 'history' and 'maturation'), and very naturally leads to a statistical test based on the randomization as it was

implemented in the design, a so-called **randomization test** [8, 9, 28]. In this way, randomization can improve both the internal and the statistical-conclusion validity of the study (*see Internal Validity*). In a nonrandomized single-case experiment (e.g., in an operant response-guided experiment), one has to be very cautious when attributing outcome changes to treatment changes. In a nonrandomized intervention study, for example, one usually does not control the variables that covary with the intervention, or with the decision to intervene, making it very difficult to rule out response-guided biases [28] or **regression artifacts**. Therefore, the control aspect of randomization might be considered as essential to single-case experiments as it is to multiple-case experiments and **clinical trials** [10, 11, 28].

We can distinguish two important schedules by which randomization can be incorporated into the design of a single-case experiment. In the first schedule, the treatment alternation is randomly determined and this gives rise to the so-called *alternation designs*. In the second schedule, the moment of intervention is randomly determined, and this gives rise to the so-called *phase designs*. Both randomized alternation and randomized phase designs will be presented in the next two sections, followed by a discussion of two types of replications: simultaneous and sequential replications (*see Interrupted Time Series Design*).

Randomized Alternation Designs

In alternation designs, any level of the independent variable could be present at each measurement occasion. For example, in the *completely randomized single-case design*, the treatment sequence is randomly determined only taking account of the number of levels of the independent variable and the number of measurement occasions for each level. If there are two levels (A and B), with three measurement occasions each, then complete randomization implies a random selection among twenty possible assignments.

If some sequences of a complete randomization are undesirable (e.g., AAABBB), then other families of alternation designs can be devised by applying the classical randomization schemes known from group designs. For example, a *randomized block single-case design* is obtained in the previous setting if the treatments are paired, with random determination

of the order of the two members of the pair. The selection occurs among the following eight possibilities: ABABAB, ABABBA, ABBAAB, ABBABA, BABABA, BABAAB, BAABBA, and BAABAB.

Although this randomized block single-case design is rampant in double-blind single-patient medication trials [16, 17, 23], it is overly restrictive if one only wants to avoid sequences of identical treatments. Therefore, a randomized version of the **alternating treatments design** was proposed, along with an algorithm to enumerate and randomly sample the set of acceptable sequences [25]. For the previous example, a constraint of two consecutive identical treatments at most results in six possibilities in addition to the eight possibilities listed for the randomized block single-case design: AABABB, AABBAB, ABAABB, BBABAA, BBAABA, and BABBAA.

This larger set of potential randomizations is of paramount importance to single-case experiments because the smallest *P* value that can be obtained with the corresponding randomization test is the inverse of the cardinality of this set. If the set is too small, then the experiments have zero statistical **power**. Randomized alternating treatments designs may guarantee sufficient power to detect treatment effects while avoiding awkward sequences of identical treatments [12, 25].

It should be noted that more complex alternation designs also can be constructed if there are two or more independent variables. For example, a *completely randomized factorial single-case design* follows on from crossing the levels of the independent variables involved.

Randomized Phase Designs

If rapid and frequent alternation of treatments is prohibitive, then researchers may opt for a phase design. In phase designs, the complete series of measurement occasions is divided into treatment phases and several consecutive measurements are taken in each phase. The simplest phase design is the *AB design* or the basic **interrupted time series design**. In this design, the first phase of consecutive measurements is taken under one condition (e.g., a baseline or control condition) and the second phase under another condition (e.g., a postintervention or treatment condition). All sorts of variations and extensions of this AB design can be conceived of:

ABA or *withdrawal and reversal designs*, ABAB, ABABACA designs, and so on [1, 15, 18, 20].

In phase designs, the order of the phases is fixed, so the randomization cannot be applied to the treatment sequence like in alternation designs. However, there is one feature that can be randomized without distorting the phase order and that is the moment of phase change (or the moment of intervention). In such randomized phase designs, only the number of available measurement occasions, the number of phases (c.q., treatments), and the minimum lengths of the phases should be specified [9, 24]. A randomized AB design with six measurement occasions and with at least one measurement occasion in each phase, for example, implies the following five possibilities: ABBBBB, AABBBB, AAABBB, AAAABB, and AAAAAB. There are, of course, many more repeated measurements in the phase designs of typical applications (e.g., by using a diary or psychophysical measures) and often it is possible to add phases and thereby increase the number of phase changes. In fact, a large number of measurement occasions and/or more than one phase change is necessary to obtain sufficient statistical power in these designs [14, 24].

Simultaneous and Sequential Replication Designs

Replication is the obvious strategy for demonstrating or testing generalizability of single-case results. If the replications are planned and part of the design, then the researcher has the option to conduct the experiments at the same time or conduct them one by one.

Simultaneous replication designs are the designs in which the replications (alternation or phase single-case designs) are carried out at the same time. The most familiar simultaneous replication design is the *multiple baseline across participants design* (see **Multiple Baseline Designs**). In such a design, several AB phase designs are implemented simultaneously and the intervention is applied for each separate participant at a different moment [1]. The purpose of the simultaneous monitoring is to control for historical confounding variables. If an intervention is introduced in one of the phase designs and produces a change for that participant, while little or no change is observed for the other participants, then it is less likely that other external events are responsible for

the observed change, than if this change was observed in an isolated phase design.

Randomization can be introduced very easily in simultaneous replication designs by just applying the randomization schedules in the several phase and/or alternation designs separately [24]. In addition, between-case constraints can be imposed, for example, to avoid simultaneous intervention points or to obtain systematic staggering of the intervention in the multiple baseline across participants design [19].

If the replications are carried out one by one, then we have a *sequential replication design*. Also for these designs, a lot of options are available to the applied researcher, depending on whether the separate designs should be very similar (e.g., the same number of measurement occasions for all or some of the designs) and whether between-case comparisons are part of the design [22].

The statistical power of the corresponding randomization tests for both simultaneous and sequential replication designs is already adequate (>0.80) for designs with four participants and a total of twenty measurement occasions (for a range of effect sizes (see **Effect Size Measures**), autocorrelations and significance levels likely to be relevant for behavioral research) [13,21]. In addition, the sequential replication design also provides an opportunity to apply powerful nonparametric or parametric meta-analytic procedures (see **Meta-Analysis**) [2, 7, 24, 29, 30].

References

- [1] Barlow, D.H. & Hersen, M., eds (1984). *Single-Case Experimental Designs: Strategies for Studying Behavior Change*, 2nd Edition, Pergamon Press, Oxford.
- [2] Busk, P.L. & Serlin, R.C. (1992). Meta-analysis for single-case research, in *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, T.R. Kratochwill & J.R. Levin, eds, Lawrence Erlbaum, Hillsdale, pp. 187–212.
- [3] Caramazza, A., ed. (1990). *Cognitive Neuropsychology and Neurolinguistics: Advances in Models of Cognitive Function and Impairment*, Lawrence Erlbaum, Hillsdale.
- [4] Conners, C.K. & Wells, K.C. (1982). Single-case designs in psychopharmacology, in *New Directions for Methodology of Social and Behavioral Sciences*, Vol. 13, *Single-Case Research Designs*, A.E. Kazdin & A.H. Tuma, eds, Jossey-Bass, San Francisco, pp. 61–77.
- [5] Cook, D.J. (1996). Randomized trials in single subjects: the N of 1 study, *Psychopharmacology Bulletin* 32, 363–367.
- [6] Dukes, W.F. (1965). $N = 1$, *Psychological Bulletin* 64, 74–79.

- [7] Edgington, E.S. (1972). An additive method for combining probability values from independent experiments, *Journal of Psychology* **80**, 351–363.
- [8] Edgington, E.S. (1986). Randomization tests, in *Encyclopedia of statistical sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 530–538.
- [9] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Dekker, New York.
- [10] Edgington, E.S. (1996). Randomized single-subject experimental designs, *Behaviour Research and Therapy* **34**, 567–574.
- [11] Ferron, J., Foster-Johnson, L. & Kromrey, J.D. (2003). The functioning of single-case randomization tests with and without random assignment, *Journal of Experimental Education* **71**, 267–288.
- [12] Ferron, J. & Onghena, P. (1996). The power of randomization tests for single-case phase designs, *Journal of Experimental Education* **64**, 231–239.
- [13] Ferron, J. & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs, *Journal of Experimental Education* **70**, 165–178.
- [14] Ferron, J. & Ware, W. (1995). Analyzing single-case data: the power of randomization tests, *Journal of Experimental Education* **63**, 167–178.
- [15] Franklin, R.D., Allison, D.B., & Gorman, B.S., eds (1996). *Design and Analysis of Single-Case Research*. Lawrence Erlbaum, Mahwah.
- [16] Gracious, B. & Wisner, K.L. (1991). Nortriptyline in chronic fatigue syndrome: a double blind, placebo-controlled single-case study, *Biological Psychiatry* **30**, 405–408.
- [17] Guyatt, G.H., Keller, J.L., Jaeschke, R., Rosenbloom, D., Adachi, J.D. & Newhouse, M.T. (1990). The n-of-1 randomized controlled trial: clinical usefulness—our three-year experience, *Annals of Internal Medicine* **112**, 293–299.
- [18] Kazdin, A.E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings*, Oxford University Press, New York.
- [19] Koehler, M.J. & Levin, J.R. (1998). Regulated randomization: a potentially sharper analytical tool for the multiple-baseline design, *Psychological Methods* **3**, 206–217.
- [20] Kratochwill, T.R. & Levin, J.R., eds (1992). *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, Lawrence Erlbaum, Hillsdale.
- [21] Lall, V.F. & Levin, J.R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests, *Journal of School Psychology* **42**, 61–86.
- [22] Levin, J.R. & Wampold, B.E. (1999). Generalized single-case randomization tests: flexible analyses for a variety of situations, *School Psychology Quarterly* **14**, 59–93.
- [23] Mahon, J., Laupacis, A., Donner, A. & Wood, T. (1996). Randomised study of n of 1 trials versus standard practice, *British Medical Journal* **312**, 1069–1074.
- [24] Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder, *Behavioral Assessment* **14**, 153–171.
- [25] Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs, *Behaviour Research and Therapy* **32**, 783–786.
- [26] Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*, Basic Books, New York.
- [27] Skinner, B.F. (1938). *The Behavior of Organisms: An Experimental Analysis*, Appleton-Century-Crofts, New York.
- [28] Todman, J.B. & Dugard, P. (2001). *Single-Case and Small-n Experimental Designs: A Practical Guide to Randomization Tests*, Lawrence Erlbaum, Mahwah.
- [29] Van den Noortgate, W. & Onghena, P. (2003). Combining single-case experimental studies using hierarchical linear models, *School Psychology Quarterly* **18**, 325–346.
- [30] Van den Noortgate, W. & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research, *Behavior Research Methods, Instruments, & Computers* **35**, 1–10.
- [31] Wilson, B. (1987). Single-case experimental designs in neuropsychological rehabilitation, *Journal of Clinical and Experimental Neuropsychology* **9**, 527–544.

PATRICK ONGHENA

Single and Double-blind Procedures

ARNOLD D. WELL

Volume 4, pp. 1854–1855

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Single and Double-blind Procedures

Suppose we want to test the effectiveness of a drug administered in the form of a pill. We could randomly assign patients to two groups, a treatment group that received the drug and a control group that did not, and then measure changes in the patients (*see Clinical Trials and Intervention Studies*). However, this design would not be adequate to evaluate the drug because we could not be sure that any changes we observed in the treatment condition were a result of the drug itself. It is well known that patients may derive some benefit simply from the belief that they are being treated, quite apart from any effects of the active ingredients in the pill.

The study could be improved by giving patients in the control group a **placebo**, pills that have the same appearance, weight, smell, and taste as those used in the drug condition but do not contain any active ingredients. Our goal would be to have a situation in which any systematic differences between groups were due only to the effects of the drug itself. We could assign patients randomly to the groups so that there were no systematic differences in patient characteristics between them. However, in order to rule out possible bias resulting from patients' expectations, it is important that they not be informed as to whether they have been assigned to the drug or the placebo/control condition. By not informing patients about condition assignments, we are employing what is called a *single-blind procedure*.

Another possible source of bias may occur because of what may be referred to as experimenter effects (*see Expectancy Effect by Experimenters*). The

expectations and hopes of researchers may influence the data by causing subtle but systematic differences in how the researchers interact with patients in different conditions, as well as how the data are recorded and analyzed. Moreover, not only do the patient and experimenter effects described above constitute possible sources of bias that might threaten the validity of the research, the two types of bias can interact. Even if the patients are not explicitly told about the details of the research design and condition assignments, they may learn about them through interactions with the research personnel who are aware of this information, thereby reintroducing the possibility of bias due to patient expectations. If we try to rule out these sources of bias by withholding information from both subjects and researchers, we are said to be using a *double-blind procedure*. The purpose of double-blind procedures is to rule out any bias that might result from knowledge about the experimental conditions or the purpose of the study by making both participants and researchers 'blind' to such information.

These ideas generalize to many kinds of behavioral research. The behavior of participants and researchers may be influenced by many factors other than the independent variable – see for example, [1]. Factors such as information or misinformation about the purpose of the experiment may combine with motivation to obtain particular findings, or to please the experimenter, thereby causing the data to differ from what they would be without this information.

Reference

- [1] Rosenthal, R. & Rosnow, R.L. (1969). *Artifacts in Behavioral Research*, Academic Press, New York.

ARNOLD D. WELL

Skewness

KARL L. WUENSCH

Volume 4, pp. 1855–1856

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Skewness

In everyday language, the terms ‘skewed’ and ‘askew’ are used to refer to something that is out of line or distorted on one side. When referring to the shape of frequency or probability distributions, ‘skewness’ refers to asymmetry of the distribution. A distribution with an asymmetric tail extending out to the right is referred to as ‘positively skewed’ or ‘skewed to the right’, while a distribution with an asymmetric tail extending out to the left is referred to as ‘negatively skewed’ or ‘skewed to the left’.

Karl Pearson [2] first suggested measuring skewness by standardizing the difference between the mean and the mode, that is, $sk = (\mu - \text{mode})/\sigma$. Population modes are not well estimated from sample modes, but one can estimate the difference between the mean and the mode as being three times the difference between the mean and the median [3], leading to the following estimate of skewness: $sk_{\text{est}} = (3(M - \text{median}))/s$. Some statisticians use this measure but with the ‘3’ eliminated.

Skewness has also been defined with respect to the third moment about the mean: $\gamma_1 = (\sum (X - \mu)^3)/n\sigma^3$, which is simply the expected value of the distribution of cubed z scores. Skewness measured in this way is sometimes referred to as ‘Fisher’s skewness’. When the deviations from the mean are greater in one direction than in the other direction, this statistic will deviate from zero in the direction of the larger deviations. From sample data, Fisher’s skewness is most often estimated by $g_1 = (n \sum z^3)/((n - 1)(n - 2))$. For large sample sizes ($n > 150$), g_1 may be distributed approximately normally, with a standard error of approximately $\sqrt{6/n}$. While one could use this sampling distribution to construct confidence

intervals for or tests of hypotheses about γ_1 , there is rarely any value in doing so.

It is important for behavioral researchers to notice skewness when it appears in their data. Great skewness may motivate the researcher to investigate outliers. When making decisions about which measure of location to report (means being drawn in the direction of the skew) and which inferential statistic to employ (one that assumes normality or one that does not), one should take into consideration the estimated skewness of the population. Normal distributions have zero skewness. Of course, a distribution can be perfectly symmetric but far from normal. Transformations commonly employed to reduce (positive) skewness include square root, log, and reciprocal transformations.

The most commonly used measures of skewness (those discussed here) may produce surprising results, such as a negative value when the shape of the distribution appears skewed to the right. There may be superior alternative measures not in common use [1].

References

- [1] Groeneveld, R.A. & Meeden, G. (1984). Measuring skewness and kurtosis, *The Statistician* **33**, 391–399.
- [2] Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London* **186**, 343–414.
- [3] Stuart, A. & Ord, J.K. (1994). *Kendall’s Advanced Theory of Statistics, Vol. 1, Distribution Theory*, 6th Edition, Edward Arnold, London.

(See also **Kurtosis**)

KARL L. WUENSCH

Slicing Inverse Regression

CHUN-HOUH CHEN

Volume 4, pp. 1856–1863

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Slicing Inverse Regression

Dimensionality and Effective Dimension Reduction Space

Dimensionality is an issue that often arises and sets a severe limitation in the study of every scientific field. In the routine practice of regression analysis (*see Multiple Linear Regression*), the curse of dimensionality [11] may come in at the early exploratory stage. For example, a 2-D or **3-D scatterplot** can be successfully applied to examine the relationship between the response variable and one or two input variables. But, when the dimension of regressors gets larger, this graphical approach could become laborious, and it is important to focus only on a selective set of projection directions. In the parametric regression setting, simple algebraic functions of $\mathbf{x}$ are used to construct a link function for applying the **least squares** or **maximum likelihood** methods. In the **nonparametric regression** setting, the class of fitted functions is enlarged. The increased flexibility in fitting via computation intensive smoothing techniques, however, also increases the modeling difficulties that are often encountered with larger number of regressors.

Li [12] introduced the following framework for dimension reduction in regression:

$$Y = g(\beta'_1 \mathbf{x}, \dots, \beta'_k \mathbf{x}, \varepsilon). \quad (1)$$

The main feature of (1) is that g is completely unknown and so is the distribution of ε , which is independent of the p -dimensional regressor $\mathbf{x}$. When k is smaller than p , (1) imposes a dimension reduction structure by claiming that the dependence of Y on $\mathbf{x}$ only comes from the k variates, $\beta'_1 \mathbf{x}, \dots, \beta'_k \mathbf{x}$, but the exact form of the dependence structure is not specified. Li called this k -dimensional space spanned by the k β vectors the e.d.r. (effective dimension reduction) space and any vector in this space is referred to as an e.d.r. direction. The aim is to estimate the base vectors of the e.d.r. space. The notion of e.d.r. space and its role in regression graphics are further explored in Cook [4]. The primary goal of Li's approach is to estimate the e.d.r. directions first so that it becomes easier to explore data further with either the graphical approach or the nonparametric curve-smoothing techniques.

Special Cases of Model (1)

Many commonly used models in regression can be considered as special cases of model (1). We separate them into one-component models ($k = 1$) and the multiple-component models ($k > 1$). One-component models ($k = 1$) include the following:

1. Multiple linear regression. $g(\beta' \mathbf{x}, \varepsilon) = a + \beta' \mathbf{x} + \varepsilon$.
2. Box-Cox transformation. $g(\beta' \mathbf{x}, \varepsilon) = h_\lambda(a + \beta' \mathbf{x} + \varepsilon)$, where $h_\lambda(\cdot)$ is the power transformation function with power parameter λ given by

$$h_\lambda(t) = \begin{cases} (t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \ln(t) & \text{if } \lambda = 0. \end{cases} \quad (2)$$

3. Additive error models. $g(\beta' \mathbf{x}, \varepsilon) = h(\beta' \mathbf{x}) + \varepsilon$, where $h(\cdot)$ is unknown.
4. Multiplicative error models. $g(\beta' \mathbf{x}, \varepsilon) = \mu + \varepsilon h(\beta' \mathbf{x})$, where $h(\cdot)$ is usually assumed to be known.

Multiple-component models ($k > 1$) include the following:

5. **Projection pursuit regression** (Friedman and Stuetzle [8]). $g(\beta'_1 \mathbf{x}, \dots, \beta'_k \mathbf{x}, \varepsilon) = h_1(\beta'_1 \mathbf{x}) + \dots + h_r(\beta'_r \mathbf{x}) + \varepsilon$, where r may be unequal to k .
6. Heterogeneous error models. $g(\beta'_1 \mathbf{x}, \beta'_2 \mathbf{x}, \varepsilon) = h_1(\beta'_1 \mathbf{x}) + \varepsilon h_2(\beta'_2 \mathbf{x})$.

More detailed discussions about these models can be found in [4, 15].

An Example

The following example will be used to illustrate the concept and implementation of SIR throughout this article. Six independent standard normal random variables, $\mathbf{x} = (x_1, \dots, x_6)$, with 200 observations each are generated. The response variable Y is generated according to the following two-component model:

$$\begin{aligned} y &= g(\beta'_1 \mathbf{x}, \beta'_2 \mathbf{x}, \varepsilon) \\ &= \frac{\beta'_1 \mathbf{x}}{0.5 + (\beta'_2 \mathbf{x} + 1.5)^2} + 0 \cdot \varepsilon, \end{aligned} \quad (3)$$

where $\beta'_1 = (1, 1, 0, 0, 0, 0)$ and $\beta'_2 = (0, 0, 1, 1, 0, 0)$. We employ this noise free model for an easier explanation.

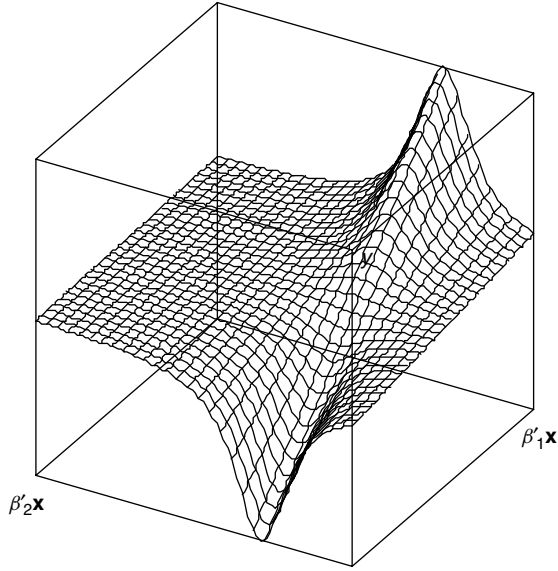


Figure 1 The response surface for function in (2)

Contour Plots and Scatterplot Matrix

The response surface of (2) is depicted in Figure 1. A different way to visualize the structure of the response

variable Y is to overlay the scatter plot of $\beta'_1 \mathbf{x}$ and $\beta'_2 \mathbf{x}$ with the contours of (3), (Figure 2). Because the vectors $\beta'_1 = (1, 1, 0, 0, 0, 0)$ and $\beta'_2 = (0, 0, 1, 1, 0, 0)$ are not given, how to identify these components is the most challenging issue in a regression problem. One possible way of constructing a contour plot of the response variable on the input variables is through the scatter-plot matrix of paired variables in $\mathbf{x}$. Here, we show only three scatter plots of (x_1, x_2) , (x_3, x_4) , and (x_5, x_6) . The upper panel of Figure 3 gives the standard scatter plots of the three paired variables. Since they are all independently generated, no interesting information can be extracted from these plots.

We can bring in the contour information about the response variable to these static scatter plots through color linkage. However, because of the black–white printing nature of the Encyclopedia, the range information about Y is coded in black and white. In the lower panel of Figure 3, each point in the scatter plots shows the relative intensity of the corresponding magnitude of the response variable Y . The linear structure of Y relative to (x_1, x_2) (the first component) can be easily detected from this plot. There also appears to be some nonlinear structure in the scatterplot of (x_3, x_4) (the second component);

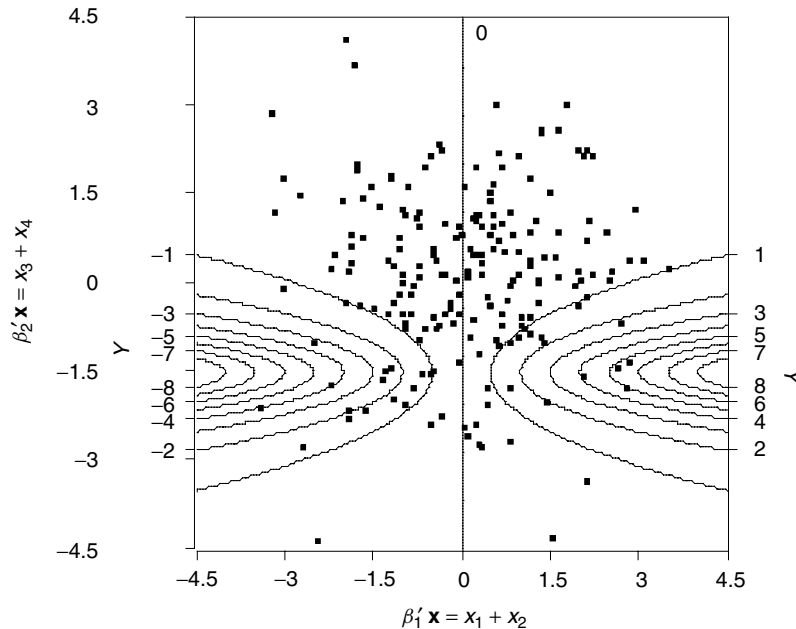


Figure 2 Scatter plot of $\beta'_1 \mathbf{x}$ and $\beta'_2 \mathbf{x}$ with contours of Y from (2)

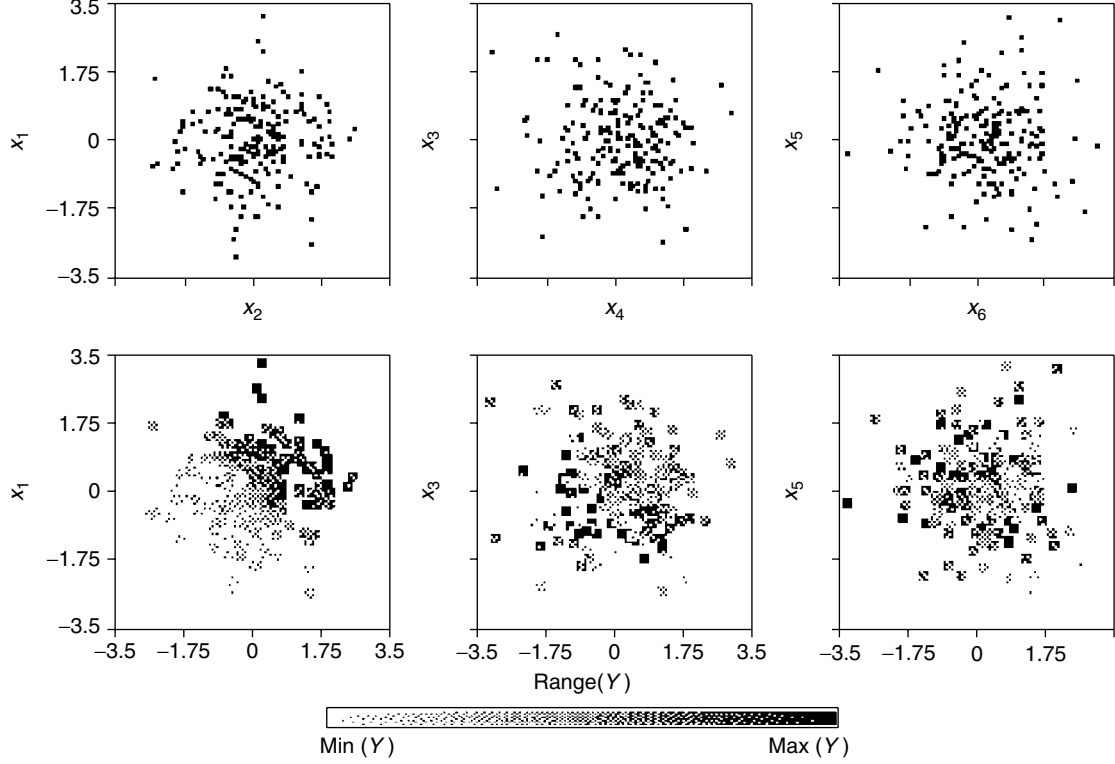


Figure 3 Upper panel: Standard scatter plots of (x_1, x_2) , (x_3, x_4) , and (x_5, x_6) . Lower panel: Same sets of scatter plots as in upper panel except that each point is coded with the relative intensity showing the corresponding value of the response variable Y

but it is not as visible as the first component. No interesting pattern can be identified from plots related to (x_5, x_6) as one would expect.

Principal Component Analysis (PCA) and Multiple Linear Regression (MLR)

For our regression problem, the matrix map of the raw data matrix $(Y, \mathbf{x})$ is plotted as Figure 4(a). In a matrix map, each numerical value in a matrix is represented by a color dot (gray shade). Owing to lack of color printing, ranges for all the variables, $(Y, \mathbf{x})$ are linearly scaled to $[0, 1]$ and coded in a white to black-gray spectrum. Please see Chen et al. [3] for an introduction to matrix map visualization.

Even for regression problems, **principal component analysis** (PCA) is often used to reduce the dimensionality of $\mathbf{x}$. This is not going to work for our example since all the input variables are independently generated. A PCA analysis utilizes only

the variation information contained in the input variables. No information about the response variable Y is taken into consideration in constructing the **eigenvalue** decomposition of the covariance matrix of $\mathbf{x}$ (see **Correlation and Covariance Matrices**). The sample covariance matrix for the six input variables in the example is also plotted as a matrix map in Figure 4(b). Since these variables are generated independently with equal variances, no structure with interesting pattern is anticipated from this map. The PCA analysis will produce six principal components with nearly equal eigenvalues. Thus, no reduction of dimension can be achieved from the performance of a PCA analysis on this example. What the SIR analysis does can be considered as a weighted PCA analysis that takes the information of response variable Y into account.

Multiple linear regression (MLR) analysis, on the other hand, can pick up some partial information from the two-component model in (2). MLR studies

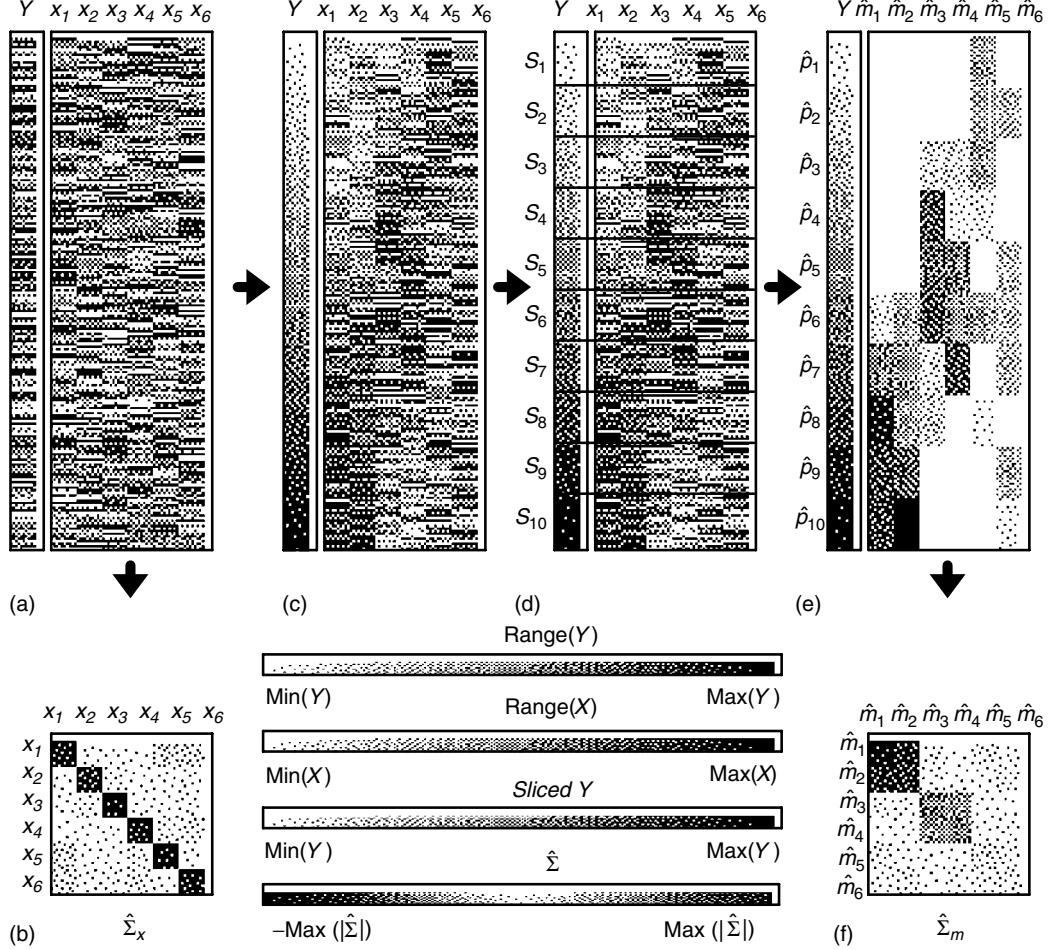


Figure 4 Matrix map of the raw data matrix ($Y, \mathbf{x}$) with a PCA analysis and the SIR algorithm. (a) Original (unsorted) matrix map. (b) Sample covariance matrix of $\mathbf{x}$ in (a), $\hat{\Sigma}_x$. (c) Sorted (by rank of Y) map. (d) Sliced sorted map. (e) Map for sliced mean matrix, $\hat{m}$. (f) Sample covariance matrix of $\hat{m}$, $\hat{\Sigma}_m$.

the linear relationship of a linear combination of the input variables $\mathbf{x}$ to the response variable Y . For our example, MLR will identify the linear relationship of Y to the first component $\beta'_1 \mathbf{x}$, but not the nonlinear structure of the second component $\beta'_2 \mathbf{x}$ on Y .

Implementation and Theoretical Foundation of SIR

Inverse Regression

Conventional functional-approximation and curve-smoothing methods regress Y against $\mathbf{x}$ (forward

regression, $E(Y|\mathbf{x})$). The contour plot in Figure 2 and gray-shaded scatter plots in Figure 3 give the hint to the basic concept of SIR; namely, to reverse the role of $\mathbf{x}$ and Y as in the general forward regression setup. We treat Y as if it were the independent variable and treat $\mathbf{x}$ as if it were the dependent variable. SIR estimates the e.d.r. directions based on inverse regression. The inverse regression curve $\eta(y) = E(\mathbf{x}|Y = y)$ is composed of p simple regressions, $E(x_j|y)$, $j = 1, \dots, p$. Thus, one essentially deals with p one-dimension to one-dimension regression problems, rather than a high-dimensional forward regression problem. Instead of asking the

question ‘given $\mathbf{x} = \mathbf{x}_o$, what value will Y take’? in the forward regression framework, SIR rephrase the problem as ‘given $Y = y$, what values will $\mathbf{x}$ take’? Instead of local smoothing, SIR intends to gain global insight on how Y changes as $\mathbf{x}$ changes by studying the reverse – how does the associated $\mathbf{x}$ region vary as Y varies.

The SIR Algorithm

Following are the steps in conducting the SIR analysis on a random sample $(Y_i, \mathbf{x}_i), i = 1, \dots, n$. Figure 4(a) gives the matrix map of our example with $n = 200$ and $p = 6$. No interesting pattern is expected from Figure 4(a) because the observations are listed in a random manner.

1. Sort the n observations $(Y_i, \mathbf{x}_i), i = 1, \dots, n$ according to the magnitudes of Y_i 's. For our example, Figure 4(c) shows a smoothed spectrum of the ranks of Y_i s with corresponding linear relationship of (x_1, x_2) and nonlinear structure of (x_3, x_4) . The sorted (x_5, x_6) do not carry information on the Y_i s.
2. Divide the range of Y into H slices S_h , for $h = 1, \dots, H$. Let $\hat{p}_h$ ($= 0.1$ in this example) be the proportion of Y_i s falling into the h th slice. $H = 10$ slices are used in our example, yielding 20 observations per slice.
3. Compute the sample mean of the $\mathbf{x}_i$ s for each slice, $\hat{\mathbf{m}}_h = (n\hat{p}_h)^{-1} \sum_{Y_i \in S_h} \mathbf{x}_i$, and form the weighted covariance matrix, $\hat{\Sigma}_m = \sum_{h=1}^H \hat{p}_h (\hat{\mathbf{m}}_h - \bar{\mathbf{x}})(\hat{\mathbf{m}}_h - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is the sample mean of all $\mathbf{x}_i$ s. Figure 4(e) gives the matrix map of the corresponding slice means. The linear and nonlinear structures of the Y_i 's on (x_1, x_2, x_3, x_4) are even more apparent now.
4. Estimate the covariance matrix of $\mathbf{x}$ with

$$\hat{\Sigma}_{\mathbf{x}} = n^{-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}}). \quad (4)$$

Find the SIR directions by conducting the eigenvalue decomposition of $\hat{\Sigma}_m$ with respect to $\hat{\Sigma}_{\mathbf{x}}$: for $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$, solve

$$\hat{\Sigma}_m \mathbf{b}_j = \hat{\lambda}_j \hat{\Sigma}_{\mathbf{x}} \mathbf{b}_j, \quad j = 1, \dots, p. \quad (5)$$

The weighted covariance matrix $\hat{\Sigma}_m$ in Figure 4(f) shows a strong two-component structure

Table 1 The first two eigenvectors (with standard deviations and ratios) and eigenvalues with P values of SIR for model (2)

First vector	(−0.72 −0.68 −0.01 −0.01 0.07 0.01)
S.D.	(0.04 0.04 0.04 0.04 0.04 0.04)
Ratio	(−17.3 −16.1 −0.2 −0.1 1.9 0.3)
Second vector	(−0.01 0.08 −0.67 −0.72 0.05 −0.02)
S.D.	(0.09 0.10 0.09 0.09 0.09 0.09)
Ratio	(−0.2 0.9 −7.6 −7.7 0.5 −0.2)
Eigenvalues	(0.76 0.38 0.06 0.03 0.02 0.02)
P values	(0.0 3.8E-7 0.62 NA NA NA)

compared to that of the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ in Figure 4(b).

5. The i th eigenvector $\mathbf{b}_i$ is called the i th SIR direction. The first few (two for the example) SIR directions can be used for dimension reduction.

The standard SIR output of the example is summarized in Table 1 along with graphical comparisons of Y against two true e.d.r. directions with two estimated directions illustrated in Figure 5.

Some Heuristics

Let us take another look at the scatter plot of $\beta'_1 \mathbf{x}$ and $\beta'_2 \mathbf{x}$ along with the contours of Y in Figure 2. When the range of Y is divided into H slices, it is possible to use either equal number of observations per slice or equal length of interval per slice. We took the former option here, yielding 20 observations per slice. This is illustrated in Figure 6(a), with the contour lines drawn from the generated Y_i s. The mean of the 20 observations contained in a slice is marked by a square with a number indicating which slice it comes from. The location variation of these sliced means along these two components suggests that the slicing-mean step of SIR actually is exploiting the relationship structure of Y against the two correct components of $\mathbf{x}$. This further leads to the well-structured sliced mean matrix shown in Figure 4(e) and the two-component weighted covariance matrix $\hat{\Sigma}_m$ shown in Figure 4(f).

Theoretical Foundation of SIR

The computation of SIR is simple and straightforward. The first three steps of the SIR algorithm

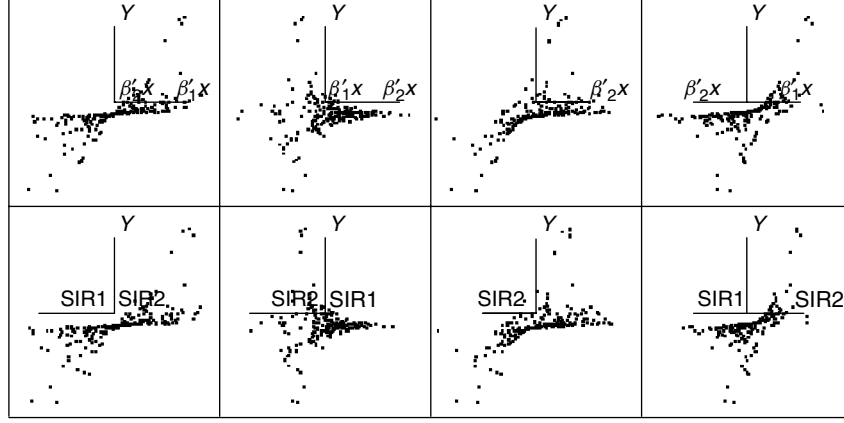


Figure 5 True view (upper panel) and SIR view (lower panel) of model (2)

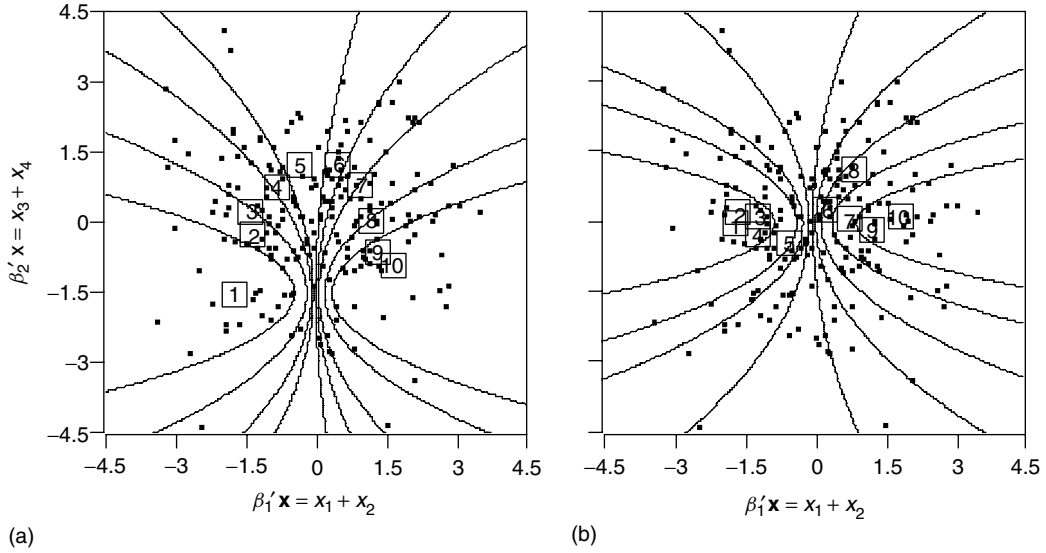


Figure 6 The scatter plot of $\beta'_1 \mathbf{x}$ and $\beta'_2 \mathbf{x}$ with the contours of Y_i s. (a) Y_i s generated from model (2). (b) Y_i s generated from model (4)

produce a crude estimate of the inverse regression curve $\eta(y) = E(\mathbf{x}|Y = y)$ through step functions from $\hat{m}_h$, $h = 1, \dots, H$. The SIR theorem (Theorem 3.1, [12]) states that under the dimension reduction assumption in model (1), the centered inverse regression curve $\eta(y) - E(\mathbf{x})$ is contained in the linear subspace spanned by $\Sigma_{\mathbf{x}} \beta_i$, $i = 1, \dots, k$, provided a linear design condition (Condition 3.1, [12]) on the distribution of $\mathbf{x}$ holds. When this is the case, the covariance matrix of $\eta(Y)$ can be written as a

linear combination of $\Sigma_{\mathbf{x}} \beta_i \beta_i' \Sigma_{\mathbf{x}}$, $i = 1, \dots, k$. Thus, any eigenvector b_i with nonzero eigenvalue λ_i from the eigenvalue decomposition

$$\text{cov}[\eta(Y)] b_i = \lambda_i \Sigma_{\mathbf{x}} b_i \quad (6)$$

must fall into the e.d.r. space. Now, because the covariance matrix of the slice average, $\hat{\Sigma}_m$, gives an estimate of $\text{cov}[\eta(Y)]$, the fourth step of the SIR algorithm is just a sample version of (6). It

is noteworthy that more sophisticated nonparametric regression methods, such as kernel, nearest neighbor, or smoothing splines can be used to yield a better estimate of the inverse regression curve.

On the basis of the theorem, SIR estimates have been shown to be root- n consistent. They are not sensitive to the number of slices used. Significance tests are available for determining the dimensionality. Further discussions on the theoretical foundation of SIR can be found in Brillinger [1], Chen and Li [2], Cook and Weisberg [5], Cook and Wetzel [6], Duan and Li [7], Hall and Li [9], Hsing and Carroll [10], Li [12, 13], Li and Duan [16], Schott [17], Zhu and Fang [18], and Zhu and Ng [19].

Extensions

Regression Graphics and Graphical Regression

One possible problem, but not necessary a disadvantage, for SIR is that it does not attempt to directly formulate (estimate) the function form of g in model (1). Instead, advocates of SIR argue that users can gain better insights about the relationship structure after visualizing how the response variable Y is associated with reduced input variables. This data analysis strategy is a reversal of the standard practice, which relies on model specification. In a high-dimensional situation, without informative graphical input, formal model specification is seldom efficient. See Cook [4] for more detailed discussion on graphical regression.

Limitations and Generalizations of SIR

SIR successfully finds the two true directions of model (2). But, sometimes, it may not work out as well as expected. To investigate the reason behind, let us change (2) to the following:

$$\begin{aligned} y &= g(\beta'_1 \mathbf{x}, \beta'_2 \mathbf{x}, \varepsilon) \\ &= \frac{\beta'_1 \mathbf{x}}{0.5 + (\beta'_2 \mathbf{x})^2} + 0 \cdot \varepsilon. \end{aligned} \quad (7)$$

We generate 200 Y_i s according to (7) while keeping the same set of input variables $\mathbf{x}$ used earlier. The same scatter plot of $\beta'_1 \mathbf{x}$ and $\beta'_2 \mathbf{x}$ is overlaid with the contours of the new Y_i s in Figure 6(b). The symmetric center of each contour is now shifted up to the horizontal axis, resulting in a symmetric

contour structure along the second component $\beta'_2 \mathbf{x}$. This symmetrical pattern causes the slice means to spread only along the first component and SIR fails in identifying the second component. However, SIR can still find the first component.

The above argument holds for any symmetric function form and the inverse regression curve may not span the entire e.d.r. space. One possible remedy to this problem is to use statistics other than the mean in each slice. For example, the covariance matrix from each slice can be computed and compared with each other. From the contour plot in Figure 6(b), we see that the magnitude of variances within each slice does vary along the second direction. This suggests that slicing the covariance matrix may be able to help. Unfortunately, this second moment-based strategy is not as effective as the first moment-based SIR in finding the first component because the slice variances do not change much along this direction. This interesting phenomenon suggests a possible hybrid technique. That is to combine the directions identified by the first moment SIR and second moment SIR in order to form the complete e.d.r. space. There are several variants of SIR-related dimension reduction strategy such as SAVE ([5]) and SIRII ([12]). These procedures are related to the method of principal Hessian directions ([14]).

References

- [1] Brillinger, D.R. (1991). Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association* **86**, 333.
- [2] Chen, C.H. & Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8**, 289–316.
- [3] Chen, C.H., Hwu, H.G., Jang, W.J., Kao, C.H., Tien, Y.J., Tzeng, S. & Wu, H.M. (2004). Matrix visualization and information mining, *Proceedings in Computational Statistics 2004 (Compstat 2004)*, Physika Verlag, Heidelberg.
- [4] Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.
- [5] Cook, R.D. & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association* **86**, 328–333.
- [6] Cook, R.D. & Wetzel, N. (1994). Exploring regression structure with graphics, (with discussion), *Test* **2**, 33–100.
- [7] Duan, N. & Li, K.C. (1991). Slicing regression: a link-free regression method, *The Annals of Statistics* **19**, 505–530.

- [8] Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.
- [9] Hall, P. & Li, K.C. (1993). On almost linearity of low dimensional projection from high dimensional data, *Annals of Statistics* **21**, 867–889.
- [10] Hsing, T. & Carroll, R.J. (1992). An asymptotic theory for sliced inverse regression, *Annals of Statistics* **20**, 1040–1061.
- [11] Huber, P. (1985). Projection pursuit, with discussion, *Annals of Statistics* **13**, 435–526.
- [12] Li, K.C. (1991). Sliced inverse regression for dimension reduction, with discussion, *Journal of the American Statistical Association* **86**, 316–412.
- [13] Li, K.C. (1992a). Uncertainty analysis for mathematical models with SIR, in *Probability and Statistics*, Z.P. Jiang, S.H. Yan, P. Cheng & R. Wu, eds, World Scientific, Singapore, pp. 138–162.
- [14] Li, K.C. (1992b). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma, *Journal of the American Statistical Association* **87**, 1025–1039.
- [15] Li, K.C. (1997). Sliced inverse regression, in *Encyclopedia of Statistical Sciences*, Vol. 1, (update), S. Kotz, C. Read & D. Banks, eds, John Wiley, New York, pp. 497–499.
- [16] Li, K.C. & Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics* **17**, 1009–1052.
- [17] Schott, J.R. (1994). Determining the dimensionality in sliced inverse regression, *Journal of the American Statistical Association* **89**, 141–148.
- [18] Zhu, L.X. & Fang, K.T. (1996). Asymptotics for kernel estimate of sliced inverse regression, *Annals of Statistics* **24**, 1053–1068.
- [19] Zhu, L.X. & Ng, K.W. (1995). Asymptotics of sliced inverse regression, *Statistica Sinica* **5**, 727–736.

CHUN-HOUH CHEN

Snedecor, George Waddell

DAVID C. HOWELL

Volume 4, pp. 1863–1864

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Snedecor, George Waddell

Born: October 20, 1881, in Memphis, USA.

Died: February 20, 1974, in Amherst, USA.

George Snedecor was born in Tennessee in 1882 and graduated from the University of Alabama in 1905 with a B.S. degree in mathematics and physics. He then moved to Michigan State, where he obtained a Master's degree in physics. In 1913, he accepted a position at Iowa State to teach algebra, and soon persuaded his department to allow him to introduce courses in the relatively new field of statistics. He remained at Iowa State until his retirement.

While George Snedecor did not make many major contributions to the theory of statistics, he was one of the field's most influential pioneers. In 1924, he paired with Henry Wallace, later to become vice president under Roosevelt, to jointly publish a manual on machine methods for computational and statistical methods [4]. Three years later, Iowa State formed the Mathematical Statistics Service headed by Snedecor and A. E. Brandt. Then, in 1935, the Iowa Agriculture Experiment Station formed a Statistical Section that later became the Department of Statistics at Iowa State. This was the first department of statistics in the United States. Again, George Snedecor was its head.

Beginning in the early 1930s, Snedecor invited eminent statisticians from Europe to spend summers at Iowa. **R. A. Fisher** was one of the first to come, and he came for several years. His influence on Snedecor's interest in experimental design and the **analysis of variance** was significant, and in 1937, Snedecor published the first of seven editions of his famous *Statistical Methods* [3]. (This work was later written jointly by Snedecor and **W. G. Cochran**, and is still in press.)

As is well-known, R. A. Fisher developed the analysis of variance, and in his 1924 book, included

an early table for evaluating the test statistic. In 1934, Snedecor published his own table of $F = \hat{\sigma}_{\text{Treatment}}^2 / \hat{\sigma}_{\text{Error}}^2$, which derives directly from the calculations of the analysis of variance [2]. He named this statistic F in honor of Fisher, and it retains that name to this day [1].

Snedecor's department included many eminent and influential statisticians of the time, among whom were **Gertrude Cox** and William Cochran, both of whom he personally recruited to Iowa. He was president of the American Statistical Association in 1948, and made an Honorary Fellow of the British Royal Statistical Society (1954). Like many other major figures in statistics at the time (e.g., Egon Pearson and Gertrude Cox), he apparently never earned a Ph. D. However, he was awarded honorary D. Sc. degrees from both North Carolina State University (1956) and Iowa State University (1958). In 1976, the Committee of Presidents of Statistical Societies established the George W. Snedecor Award. This honors individuals who were instrumental in the development of statistical theory in biometry.

References

- [1] Kirk, R.E. (2003). Experimental design, in J. Schinka & W.F. Velicer, eds, *Handbook of Psychology, Research Methods in Psychology*, Vol. 2. (I.B. Weiner, Editor-in-Chief). New York, Wiley, Available at http://media.wiley.com/product_data/excerpt/31/04713851/0471385131.pdf.
- [2] Snedecor, G.W. (1934). *Analysis of Variance and Covariance*, Iowa State University Press, Ames.
- [3] Snedecor, G.W. (1937). *Statistical Methods*, Iowa State University Press, Ames.
- [4] Wallace, H.A. & Snedecor, G.W. (1925). *Correlation and Machine Calculation*, Iowa State College Press, Ames.

DAVID C. HOWELL

Social Interaction Models

JOHN K. HEWITT

Volume 4, pp. 1864–1865

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Social Interaction Models

Social interaction is modeled in behavior genetics as the influence of one individual's **phenotype** on another, usually within the same family [1, 2]. The principle feature of social interaction in such models is that the phenotype of a given individual (P_1) has an additional source of influence besides the more usually considered additive genetic (A_1), shared family environmental (C_1), and nonshared environmental (E_1) influences (*see ACE Model*). This additional source of influence is the phenotype of another individual (P_2), who is often a sibling and, in the traditional twin study, a monozygotic or dizygotic twin.

Thus, the linear path model changes from: $P_1 = aA_1 + cC_1 + eE_1$ to $P_1 = sP_2 + aA_1 + cC_1 + eE_1$. Of course, in a sibling pair, there is usually no reason to suppose asymmetry of influence (although allowance can be made for such asymmetry in the case of, for example, parents and offspring, or siblings of different ages, or of different sex.) In the symmetrical case, $P_2 = sP_1 + aA_2 + cC_2 + eE_2$. If s is positive, then we would be modeling *cooperative social interactions or imitation*; if s is negative, then we would be modeling *competitive social interactions or contrast*.

We have

$$\begin{aligned} P_1 &= sP_2 + aA_1 + cC_1 + eE_1 \\ P_2 &= sP_1 + aA_2 + cC_2 + eE_2 \end{aligned} \quad (1)$$

or

$$\begin{aligned} P_1 - sP_2 &= aA_1 + cC_1 + eE_1 \\ P_2 - sP_1 &= aA_2 + cC_2 + eE_2 \end{aligned} \quad (2)$$

or

$$(\mathbf{I} - \mathbf{B})\mathbf{P} = \mathbf{aA} + \mathbf{cC} + \mathbf{eE}, \quad (3)$$

where $\mathbf{I}$ is a 2 by 2 identity matrix, $\mathbf{B}$ is a 2 by 2 matrix with zeros on the leading diagonal and s in each of the off diagonal positions, and $\mathbf{P}$, $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{E}$ are each 2 by 1 vectors of the corresponding phenotypes, and genetic and environmental influences.

With a little rearrangement, we see that

$$\mathbf{P} = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{aA} + \mathbf{cC} + \mathbf{eE}), \quad (4)$$

and then the expected covariance matrix (*see Covariance/variance/correlation*) of the phenotypes is the usual expectation for the given type of pair of relatives, premultiplied by $(\mathbf{I} - \mathbf{B})^{-1}$ and postmultiplied by its transpose.

As a consequence, both the expected phenotypic variances of individuals, and the expected covariances between pairs of individuals, are changed by social interaction *and the extent of those changes depends on the a priori correlation of the interacting pair*. Thus, in the presence of cooperative social interaction, the phenotypic variance will be increased, but more so for monozygotic twin pairs than dizygotic twin pairs or full siblings and, in turn, more so for these individuals than for pairs of adopted (not biologically related) siblings. The family covariance will also be increased in the same general pattern, but the proportional increase will be greater for the less closely related pairs. Thus, the overall effect of cooperative social interaction will be similar to that of a shared family environment (which in a sense it is), but will differ in its telltale differential impact on the phenotypic variances of individuals from different types of interacting pairs. For categorical traits, such as disease diagnoses, different **prevalences** in different types of interacting pairs may be detected [2].

In the presence of competitive social interactions, the consequences depend on the details of the parameters of the model but the signature characteristic is that, in addition to phenotypic variances differing for pairs of different relationship, typically in the opposite pattern than for cooperative interactions, the pattern of pair resemblance suggests nonadditive genetic influences. The model predicts negative family correlations under strong competition. These have sometimes been reported for such traits as toddler temperament [4] or childhood hyperactivity, but they may result from the rater contrasting one child with another, especially when a single reporter, such as a parent, is rating both members of a pair of siblings [5].

Although social interactions are of considerable theoretical interest, convincing evidence for the effects of such interactions is hard to find in the behavior genetic literature. The prediction that phenotypic variances are dependent on what kind of pair is being considered is implicitly tested whenever a standard genetic and environmental model is fitted to empirical data, and it has rarely been found

2 Social Interaction Models

wanting. 'For IQ, educational attainment, psychometric assessments of personality, social attitudes, body mass index, heart rate reactivity, and so on, the behavior genetic literature is replete with evidence for the *absence* of the effects of social interactions.' ([3], p. 209.) Thus, the take home message from behavior genetics may be that social interactions, at least those occurring within the twin or sibling context, appear to have surprisingly little effect on individual differences in the traits routinely studied by psychologists.

References

- [1] Carey, G. (1986). Sibling imitation and contrast effects, *Behavior Genetics* **16**, 319–341.
- [2] Carey, G. (1992). Twin imitation for antisocial behavior: implications for genetic and family environment research, *Journal of Abnormal Psychology* **101**, 18–25.
- [3] Neale, M.C. & Cardon, L. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, Boston.
- [4] Saudino, K.J., Cherny, S.S. & Plomin, R. (2000). Parent ratings of temperament in twins: explaining the 'too low' DZ correlations, *Twin Research* **3**, 224–233.
- [5] Simonoff, E., Pickles, A., Hervas, A., Silberg, J., Rutter, M. & Eaves, L.J. (1998). Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction, *Psychological Medicine* **28**, 825–837.

JOHN K. HEWITT

Social Networks

STANLEY WASSERMAN, PHILIPPA PATTISON AND DOUGLAS STEINLEY

Volume 4, pp. 1866–1871

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Social Networks

Network analysis is the interdisciplinary study of social relations and has roots in anthropology, sociology, psychology, and applied mathematics. It conceives of social structure in relational terms, and its most fundamental construct is that of a *social network*, comprising at the most basic level a set of *social actors* and a set of *relational ties* connecting pairs of these actors. A primary assumption is that social actors are interdependent, and that the relational ties among them have important consequences for each social actor as well as for the larger social groupings that they comprise.

The *nodes* or *members* of the network can be groups or organizations as well as people. Network analysis involves a combination of theorizing, model building, and empirical research, including (possibly) sophisticated data analysis. The goal is to study network structure, often analyzed using such concepts as density, centrality, prestige, mutuality, and role. Social network data sets are occasionally multidimensional and/or longitudinal, and they often include information about *actor attributes*, such as actor age, gender, ethnicity, attitudes, and beliefs.

A basic premise of the social network paradigm is that knowledge about the structure of social relationships enriches explanations based on knowledge about the attributes of the actors alone. Whenever the social context of individual actors under study is relevant, relational information can be gathered and studied. Network analysis goes beyond measurements taken on individuals to analyze data on patterns of relational ties and to examine how the existence and functioning of such ties are constrained by the social networks in which individual actors are embedded. For example, one might measure the relations ‘communicate with’, ‘live near’, ‘feel hostility toward’, and ‘go to for social support’ on a group of workers. Some network analyses are longitudinal, viewing changing social structure as an outcome of underlying processes. Others link individuals to events (affiliation networks), such as a set of individuals participating in a set of community activities.

Network structure can be studied at many different levels: the dyad, triad, subgroup, or even the entire network. Furthermore, network theories can be postulated at a variety of different levels. Although this multilevel aspect of network analysis allows different

structural questions to be posed and studied simultaneously, it usually requires the use of methods that go beyond the standard approach of treating each individual as an independent unit of analysis. This is especially true for studying a *complete* or whole network: a census of a well-defined population of social actors in which all ties, of various types, among all the actors are measured. Such analyses might study structural balance in small groups, transitive flows of information through indirect ties, structural equivalence in organizations, or patterns of relations in a set of organizations.

For example, network analysis allows a researcher to model the interdependencies of organization members. The paradigm provides concepts, theories, and methods to investigate how informal organizational structures intersect with formal bureaucratic structures in the unfolding flow of work-related actions of organizational members and in their evolving sets of knowledge and beliefs. Hence, it has informed many of the topics of organizational behavior, such as leadership, attitudes, work roles, turnover, and computer-supported cooperative work.

Historical Background

Network analysis has developed out of several research traditions, including (a) the birth of sociometry in the 1930s spawned by the work of the psychiatrist Jacob L. Moreno; (b) ethnographic efforts in the 1950s and 1960s to understand migrations from tribal villages to polyglot cities, especially the research of A. R. Radcliffe-Brown; (c) survey research since the 1950s to describe the nature of personal communities, social support, and social mobilization; and (d) archival analysis to understand the structure of interorganizational and international ties. Also noteworthy is the work of Claude Lévi-Strauss, who was the first to introduce formal notions of kinship, thereby leading to a mathematical algebraic theory of relations, and the work of Anatol Rapoport, perhaps the first to propose an elaborate statistical model of relational ties and flow through various nodes.

Highlights of the field include the adoption of sophisticated mathematical models, especially discrete mathematics and graph theory, in the 1940s and 1950s. Concepts such as transitivity, structural equivalence, the strength of weak ties, and centrality arose from network research by James A. Davis,

Samuel Leinhardt, Paul Holland, Harrison White, Mark Granovetter, and Linton Freeman in the 1960s and 1970s. Despite the separateness of these many research beginnings, the field grew and was drawn together in the 1970s by formulations in graph theory and advances in computing. Network analysis, as a distinct research endeavor, was born in the early 1970s. Noteworthy in its birth is the pioneering text by Harary, Norman, and Cartwright [4]; the appearance in the late 1970s of network analysis software, much of it arising at the University of California, Irvine; and annual conferences of network analysts, now sponsored by the International Network for Social Network Analysis. These well-known ‘Sunbelt’ Social Network Conferences now draw as many as 400 international participants. A number of fields, such as organizational science, have experienced rapid growth through the adoption of a network perspective.

Over the years, the social network analytic perspective has been used to gain increased understanding of many diverse phenomena in the social and behavioral sciences, including (taken from [8])

- Occupational mobility
- Urbanization
- World political and economic systems
- Community elite decision making
- Social support
- Community psychology
- Group problem solving
- Diffusion and adoption of information
- Corporate interlocking
- Belief systems
- Social cognition
- Markets
- Sociology of science
- Exchange and power
- Consensus and social influence
- Coalition formation

In addition, it offers the potential to understand many contemporary issues, including (see [1])

- The Internet
- Knowledge and distributed intelligence
- Computer-mediated communication
- Terrorism
- Metabolic systems
- Health, illness, and epidemiology, especially of HIV

Before a discussion of the details of various network research methods, we mention in passing a number of important measurement approaches.

Measurement

Complete Networks

In complete network studies, a census of network ties is taken for all members of a prespecified population of network members. A variety of methods may be used to observe the network ties (e.g., survey, archival, participant observation), and observations may be made on a number of different types of network tie. Studies of complete networks are often appropriate when it is desirable to understand the action of network members in terms of their location in a broader social system (e.g., their centrality in the network, or more generally in terms of their patterns of connections to other network members). Likewise, it may be necessary to observe a complete network when properties of the network as a whole are of interest (e.g., its degree of centralization, fragmentation, or connectedness).

Ego-centered Networks

The size and scope of complete networks generally preclude the study of all the ties and possibly all the nodes in a large, possibly unbounded population. To study such phenomena, researchers often use survey research to study a sample of personal networks (often called *ego-centered* or *local* networks). These smaller networks consist of the set of specified ties that links *focal persons* (or *egos*) at the centers of these networks to a set of close ‘associates’ or *alters*. Such studies focus on an ego’s ties and on ties among ego’s alters. Ego-centered networks can include relations such as kinship, weak ties, frequent contact, and provision of emotional or instrumental aid. These relations can be characterized by their variety, content, strength, and structure. Thus, analysts might study network member *composition* (such as the percentage of women providing social or emotional support, for example, or basic actor attributes more generally); network characteristics (e.g., percentage of dyads that are mutual); measures of *relational association* (do strong ties with immediate kin also imply supportive relationships?); and network

structure (how densely knit are various relations? do actors cluster in any meaningful way?).

Snowball Sampling and Link Tracing Studies

Another possibility, to study large networks, is simply to sample nodes or ties. Sampling theory for networks contains a small number of important results (e.g., estimation of subgraphs or subcomponents; many originated with Ove Frank) as well as a number of unique techniques or strategies such as snowball sampling, in which a number of nodes are sampled, then those linked to this original sample are sampled, and so forth, in a multistage process. In a link-tracing sampling design, emphasis is on the links rather than the actors – a set of social links is followed from one respondent to another. For hard-to-access or hidden populations, such designs are considered the most practical way to obtain a sample of nodes. Related are recent techniques that obtain samples, and based on knowledge of certain characteristics in the population and the structure of the sampled network, make inferences about the population as a whole (see [5, 6]).

Cognitive Social Structures

Social network studies of *social cognition* investigate how individual network actors perceive the ties of others and the social structures in which they are contained. Such studies often involve the measurement of multiple perspectives on a network, for instance, by observing each network member's view of who is tied to whom in the network. David Krackhardt referred to the resulting data arrays as *cognitive social structures*. Research has focused on clarifying the various ways in which social cognition may be related to network locations: (a) People's positions in social structures may determine the specific information to which they are exposed, and hence, their perception; (b) structural position may be related to characteristic patterns of social interactions; (c) structural position may frame social cognitions by affecting people's perceptions of their social locales.

Methods

Social network analysts have developed methods and tools for the study of relational data. The techniques

include graph theoretic methods developed by mathematicians (many of which involve counting various types of subgraphs); algebraic models popularized by mathematical sociologists and psychologists; and statistical models, which include the social relations model from social psychology and the recent family of random graphs first introduced into the network literature by Ove Frank and David Strauss. Software packages to fit these models are widely available.

Exciting recent developments in network methods have occurred in the statistical arena and reflect the increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. Therefore, a growing importance has been accorded the problem of constructing theoretically and empirically plausible parametric models for structural network phenomena and their changes over time. Substantial advances in statistical computing are now allowing researchers to more easily fit these more complex models to data.

Some Notation

In the simplest case, network studies involve a single type of directed or nondirected tie measured for all pairs of a node set $N = \{1, 2, \dots, n\}$ of individual actors. The observed tie linking node i to node j ($i, j \in N$) can be denoted by x_{ij} and is often defined to take the value 1 if the tie is observed to be present and 0 otherwise. The network may be either *directed* (in which case x_{ij} and x_{ji} are distinguished and may take different values) or *nondirected* (in which case x_{ij} and x_{ji} are not distinguished and are necessarily equal in value). Other cases of interest include the following:

1. *Valued* networks, where x_{ij} takes values in the set $\{0, 1, \dots, C - 1\}$.
2. *Time-dependent* networks, where x_{ijt} represents the tie from node i to node j at time t .
3. *Multiple relational* or *multivariate* networks, where x_{ijk} represents the tie of type k from node i to node j (with $k \in R = \{1, 2, \dots, r\}$, a fixed set of *types* of tie).

In most of the statistical literature on network methods, the set N is regarded as fixed and the network ties are assumed to be random. In this case, the tie linking node i to node j may be denoted by the random variable X_{ij} and the $n \times n$ array $X = [X_{ij}]$

of random variables can be regarded as the adjacency matrix of a *random (directed) graph* on N .

Graph Theoretic Techniques

Graph theory has played a critical role in the development of network analysis. Graph theoretical techniques underlie approaches to understanding cohesiveness, connectedness, and fragmentation in networks. Fundamental measures of a network include its *density* (the proportion of possible ties in the network that are actually observed) and the degree sequence of its nodes. In a nondirected network, the *degree* d_i of node i is the number of distinct nodes to which node i is connected. Methods for characterizing and identifying cohesive subsets in a network have depended on the notion of a *clique* (a subgraph of network nodes, every pair of which is connected) as well as on a variety of generalizations (including k -clique, k -plex, k -core, LS -set, and k -connected subgraph).

Our understanding of connectedness, connectivity, and centralization is also informed by the distribution of path lengths in a network. A *path* of length k from one node i to another node j is defined by a sequence $i = i_1, i_2, \dots, i_{k+1} = j$ of distinct nodes such that i_h and i_{h+1} are connected by a network tie. If there is no path from i to j of length $n - 1$ or less, then j is not reachable from i and the distance from i to j is said to be infinite; otherwise, the distance from i to j is the length of the shortest path from i to j . A directed network is *strongly connected* if each node is reachable from each other node; it is *weakly connected* if, for every pair of nodes, at least one of the pair is reachable from the other. For nondirected networks, a network is connected if each node is reachable from each other node, and the *connectivity*, κ , is the least number of nodes whose removal results in a disconnected (or trivial) subgraph.

Graphs that contain many cohesive subsets as well as short paths, on average, are often termed *small world* networks, following early work by Stanley Milgram, and more recent work by Duncan Watts. Characterizations of the centrality of each actor in the network are typically based on the actor's degree (*degree* centrality), on the lengths of paths from the actor to all other actors (*closeness* centrality), or on the extent to which the shortest paths between other actors pass through the given

actor (*betweenness* centrality). Measures of network *centralization* signify the extent of heterogeneity among actors in these different forms of centrality.

Algebraic Techniques

Closely related to graph theoretic approaches is a collection of algebraic techniques that has been developed to understand social roles and structural regularities in networks. Characterizations of role have developed in terms of mappings on networks, and descriptions of structural regularities have been facilitated by the construction of algebras among labeled network walks. An important proposition about what it means for two actors to have the same social role is embedded in the notion of *structural equivalence*: Two actors are said to be *structurally equivalent* if they are related to and are related to by every other network actor in exactly the same way (thus, nodes i and j are structurally equivalent if, for all $k \in N$, $x_{ik} = x_{jk}$ and $x_{ki} = x_{kj}$). Generalizations to automorphic and regular equivalence are based on more general mappings on N and capture the notion that similarly positioned network nodes are related to *similar* others in the same way.

Statistical Techniques

A simple statistical model for a (directed) graph assumes a Bernoulli distribution (see **Catalogue of Probability Density Functions**), in which each edge, or tie, is statistically independent of all others and governed by a theoretical probability P_{ij} . In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the probabilities to depend on structural parameters. These distributions often have been used as models for at least 40 years, but are of questionable utility because of the independence assumption.

Dyadic Structure in Networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. The p_1 model recognized the theoretical and empirical importance of dyadic structure in social networks, that is, of the interdependence of the variables X_{ij} and X_{ji} .

This class of Bernoulli dyad distributions and their generalization to valued, multivariate, and time-dependent forms gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time. The model assumes that each dyad (X_{ij}, X_{ji}) is independent of every other, resulting in a **log-linear model** that is easily fit. Generalizations of this model are numerous, and include *stochastic block models*, representing hypotheses about the interdependence of social positions and the patterning of network ties; mixed models, such as p_2 ; and *latent space models* for networks.

Null Models for Networks

The assumption of dyadic independence is questionable. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. A number of *conditional uniform* random graph distributions were introduced as null models for exploring the structural features of social networks. These distributions, denoted by $U|Q$, are defined over subsets Q of the state space Ω_n of directed graphs and assign equal probability to each member of Q . The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric, and null dyads). When Q is equal to Ω_n , the distribution is referred to as the *uniform* (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the members of Q and simulation of $U|Q$ are often straightforward, although certain cases, such as the distribution that is conditional on the indegree and outdegree of each node i in the network, require more complicated approaches.

A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features. This general approach has also been developed for the analysis of multiple networks. The best known example is probably Frank Baker and Larry Hubert's quadratic assignment procedure (QAP) for networks. In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multigraph distribution that is conditional

on the unlabeled graph structure of each constituent graph.

Extradynamic Local Structure in Networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss [3] with the introduction of the class of *Markov random graphs*, denoted as p^* by later researchers. This class of models permitted the parameterization of extradynamic local structural forms, allowing a more explicit link between some important theoretical propositions and statistical network models. These models are based on the fact that the Hammersley–Clifford theorem provides a general probability distribution for X from a specification of which pairs (X_{ij}, X_{kl}) of tie random variables are conditionally dependent, given the values of all other random variables.

These random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity, local clustering, degree variability, and centralization. Valued, multiple, and temporal generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability, generalized transitivity and exchange, and the strength of weak ties. Pseudomaximum likelihood estimation is easy; **maximum likelihood estimation** is difficult, but not impossible.

Dynamic Models

A significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of long-standing relational ties. Early attempts to model the evolution of networks in either discrete or continuous time assumed dyad independence and Markov processes in time. A step towards continuous time **Markov chain** models for network evolution that relaxes the assumption of dyad independence has been taken by Tom Snijders and colleagues. This approach also illustrates the potentially valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such

methods provide a promising focus for future development. Computational models based on simulations are becoming increasingly popular in network analysis; however, the development of associated model evaluation approaches poses a significant challenge.

Current research, including future challenges, such as statistical estimation of complex model parameters, model evaluation, and dynamic statistical models for longitudinal data, can be found in [2]. Applications of the techniques and definitions mentioned here can be found in [7] and [8].

References

- [1] Breiger, R., Carley, K. & Pattison, P., eds (2003). *Dynamic Social Network Modeling and Analysis*, The National Academies Press, Washington.
- [2] Carrington, P.J., Scott, J. & Wasserman, S., eds (2004). *Models and Methods in Social Network Analysis*, Cambridge University Press, New York.
- [3] Frank, O. & Strauss, D. (1986). Markov graphs, *Journal of the American Statistical Association* **81**, 832–842.
- [4] Harary, F., Norman, D. & Cartwright, D. (1965). *Structural Models for Directed Graphs*, Free Press, New York.
- [5] Killworth, P.D., McCarty, C., Bernard, H.R., Shelley, G.A. & Johnsen, E.C. (1998). Estimation of seroprevalence, rape, and homelessness in the U.S. using a social network approach, *Evaluation Review* **22**, 289–308.
- [6] McCarty, C., Killworth, P.D., Bernard, H.R., Johnsen, E.C. & Shelley, G.A. (2001). Comparing two methods for estimating network size, *Human Organization* **60**, 28–39.
- [7] Scott, J. (1992). *Social Network Analysis*, Sage Publications, London.
- [8] Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York.

Further Reading

- Boyd, J.P. (1990). *Social Semigroups: A Unified Theory of Scaling and Bockmodeling as Applied to Social Networks*, George Mason University Press, Fairfax.
- Friedkin, N. (1998). *A Structural Theory of Social Influence*, Cambridge University Press, New York.
- Monge, P. & Contractor, N. (2003). *Theories of Communication Networks*, Oxford University Press, New York.
- Pattison, P.E. (1993). *Algebraic Models for Social Networks*, Cambridge University Press, New York.
- Wasserman, S. & Galaskiewicz, J., eds (1994). *Advances in Social Network Analysis*, Sage Publications, Thousand Oaks.
- Watts, D. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton.
- Wellman, B. & Berkowitz, S.D., eds (1997). *Social Structures: A Network Approach*, (updated Edition), JAI, Greenwich.

STANLEY WASSERMAN, PHILIPPA PATTISON
AND DOUGLAS STEINLEY

Social Psychology

CHARLES M. JUDD AND DOMINIQUE MULLER

Volume 4, pp. 1871–1874

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Social Psychology

Social psychologists are interested in the ways behavior is affected by the social situations in which people find themselves. Recurring topics of research interest in this subdiscipline include (a) stereotyping, prejudice, and intergroup behavior; (b) interpersonal attraction and close relationships; (c) behavior in small groups and group decision making; (d) social influence, conformity, and social norms; (e) attitudes and affective responses to social stimuli; and (f) dyadic interaction and communication. These topics do not exhaust the list of social psychological interests, but they are representative of the broad array of social behaviors that have attracted research attention.

As an empirical discipline, social psychology has taken pride in its use of systematic observation methods to reach conclusions about the causes, processes, and consequences of social behavior. In this endeavor, it has historically adopted an experimental perspective by simulating social environments in the laboratory and other settings, and gathering data on social behavior in those settings. The impetus for this approach came from the seminal influence of Kurt Lewin (and his students) who showed how important social phenomena can be captured in simulated experimental environments. For example, to study the effects of different leadership styles in small groups, Lewin and colleagues [7] examined the consequences of democratic and autocratic leadership styles in simulated small group interactions.

In this empirical approach, participants are typically randomly assigned to various experimental conditions (*see* **Randomization**) (e.g., a democratic or autocratic leader) and differences in their responses (e.g., observed behaviors, self-reports on questionnaires) are calculated. Accordingly, statistical procedures in experimental social psychology have been dominated by examining and testing mean differences using *t* Tests and analysis of variance procedures. A typical report of experimental results in social psychological journals gives condition means (and standard deviations) and the inferential statistic that shows that the observed mean differences are significantly different from zero. Accordingly, traditional null hypothesis testing has been the dominant approach, with attention given only recently to reporting **effect size estimates** or **confidence intervals** for observed means (and for their differences).

Experimental designs in social psychology have become increasingly complex over the years, incorporating multiple (crossed) experimental factors (*see* **Factorial Designs**), as researchers have hypothesized that the effects of some manipulations depend on other circumstances or events. Accordingly, **analysis of variance** models with multiple factors are routinely used, with considerable attention focused on statistical interactions (*see* **Interaction Effects**): Does the effect of one experimental factor depend upon the level of another? In fact, it is not unusual for social psychological experiments to cross three or more experimental factors, with resulting higher order interactions that researchers attempt to meaningfully interpret. One of our colleagues jokingly suggests that a criterion for publication in the leading social psychological journals is a significant (at least two-way) interaction.

For some independent variables of interest to social psychologists, manipulations are done ‘within participants’, exposing each participant in turn to multiple levels of a factor, and measuring responses under each level (*see* **Repeated Measures Analysis of Variance**). This has seemed particularly appropriate where the effects of those manipulations are short-lived, and where one can justify assumptions about the lack of **carry-over effects**. Standard analysis of variance courses taught to social psychologists have typically included procedures to analyze data from within-participant designs, which are routinely used in experimental and cognitive psychological research as well. Thus, beyond multifactorial analyses of variance, standard analytic tools have included **repeated measures analysis of variance** models and split-plot designs.

While social psychologists have traditionally been very concerned about the integrity of their experimental manipulations and the need to randomly assign participants to experimental conditions (e.g., [3]), they have historically been much less concerned about sampling issues (*see* **Experimental Design**). In fact, the discipline has occasionally been subjected to substantial criticism for its tendency to rely on undergraduate research participants, recruited from psychology courses where ‘research participation’ is a course requirement (*see* [9]). The question raised is about the degree to which results from social psychological experiments can be generalized to a public broader than college undergraduates. From a

statistical point of view, the use of convenience samples such as undergraduates who happen to enroll in psychology classes means that standard inferential statistical procedures are difficult to interpret since there is no known population from which one has sampled. Rather than concluding that a statistically significant difference suggests a true difference in some known population, one is left with hypothetical populations and unclear inferences such as 'If I were to repeat this study over and over again with samples randomly chosen from the unknown hypothetical population from which I sampled, I would expect to continue to find a difference.'

A classic interaction perspective in social psychology, again owing its heritage to Kurt Lewin, is that situational influences on social behavior depend on the motives, aspirations, and attributes of the participants: Behavior results from the interaction of the personality of the actor and the characteristics of the situation. Accordingly, many social psychological studies involve at least one measured, rather than manipulated, independent variable, which is typically a variable that characterizes a difference among participants. Because many social psychologists received training exclusively in classic analysis of variance procedures, the analysis of designs, where one or more measured independent variables are crossed with other factors, has routinely been accomplished by dividing the measured variable into discrete categories, by, for example, a median split (*see Categorizing Data*), and including this variable in a factorial analysis of variance. This practice leads to a loss of statistical **power** and, occasionally, to biased estimates [8]. Recently, however, because of the pioneering effort of **Jacob Cohen** [4] and others, social psychologists began to appreciate that analysis of variance is a particular instantiation of the **generalized linear model**. Accordingly, both categorical and more continuously measured independent variables (and their interactions) can be readily included in models that are estimated by standard ordinary least squares regression programs. As social psychologists have become more familiar with the wide array of models estimable under the general linear model, they have increasingly strayed from the laboratory, measuring and manipulating independent variables and using longitudinal designs to assess naturally occurring changes in behavior overtime and in different situations.

With these developments, social psychologists have been forced to confront the limitations inherent in their standard analytic toolbox, specifically assumptions concerning the independence of errors or residuals in both standard analysis of variance and ordinary least squares regression procedures (*see Regression Models*). Admittedly, they use specific techniques for dealing with nonindependence in repeated measures designs, in which participants are repeatedly measured under different levels of one or more independent variables. But the assumptions underlying **repeated measures analysis of variance** models, specifically the assumption of equal **covariances** among all repeated observations, seems too restrictive for many research designs where data are collected overtime from participants in naturally occurring environments.

In many situations of interest to social psychologists, data are likely to exhibit nested structures, wherein observations come from dyads, small groups, individuals who are located in families, and so forth. These nested structures mean that observations within groupings are likely to be more similar to each other than observations that are from different groupings. For instance, in collecting data on close relationships, it is typical to measure both members of a number of couples, asking, for instance, about levels of marital satisfaction. Unsurprisingly, observations are likely to show resemblances due to couples. Or, in small group research, we may ask for the opinions of group members who come from multiple groups. Again, it is likely that group members agree with each other to some extent. Or, on a larger scale, if we are interested in behavior in larger organizations, we may sample individuals who are each located in different organizations.

Additional complications arise when individuals or dyads or groups are measured overtime, with, perhaps, both the interval of observations and the frequency of observations varying. For instance, a social psychologist might ask each member of a couple to indicate emotional responses each time an interaction between them lasted more than five minutes. What might be of interest here is the consistency of the emotional responses of one member of the couple compared to the other's and the degree to which that consistency varied with other known factors about the couple. Clearly, standard analysis of variance and ordinary least squares regression procedures cannot be used in such situations. Accordingly,

social psychologists are increasingly making use of more general models for nested and dependent data, widely known as multilevel modeling procedures (*see* **Linear Multilevel Models**) or random regression models [2, 5]. Although the use of such analytic approaches to data is still not widespread in social psychology, their utility for the analysis of complex data structures involving dependent observations suggests that they will become standard analytic tools in social psychology in the near future.

In addition to these advances in analytic practices, social psychologists have also become more sophisticated in their use of analytic models for dichotomous-dependent variables. The use of **logistic regression** procedures, for instance, is now fairly widespread in the discipline. Such procedures have significantly extended the range of questions that can be asked about dichotomous variables, permitting, for instance, tests of higher order interactions.

Additionally, while the focus has historically been on the assessment of the overall impact of one or more independent variables on the dependent variable, social psychologists have become increasingly interested in process questions such as what is the mediating process by which the impact is produced? Thus, social psychologists have increasingly made use of analytic procedures designed to assess **mediation**. Given this and the long-standing interest in the discipline in statistical interactions, it is no accident that the classic article on procedures for assessing mediation and **moderation** was written by two social psychologists and published in a leading social psychological outlet [1].

All of the procedures discussed up to this point involve the estimation of relationships between variables thought to assess different theoretical constructs such as the effect of a manipulated independent variable on an outcome variable. Additionally, statistical procedures are used in social psychology to support measurement claims, such as that a particular variable successfully measures a construct of theoretical interest [6]. For these measurement claims, both exploratory and, more recently, confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) procedures are used. Exploratory factor analysis (*see* **Factor Analysis: Exploratory**) is routinely used when one develops a set of self-report questions, designed, for instance, to measure a particular attitude, and one wants to verify that the questions exhibit a pattern of correlations that suggests they

have a single underlying factor in common. In this regard, social psychologists are likely to conduct a principal factoring (*see* **Principal Component Analysis**) or components analysis to demonstrate that the first factor or component explains a substantial amount of the variance in all of the items. Occasionally, researchers put together items that they suspect may measure a number of different latent factors (*see* **Latent Class Analysis; Latent Variable**) or constructs that they are interested in ‘uncovering’. This sort of approach has a long history in intelligence testing, where researchers attempted to uncover the ‘true’ dimensions of intelligence. However, as that literature suggests, this use of factor analysis is filled with pitfalls. Not surprisingly, if items that tap a given factor are not included, then that factor cannot be ‘uncovered’. Additionally, a factor analytic solution is indeterminate, with different rotations yielding different definitions of underlying dimensions.

Recently, confirmatory factor analytic models have become the approach of choice to demonstrate that items measure a hypothesized underlying construct. In this approach, one hypothesizes a latent factor structure that involves one or more factors, and then examines whether the item covariances are consistent with that structure. Additionally, **structural equation modeling** procedures are sometimes used to model both the relationships between the latent constructs and the measured variables, and also the linear structural relations among those constructs. This has the advantage of estimating the relationships among constructs potentially unbiased by measurement error, because those errors of measurement are modeled in the confirmatory factor analysis part of the estimation. The use of structural equation modeling procedures will remain limited in social psychology, however, for they are not efficient at examining interactive and nonlinear predictions.

In summary, social psychologists are abundant users of statistical and data analytic tools. They pride themselves on the fact that theory evaluation in their discipline ultimately rests on gathering data through systematic observation that can be used to either bolster theoretical conjectures or argue against them. As a function of being usually trained in psychology departments, their standard analytic tools have been those taught in experimental design courses. However, social psychologists often collect data that demand other data analytic approaches. Gradually,

social psychologists are becoming sophisticated users of these more flexible approaches. In fact, the statistical demands of some of the data routinely collected by social psychologists means that many new developments in statistical tools for behavioral and social scientists are being developed by social psychologists. And it is no accident that in many psychology departments, the quantitative and analytic courses are now being taught by social psychologists, considerably expanding the traditional analysis of variance and experimental design emphases of such courses.

References

- [1] Baron, R.M. & Kenny, D.A. (1986). The moderator – mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.
- [2] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Newbury Park.
- [3] Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research, in *Handbook of Research on Teaching*, N.L. Gage, ed., Rand McNally, Chicago, pp. 171–246.
- [4] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.
- [5] Goldstein, H. (1995). *Multilevel Statistical Models*, Halstead Press, New York.
- [6] Judd, C.M. & McClelland, G.H. (1998). Measurement, in *The Handbook of Social Psychology*, Vol. 1, 4th Edition, D.T. Gilbert, S.T. Fiske & G. Lindzey, eds, McGraw-Hill, Boston, pp. 180–232.
- [7] Lewin, K., Lippitt, R. & White, R. (1939). Patterns of aggressive behavior in experimental created “social climates.”, *Journal of Social Psychology* **10**, 271–299.
- [8] Maxwell, S.E. & Delaney, H.D. (1993). Bivariate mediation splits and spurious statistical significance, *Psychological Bulletin* **113**, 181–190.
- [9] Sears, D.O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology’s view of human nature, *Journal of Personality and Social Psychology* **51**, 515–530.

CHARLES M. JUDD AND DOMINIQUE MULLER

Social Validity

ALAN E. KAZDIN

Volume 4, pp. 1875–1876

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Social Validity

Social validity is a concept that is used in intervention research in which the focus is on treatment, prevention, rehabilitation, education, and, indeed, any area in which the goal is to produce change in human behavior and adaptive functioning (*see Clinical Trials and Intervention Studies*). The concept of 'social' in the term emphasizes the views and perspectives of individuals who are stakeholders, consumers, or recipients of the intervention. Social validity raises three questions in the context of designing, implementing, and evaluating interventions: (a) Are the *goals* of the intervention relevant to everyday life? (b) Are the *intervention procedures* acceptable to consumers and to the community at large? (c) Are the *outcomes* of the intervention important; that is, do the changes make a difference in the everyday lives of individuals to whom the intervention was directed or those who are in contact with them? The focus is on whether consumers of the interventions find the goals, intervention, and outcomes, reasonable, acceptable, and useful.

Background

Social validation grew out of work in *applied behavior analysis*, an area of behavior modification within psychology [1, 2]. Applied behavior analysis draws on operant conditioning, a type of learning elaborated by B. F. Skinner [7] and that focuses on antecedents, behaviors, and consequences. In the late 1950s and early 1960s, the principles, methods, and techniques of operant conditioning, developed in animal and human laboratory research, were extended to problems of treatment, education, and rehabilitation in applied settings. The applications have included a rather broad array of populations (infants, children adolescents, adults), settings (e.g., medical hospitals, psychiatric hospitals, schools [preschool through college], nursing homes), and contexts (e.g., professional sports, business and industry, the armed forces) [3].

Social validation was initially developed by Montrose Wolf [8], a pioneer in applied behavior analysis, to consider the extent to which the intervention was addressing key concerns of individuals in everyday life. In applying interventions to help people, he reasoned, it was invariably important to ensure that the

interventions, their goals, and the effects that were attained were in keeping with the interests of individuals affected by them.

An Example

Consider as an example the application of operant conditioning principles to reduce self-injurious behavior in seriously disturbed children. Many children with pervasive developmental disorder or severe mental retardation hit or bite themselves with such frequency and intensity that physical damage can result. Using techniques to alter antecedents (e.g., prompts, cues, and events presented before the behavior occurs) and consequences (e.g., carefully arranged incentives and activities), these behaviors have been reduced and eliminated [6].

As a hypothetical example, assume for a moment that we have several institutionalized children who engage in severe self-injury such as head banging (banging head against a wall or pounding one's head) or biting (biting one's hands or arms sufficiently to draw blood). We wish to intervene to reduce self-injury. The initial social validity question asks if the goals are relevant or important to everyday life. Clearly they are. Children with self-injury cannot function very well even in special settings, often must be restrained, and are kept from a range of interactions because of the risk to themselves and to others. The next question is whether the interventions are acceptable to consumers (e.g., children and their families, professionals who administer the procedures). Aversive procedures (e.g., punishment, physical restraint, or isolation) might be used, but they generally are unacceptable to most professionals and consumers. Fortunately, effective procedures are available that rely on variations of reinforcement and are usually quite acceptable to consumers [3, 6].

Finally, let us say we apply the intervention and the children show a reduction of self-injury. Let us go further and say that before the intervention, the children of this example hit themselves a mean of 100 times per hour, as observed directly in a special classroom of a hospital facility. Let us say further that treatment reduced this to a mean of 50 times. Does the change make a difference in everyday life? To be sure, a 50% reduction is large, but still not likely to improve adjustment and functioning of the individuals in everyday life. Much larger reductions,

indeed, elimination of the behaviors, are needed to have clear social impact.

Extensions

The primary application of social validity has evolved into a related concept, 'clinical significance' that focuses on the outcomes of treatment in the context of psychotherapy for children, adolescents, and adults. The key question of clinical significance is the third one of social validity, namely, does treatment make a difference to the lives of those treated? **Clinical trials** of therapy (e.g., for depression, anxiety, marital discord) often compare various treatment conditions or treatment and control conditions. At the end of the study, statistical tests are usually used to determine whether the group differences and whether the changes from pre- to post-treatment are statistically significant. Statistical significance is not intended to reflect important effects in relation to the functioning of individuals. For example, a group of obese individuals (e.g., >90 kilograms overweight) who receive treatment may lose a mean of nine kilograms, and this change could be statistically significant in comparison to a control group that did not receive treatment. Yet, the amount of weight lost is not very important or relevant from the standpoint of clinical significance. Health (morbidity and mortality) and adaptive functioning (e.g., activities in everyday life) are not likely to be materially improved with such small changes (see **Effect Size Measures**).

Treatment evaluation increasingly supplements statistical significance with indices of clinical significance to evaluate whether the changes are actually important to the patients or clients, and those with whom they interact (e.g., spouses, coworkers). There are several indices currently in use such as evaluating whether the level of symptoms at the end of treatment

falls within a normative range of individuals functioning well in everyday life, whether the condition that served as the basis for treatment (e.g., depression, panic attacks) is no longer present, and whether the changes made by the individual are especially large [4, 5]. Social validity and its related but more focused concept of clinical significance have fostered increased attention to whether treatment outcomes actually help people in everyday life. The concept has not replaced other ways of evaluating treatment, for example, statistical significance, magnitude of change, but has expanded the criteria by which to judge intervention effects.

References

- [1] Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **1**, 91–97.
- [2] Baer, D.M., Wolf, M.M. & Risley, T.R. (1987). Some still-current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **20**, 313–328.
- [3] Kazdin, A.E. (2001). *Behavior Modification in Applied Settings*, 6th Edition, Wadsworth, Belmont.
- [4] Kazdin, A.E. (2003). *Research Design in Clinical Psychology*, 4th Edition, Allyn & Bacon, Needham Heights.
- [5] Kendall, P.C., ed. (1999). Special section: clinical significance. *Journal of Consulting and Clinical Psychology* **67**, 283–339.
- [6] Repp, A.C. & Singh, N.N., eds. (1990). *Perspectives on the Use of Nonaversive and Aversive Interventions for Persons with Developmental Disabilities*, Sycamore Publishing, Sycamore.
- [7] Skinner, B.F. (1938). *The Behavior of Organisms: An Experimental Analysis*, Appleton-Century, New York.
- [8] Wolf, M.M. (1978). Social validity: the case for subjective measurement, or how applied behavior analysis is finding its heart, *Journal of Applied Behavior Analysis* **11**, 203–214.

ALAN E. KAZDIN

Software for Behavioral Genetics

MICHAEL C. NEALE

Volume 4, pp. 1876–1880

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Software for Behavioral Genetics

Historical Background

The term software was coined in the 1950s by the eminent statistician **John Tukey** (1915–2000). It usually refers to the program and algorithms used to control the electronic machinery (hardware) of a computer, and may include the documentation. Typically, software consists of source code, which is then compiled into machine-executable code which the end user applies to a specific problem. This general scenario applies to software for genetically informative studies, and might be considered to have existed before Professor Tukey invented the term in the mid twentieth Century. Algorithms are at the heart of software, and this term dates back to the ninth century Iranian mathematician, Al-Khawarizmi. Although formal analysis of data collected from twins did not begin until the 1920s, it was, nevertheless, algorithmic in form. A heuristic estimate of **heritability**, such as twice the difference between the MZ and the DZ correlations, may be implemented using mental arithmetic, the back of an envelope, or on a supercomputer. In all cases the algorithm constitutes software; it is only the hardware that differs.

Software for Model-fitting

Much current behavior genetic analysis is built upon the statistical framework of **maximum likelihood**, attributed to **Ronald Fisher** [4]. As its name suggests, maximum likelihood requires an algorithm for *optimization*, of which there are many: some general and some specific to particular applications. All such methods use *input data* whose likelihood is computed under a particular statistical model. The values of the parameters of this model are not known, but it is often possible to obtain the set of values that maximize the likelihood. These *maximum likelihood estimates* have two especially desirable statistical properties; they are asymptotically unbiased, and have minimum variance of all asymptotically unbiased estimates. Therefore, in the analysis of both genetic linkage (see **Linkage Analysis**) using genetic markers, and of **twin studies** to estimate variance components,

there was motivation to pursue these more complex methods. This section focuses on twin studies and their extensions.

Before the advent of high-speed computers, maximum likelihood estimation would typically involve: (a) writing out the formula for the likelihood; (b) finding the first and second derivatives of this function with respect to the parameters of the model; and (c) solving the (often nonlinear) simultaneous equations to find those values of the parameters that maximize the likelihood, that is, where the first derivatives are zero and the second derivatives are negative. The first of these steps is often relatively simple, as it typically involves writing out the formula for the probability density function (pdf) (see **Catalogue of Probability Density Functions**) of the parameters of the model. In many cases, however, the second and third steps can prove to be challenging or intractable. Therefore, the past 25 years has seen the advent of software designed to estimate parameters under increasingly general conditions.

Early applications of software for numerical optimization (see **Optimization Methods**) to behavior genetic data primarily consisted of purpose-built computer programs which were usually written in the high-level language FORTRAN, originally developed in the 1950s by John Backus. From the 1960s to the 1980s this was very much the language of choice, mainly because a large library of numerical algorithms had been developed with it. The availability of these libraries saved behavior geneticists from having to write quite complex code for optimization themselves. Two widely used libraries were MINUIT from the (Centre Européen de Recherche Nucléaire) (CERN) and certain routines from the E04 library of the Numerical Algorithms group (NAg). The latter were developed by Professor Murray and colleagues in the Systems Optimization Laboratory at Stanford University. A key advantage of these routines was that they incorporated methods to obtain numerical estimates of the first and second derivatives, rather than requiring the user to provide them. Alleviated of the burden of finding algebraic expressions for the derivatives, behavior geneticists in the 1970s and 1980s were able to tackle a wider variety of both statistical and substantive problems [3, 7].

Nevertheless, some problems remained which curtailed the widespread adoption of model-fitting by maximum likelihood. Not least of these was that

2 Software for Behavioral Genetics

the geneticist had to learn to use FORTRAN or a similar programming language in order to fit models to their data, particularly if they wished to fit models for which no suitable software was already available. Those skilled in programing were able to assemble loose collections of programs, but these typically involved idiosyncratic formats for data input, program control and interpretation of output. These limitations in turn made it difficult to communicate use of the software to other users, difficult to modify the code for alternative types of data or pedigree structure, and difficult to fit alternative statistical models. Fortunately, the development by Karl Jöreskog and Dag Sörbom of a more general program for maximum likelihood estimation, called LISREL, alleviated many of these problems [1, 8]. Although other programs, such as COSAN, developed by C. Fraser & R. P. McDonald [5] existed, these proved to be less popular with the behavior genetic research community. In part, this was because they did not facilitate the simultaneous analysis of data collected from multiple groups, such as from MZ and DZ twin pairs, which is a prerequisite for estimating heritability and other components of variance. The heart of LISREL's flexibility was its matrix algebra formula for the specification of what are now usually called **structural equation models**. In essence, early versions of the program allowed the user to specify the elements of matrices in the formula:

$$\Sigma = \begin{pmatrix} \Lambda_y \mathbf{A} (\Gamma \Phi \Gamma' + \Psi) \mathbf{A}' \Lambda_y' + \Theta_\epsilon & \Lambda_y \mathbf{A} \Gamma \Phi \Lambda_x' + \Theta_{\delta\epsilon}' \\ \Lambda_x \Phi \Gamma' \mathbf{A}' \Lambda_y' + \Theta_{\delta\epsilon} & \Lambda_x \Phi \Lambda_x' + \Theta_\delta \end{pmatrix}, \quad (1)$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$. The somewhat cumbersome expression (1) is the predicted covariance within a set of dependent y variables (upper left), within a set of independent variables, x (lower right) and between these two sets (lower left and upper right). Using this framework, a wide variety of models may be specified. The program was used in many of the early (1987–1993) workshops on Methodology for Genetic Studies of Twins and Families, and was used in the Neale and Cardon [12] text. What was particularly remarkable about LISREL, COSAN, and similar products that emerged in the 1980s was that they formed a bridge between a completely general programming language such as FORTRAN or

C, and a purpose-built piece of software that was limited to one or two models or types of input data. These programs allowed the genetic epidemiologist to fit, by maximum likelihood, a vast number of models to several types of summary statistic, primarily means and correlation and covariance matrices, and all without the need to write or compile FORTRAN.

Although quite general, some problems could not be tackled easily within the LISREL framework, and others appeared insurmountable. These problems included the following:

1. A common complication in behavior genetic studies is that human families vary in size and configuration, whereas the covariance matrices used as input data assumed an identical structure for each family.
2. Data collected from large surveys and interviews are often incomplete, lacking responses on one or more items from one or more relatives.
3. Genetic models typically specify many linear constraints among the parameters; for example, in the **ACE model** the impact of genetic and environmental factors on the phenotypes is expected to be the same for twin 1 and twin 2 within a pair, and also for MZ and DZ pairs.
4. Certain models – such as those involving mixed genetic and cultural transmission from parent to child [12] – require nonlinear constraints among the parameters.
5. Some models use a likelihood framework that is not based on normal theory.
6. Methods to handle, for example, imperfect diagnosis of identity-by-descent in sib pairs, or zygosity in twin pairs, require the specification of the likelihood as a finite mixture distribution.
7. Tests for **genotype X environment** (or age or sex) interactions may involve continuous or discrete moderator variables.
8. Model specification using matrices is not straightforward especially for the novice.

These issues led to the development, by the author of this article and his colleagues, of the Mx software [10, 11]. Many of the limitations encountered in the version of LISREL available in the early 1990s have been lifted in recent versions (and in LISREL's competitors in the marketplace such as EQS, AMOS and MPLUS). However, at the time of writing, Mx

seems to be the most popular program for the analysis of data from twins and adoptees. In part, this may be due to economic factors, since Mx is freely distributed while the commercial programs cost up to \$900 per user.

Mx was initially developed in 1990, using FORTRAN and the NPSOL numerical optimizer from Professor Walter Murray's group [6]. Fortunately, compiled FORTRAN programs still generate some of the fastest executable programs of any high-level programming language. An early goal of Mx was to liberate the user from the requirement to use the single (albeit quite general) matrix algebra formula that LISREL provided. Therefore, the software included a matrix algebra interpreter, which addressed problem three because large numbers of equality constraints could be expressed in matrix form. It also permitted the analysis of raw data with missing values, by maximizing the likelihood of the observed data, instead of the likelihood of summary statistics, which simultaneously addressed problems one and two. Facilities for matrix specification of linear and non-linear constraints addressed problem four.

Problem five was partially addressed by the advent of user-defined fit functions, which permit parameter estimation under a wider variety of models and statistical theory. In 1994, raw data analysis was extended by permitting the specification of matrix elements to contain variables from the raw data to overcome problem seven. This year also saw the development of a graphical interface to draw path diagrams and fit models directly to the data (problem eight) and the following year mixture distributions were added to address problem six. More recently, developments have focused on the analysis of binary and ordinal data; these permit a variety of Item Response Theory (*see* **Item Response Theory (IRT) Models for Dichotomous Data**) and **Latent Class models** to be fitted relatively efficiently, while retaining the genetic information in studies of twins and other relatives. These developments are especially important for behavior genetic studies, since conclusions about sex-limitation and genotype-environment interaction may be biased by inconsistent measurement [9].

The development of both commercial and non-commercial software continues today. Many of the features developed in Mx have been adopted by the commercial packages, particularly the analysis of raw data. There is also some progress in the development of a package for structural equation model-fitting

package written in the R language (<http://r-project.org>). Being an open-source project, this development is should prove readily extensible. Overall, the feature sets of these programs overlap, so that each program has some unique features and some that are in common with some of the others.

Several new methods, most notably Bayesian approaches involving Monte Carlo Markov Chain (MCMC) (*see* **Markov Chain Monte Carlo and Bayesian Statistics**) algorithms permit greater flexibility in model specification, and in some instances have more desirable statistical properties [2]. For example, estimation of (genetic or environmental or phenotypic) factor scores is an area where the MCMC approach has some clear advantages. Bayesian factor score estimates will incorporate the error inherent in the estimates of the factor loadings, whereas traditional methods will assume that the factor loadings are known without error and are thus artificially precise. Future developments in this area seem highly likely.

Acknowledgment

Michael C. Neale is primarily supported from PHS grants MH-65322 and DA018673.

References

- [1] Baker, L.A. (1986). Estimating genetic correlations among discontinuous phenotypes: an analysis of criminal convictions and psychiatric-hospital diagnoses in Danish adoptees, *Behavior Genetics* **16**, 127–142.
- [2] Eaves, L., & Erkanli, A. (2003). Markov chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and $g \times e$ interaction, *Behavior Genetics* **33**(3), 279–299.
- [3] Eaves, L.J., Last, K., Young, P.A., & Martin, N.G. (1978). Model fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.
- [4] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.
- [5] Fraser, C. (1988). *Cosan User's Guide*. Unpublished Documentation, Centre for Behavioural Studies in Education, University of England, Armidale Unpublished documentation, p. 2351.
- [6] Gill, P.E., Murray, W., & Wright, M.H. (1991). *Numerical Linear Algebra and Optimization*, Vol. 1, Addison-Wesley, New York.
- [7] Jinks, J.L., & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches

4 Software for Behavioral Genetics

- to the analysis of human behavior, *Psychological Bulletin* **73**, 311–349.
- [8] Jöreskog, K.G., & Sörbom, D. (1986). *LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, National Educational Resources, Chicago.
- [9] Lubke, G.H., Dolan, C.V., & Neale, M.C. (2004). Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction, *Twin Research* **7**(3), 292–298.
- [10] Neale, M.C. (1991). *Mx: Statistical Modeling*, Department of Human Genetics, Virginia Commonwealth University, Richmond, VA.
- [11] Neale, M., Boker, S., Xie, G., & Maes, H. (2003). *Mx: Statistical Modeling*, 6th Edition, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA.
- [12] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers.

(See also **Structural Equation Modeling: Software**)

MICHAEL C. NEALE

Software for Statistical Analyses

N. CLAYTON SILVER

Volume 4, pp. 1880–1885

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Software for Statistical Analyses

Introduction

One of the first statistical packages on the market was the Biomedical Statistical Software Package (BMDP), developed in 1960 at the University of California, Los Angeles. One of the reasons for its popularity was that it was written in FORTRAN, which was a major computer language in the 1950s and 1960s.

In 1968, three individuals from Stanford University, Norman Nie, a social scientist, 'Tex' Hull, a programmer with an MBA, and Dale Bent, an operations researcher, developed the Statistical Package for the Social Sciences (SPSS) for analyzing a multitude of social science research data. McGraw-Hill published the first manual for SPSS in 1970. In the mid-1980s, SPSS was first sold for personal computer use [18].

Statistical Analysis Systems (SAS) software was developed in the early 1970s at North Carolina State University by a number of students who discovered that there was no software for managing and analyzing agricultural data. These students wrote the software for a variety of student projects, which provided the impetus for SAS [16].

MINITAB was developed by Barbara Ryan, Thomas Ryan, and Brian Joiner in 1972 in an attempt to make statistics more interesting to students. Because SPSS, SAS, and BMDP were difficult for undergraduates, these innovators constructed a software program that could be learned in about one hour of class time [13]. For an overview of the history of the major statistical software companies, see the statistics site at George Mason University [6].

Modern Statistical Software

The eight general purpose software packages listed in Table 1 perform descriptive statistics, **multiple linear regression**, **analysis of variance (ANOVA)**, **analysis of covariance (ANCOVA)**, **multivariate analysis**, and nonparametric methods (*see* **Distribution-free Inference, an Overview**). For each package, a website and system compatibility are shown.

The hierarchical linear modeling program described in Table 2 estimates multivariate linear models from incomplete data and imports data from other statistical packages. It computes **latent variable** analysis, ordinal and multinomial regression for two-level data (*see* **Logistic Regression**), and **generalized estimating equations** with robust standard errors.

Table 3 lists six programs that perform meta-analyses. Biostat has a variety of **meta-analysis**

Table 1 General statistical packages

Name	Website	Compatibility
BMDP	http://www.statsol.ie/bmdp/bmdp.htm	Windows 95, 98, 2000, NT
JMP	www.jmpdiscovery.com	Windows 95, 98, 2000, NT; Macintosh OS 9.1 or higher
MINITAB	www.minitab.com	Windows 98, 2000, NT, Me, XP
SAS	http://www.sas.com	Windows 95, 98, 2000, NT, XP
SPSS	http://www.spss.com	Windows 95, 98, 2000, NT, XP
STATA	http://www.stata.com	Windows 98, 2000, NT, XP; Macintosh OS X 10.1; UNIX
STATISTICA	http://www.statsoftinc.com	Windows 95, 98, 2000, NT, Me, XP; Macintosh
SYSTAT	http://www.systat.com	Windows 95, 98, 2000, NT, XP

Table 2 Hierarchical linear modeling

Name	Website	Compatibility
HLM-5	http://www.ssicentral.com/hlm/hlm.htm	Windows 95, 98, 2000, NT, XP; UNIX

2 Software for Statistical Analyses

Table 3 Meta-analysis

Name	Website	Compatibility
Biostat	http://www.meta-analysis.com	Windows 95, 98, 2000, NT
Meta	http://users.rcn.com/dakenny/meta.htm	DOS
Meta-analysis	http://www.lyonsmorris.com/MetaA/links.htm	Windows 95, 98, NT, Me; MacIntosh OS
Meta-analysis 5.3	http://www.fu.berlin.de/gesund_engl/meta_e.htm	IBM compatibles (DOS 3.3 or higher)
MetaWin 2.0	http://www.metawinsoft.com	Windows 95, 98, NT
WeasyMA	http://www.weasyrna.com	Windows 95, 98, 2000, NT, XP

Table 4 Power analysis

Name	Website	Compatibility
G*Power	http://www.psych.uni-duesseldorf.de/aap/projects/gpower	DOS or MacIntosh OS systems
Power analysis	http://www.math.yorku.ca/SCS/online/power/	Program on website
PASS-2002	http://www.ncss.com	Windows 95, 98, 2000, NT, Me, XP
Java applets for power	http://www.stat.uiowa.edu/~rlenth/power/index.html	Program on website
Power and precision	http://www.power-analysis.com	Windows 95, 98, 2000, NT
Power on	http://www.macupdate.com/info.php/id/7624	MacIntosh OS X 10.1 or later
StudySize 1.0	http://www.studysize.com/index.htm	Windows 95, 98, 2000, NT, XP

algorithms. It provides **effect size** indices, moderator variables (see **Moderation**), and forest plots. Meta computes and pools effect sizes, and tests whether the average effect size differs from zero. Meta-analysis performs the Hunter-Schmidt method. It computes effect sizes and corrects for range restriction and sampling error. Meta-analysis 5.3 has algorithms utilizing exact probabilities and effect sizes (d or r). The program also provides **cluster analysis** and **stem-and-leaf displays** of correlation coefficients. MetaWin 2.0 computes six different effect sizes and performs cumulative and graphical analyses. It reads text, EXCEL, and Lotus files. WeasyMA performs cumulative analyses and it provides funnel, radial, and forest plots.

Rothstein, McDaniel, and Borenstein [15] provided a summary and a brief evaluative statement (e.g., user friendliness) of eight meta-analytic software programs. They found the programs Meta, True EPISTAT, and DSTAT to be user friendly.

The seven programs listed in Table 4 compute **power analyses**. G*Power, PASS-2002, and Power and Precision compute power, sample size, and effect sizes for t Tests, ANOVA, regression, and chi-square. G*power is downloadable freeware. Power Analysis by Michael Friendly and the Java Applet for Power by Russ Lenth are interactive programs found on their websites. Power Analysis computes power and sample size for one effect in a factorial ANOVA design, whereas the Java Applet for Power computes power for the t Test, ANOVA, and proportions. Power On computes the power and sample sizes needed for t Tests. StudySize 1.0 computes power and sample size for t Tests, ANOVA, and chi-square. It also computes **confidence intervals** and performs **Monte Carlo simulations**.

The website http://www.insp.mx/dinf/stat_list.html lists a number of additional power programs. Yu [21] compared the Power and Precision package [2], PASS, G*Power, and a SAS

Table 5 Qualitative data analysis packages

Name	Website	Compatibility
AtlasT.ti	http://www.atlasti.de/features.shtml	Windows 95, 98, 2000, NT, XP; MacIntosh; SUN
NUD*IST – N6	http://www.qsr.com	Windows 2000, Me, XP
Nvivo 2.0	http://www.qsr.com	Windows 2000, Me, XP

Table 6 Structural equation modeling

Name	Website	Compatibility
AMOS 5	http://www.smallwaters.com/amos/features.html	Windows 98, Me, NT4, 2000, XP
EQS 6.0	http://www.mvsoft.com/eqs60.htm	Windows 95, 98, 2000, NT, XP; UNIX
LISREL 8.5	http://www.ssicentral.com/lisrel/mainlis.htm	Windows 95, 98, 2000, NT, XP; MacIntosh OS 9

Macro developed by Friendly [5]. Yu recommended the Power and Precision package (which is marketed by SPSS, Inc. under the name Sample Power) because of its user-friendliness and versatility. In a review of 29 different power programs [20], the programs nQuery advisor, PASS, Power and Precision, Statistical Power Analysis, Stat Power, and True EPISTAT were rated highest with regard to ease of learning. PASS received the highest mark for ease of use.

Three packages that analyze qualitative data (*see Qualitative Research*) are listed in Table 5. Atlas.ti generates PROLOG code for building knowledge based systems and performs semiautomatic coding with multistring text search and pattern matching. It integrates all relevant material into Hermeneutic Units, and creates and transfers knowledge networks between projects. NUD*IST – N6 provides rapid handling of text records, automated data processing, and the integration of qualitative and quantitative data. It codes questionnaires or focus group data. NVivo 2.0 performs qualitative modeling and integrated searches for qualitative questioning. It provides immediate access to interpretations and insights, and tools that show, shape, filter, and assay data.

Barry [1] compared Atlas.ti and NUD*IST across project complexity, interconnected versus sequential structure, and software design. According to Barry, Atlas.ti's strengths were a well-designed interface that was visually attractive and creative, and its handling of simple sample, one timepoint projects. She believed that NUD*IST had a better searching

structure and was more suitable for complex projects, although it was not as visually appealing as Atlas.ti.

Structural equation modeling packages are listed in Table 6. AMOS 5 fits multiple models into a single analysis. It performs **missing data** modeling (via casewise maximum likelihood), **bootstrap** simulation, **outlier detection**, and multiple fit statistics such as Bentler–Bonnet and Tucker–Lewis indices. EQS 6.0 performs EM-type missing data methods, heterogeneous kurtosis methods, subject weighting methods, multilevel methods, and resampling and simulation methods and statistics. LISREL 8.5 performs structural equation modeling with incomplete data, multilevel structural equation modeling, formal inference based recursive modeling, and multiple imputation and nonlinear multilevel regression modeling.

Kline [8] analyzed the features of AMOS, EQS, and LISREL and concluded that AMOS had the most user-friendly graphical interface; EQS had numerous test statistics and was useful for nonnormal data; LISREL had flexibility in displaying results under a variety of graphical views. He concluded that all three programs were capable of handling many SEM situations (*see Structural Equation Modeling: Software*).

The package in Table 7 not only computes **confidence intervals** for **effect sizes** but it also performs six different simulations that could be used for teaching concepts such as meta-analysis and power. The program runs under Microsoft Excel97 or Excel2000.

4 Software for Statistical Analyses

Table 7 Confidence intervals for effect sizes

Name	Website	Compatibility
ESCI	http://www.latrobe.edu.au/psy/esci/	Windows 95, 98, 2000, NT, Me, XP

Evaluating Statistical Software: A Consumer Viewpoint

Articles comparing statistical software often focus on the accuracy of the statistical calculations and tests of random number generators (e.g., [12]). But, Kuonen and Roehrl [9] list characteristics that may be of more utility to the behavioral researcher. These characteristics are performance (speed and memory); scalability (maximizing computing power); predictability (computational time); compatibility (generalizable code for statistical packages or computers); user-friendliness (easy to learn interfaces and commands); extensibility ('rich, high-level, object-oriented, extensive, and open language' (p. 10)); intelligent agents (understandable error messages); and good presentation of results (easy-to-understand, orderly output). All information should be labeled and presented in an easy to read font type and relatively large font size. Moreover, it should fit neatly and appropriately paged on standard paper sizes.

Kuonen and Roehrl [9] evaluated nine statistical software packages on seven of these characteristics (excluding predictability). They found that no major statistical software package stands out from the rest and that many of them have poor performance, scalability, intelligent agents, and compatibility. For additional reviews of a wider range of major statistical software packages, see www.stats.gla.ac.uk/cti/activities/reviews/alphabet.html.

Problems with Statistical Software Packages

Researchers often blindly follow the statistical software output when reporting results. Many assume that the output is appropriately labeled, perfectly computed, and is based on up-to-date procedures. Indeed, each subsequent version of a statistical software package generally has more cutting edge procedures and is also more user friendly. Unfortunately, there are

still a number of difficulties that these packages have not solved. The examples that follow are certainly not exhaustive, but they demonstrate that researchers should be more cognizant of the theory and concepts surrounding a statistical technique rather than simply parroting output. As one example, behavioral science researchers often correlate numerous measures. Many of these measures contain individual factors that are also contained in the correlation matrix. It is common to have, for example, a 15×15 correlation matrix and associated probability values of tests of the null hypothesis that $\rho = 0$ for each correlation. These probabilities are associated with the standard F , t , or z Tests. Numerous researchers (e.g., [4]) have suggested that multiple F , t , or z tests for testing the null hypothesis that $\rho = 0$, may lead to an inflation of Type I error (*see Multiple Comparison Procedures*). In some cases, the largest correlation in the matrix may have a Type I error rate that is above .40! To guard against this Type I error rate inflation, procedures such as the multistage Bonferroni [10], step up Bonferroni [14] and step down Bonferroni [7], and the rank order method [19] have been proposed. In standard statistical software packages, these and other options are not available, leaving the researcher to resort to either a stand-alone software program or to ignore the issue completely.

As a second example, Levine and Hullett [11] reported that in SPSS for Windows 9.0 (1998), the measure of effect size in the **generalized linear models** procedure was labeled as eta squared, but it should have been labeled as partial eta squared. According to Levine and Hullett, the measure of effect size was correctly labeled in the documentation, but not in the actual printouts. Hence, researchers were more likely to misreport effect sizes for two-way or larger ANOVAs. In some cases, they noted that the effect sizes summed to more than 1.0. Fortunately, in later editions of SPSS for Windows, the label was changed to partial eta squared, so effect size reporting errors should be reduced for output from this and future versions of SPSS.

Finding Other Behavioral Statistical Software

Although many statistical software packages perform a wide variety of statistics, there are times when software packages may not provide a specific hypothesis test. For example, suppose an industrial psychologist correlated salary with job performance at time 1, and after providing a 5% raise, correlated salary and job performance six months later. This constitutes testing the null hypothesis of no statistically significant difference between dependent population correlations with zero elements in common ($\rho_{12} = \rho_{34}$). As an additional example, suppose an educational psychologist was interested in determining if the correlation between grade point average and scholastic aptitude test scores are different for freshmen, sophomores, juniors, and seniors. This tests the null hypothesis that there are no statistically significant differences among independent population correlations ($\rho_1 = \rho_2 = \rho_3 = \rho_4$). Finally, suppose one wanted to orthogonally break a large chi-square contingency table into individual 2×2 contingency tables using the Bresnahan and Shapiro [3] methodology. In all three cases, researchers either need to perform the computational procedures by hand, which is extremely cumbersome, or ask their local statistician to program them. Fortunately, during the past 25 years, many behavioral statisticians published stand-alone programs that augment the standard statistical software packages. In many cases, the programs are available free or a nominal cost. One source of statistical software is a book by Silver and Hittner [17], a compendium of more than 400 peer-reviewed stand-alone statistical programs. They provided a description of each program, its source, compatibility and memory requirements, and information for obtaining the program.

A number of journals publish statistical software. *Applied Psychological Measurement* sporadically includes the Computer Program Exchange that features one-page abstracts of statistical software. *Behavior Research Methods, Instruments, and Computers* is a journal that is devoted entirely to computer applications. The *Journal of Statistical Software* is an Internet peer-reviewed journal that publishes and reviews statistical software, manuals, and user's guides.

A number of website links provide free or nominal cost statistical software. The website ([http://](http://members.aol.com/johnp71/index.html)

members.aol.com/johnp71/index.html), developed by John C. Pezzullo, guides users to free statistical software, some of which are downloads available on a 30-day free trial basis. This extremely well-documented, highly encompassing website lists a wide variety of general packages, subset packages, survey, testing, and measurement software, programming languages, curve fitting and modeling software, and links to other free software. The websites <http://www.statserve.com/software.html> and <http://www.statistics.com/content/commsoft/fulllist.php3> provide lists of statistical software packages that range from general purpose programs such as SPSS to more specific purpose programs such as BILOG3 for Item Response Theory. These websites provide brief descriptions and system compatibilities for more than 100 statistical programs. The website <http://sociology.ca/sociologycalinks.html> has links to a variety of statistical software package tutorials.

The Future of Statistical Software

Although it is difficult to prognosticate the future, there have been trends over the past few years that may continue. First, the major statistical software packages will continue to be upgraded in terms of user-friendly interface and general statistical techniques. This upgrading may include asking questions of the user, similar to that of income tax software, to assure that the design and analysis are appropriate. Help files and error messages will also be more user friendly. Moreover, widely used individual statistical techniques (e.g., meta-analysis and reliability generalization) will continue to be provided as separate programs offered by statistical software companies or by individual statisticians. The programming of less widely used techniques (e.g., testing the difference between two independent **intraclass correlations**), will still be performed by individual statisticians, although, there may be fewer outlets for their publication. Hopefully, there will be additional print or online statistical software journals that will describe computer programs understandably for the applied researcher. Without additional peer-reviews of statistical software for seldom-used techniques, these statistical programs may not meet acceptable standards for quality and quantity.

References

- [1] Barry, C.A. (1998). Choosing qualitative data analysis software: Atlas/ti and NUD*IST compared. *Sociological Research Online*, 3. Retrieved August 18, 2003 from <http://www.socresonline.org.uk/socresonline/3/3/4.html>
- [2] Borenstein, M., Cohen, J. & Rothstein, H. (1997). Power and precision. [Computer software]. Retrieved from <http://www.dataxiom.com>
- [3] Bresnahan, J.L. & Shapiro, M.M. (1966). A general equation and technique for partitioning of chi-square contingency tables, *Psychological Bulletin* **66**, 252–262.
- [4] Collis, B.A. & Rosenblood, L.K. (1984). The problem of inflated significance when testing individual correlations from a correlation matrix, *Journal of Research in Mathematics Education* **16**, 52–55.
- [5] Friendly, M. (1991). SAS Macro programs: mpower [Computer software]. Retrieved from <http://www.math.yorku.ca/SCS/sasmac/mpower/html>
- [6] George Mason University (n.d.). A guide to statistical software. Retrieved October 22, 2003 from www.galaxy.gmu.edu/papers/ast1.html
- [7] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.
- [8] Kline, R.B. (1998). Software programs for structural equation modeling: AMOS, EQS, and LISREL, *Journal of Psychoeducational Assessment* **16**, 343–364.
- [9] Kuonen, D. & Roehrl, A.S.A. (n.d.). Identification of key steps to evaluate statistical software. Retrieved October 20, 2003 from <http://interstat.stat.vt.edu/InterStat/articles/1998/abstracts>
- [10] Larzelere, R.E. & Mulaik, S.A. (1977). Single-sample tests for many correlations, *Psychological Bulletin* **84**, 557–569.
- [11] Levine, T.R. & Hullett, C.R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research, *Human Communication Research* **28**, 612–625.
- [12] McCullough, B.D. (1999). Assessing the reliability of statistical software: part II, *The American Statistician* **53**, 149–159.
- [13] MINITAB (n.d.). Retrieved September 17, 2003, from <http://www.minitab.com/company/identity/History.htm>
- [14] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 663–665.
- [15] Rothstein, H.R., McDaniel, M.A. & Borenstein, M. (2002). Meta-analysis: a review of quantitative cumulation methods, in *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data-Analysis*, F. Drasgow & N. Schmitt, eds, Jossey-Bass, San Francisco, pp. 534–570.
- [16] SAS (n.d.). Retrieved September 17, 2003, from <http://www.sas.com/offices/europe/uk/academic/history/>
- [17] Silver, N.C. & Hittner, J.B. (1998). *A Guidebook of Statistical Software for the Social and Behavioral Sciences*, Allyn & Bacon, Boston.
- [18] SPSS (n.d.). Retrieved September 15, 2003, from <http://www.spss.com/corpinfo/history.htm>
- [19] Stavig, G.R. & Acock, A.C. (1976). Evaluating the degree of dependence for a set of correlations, *Psychological Bulletin* **83**, 236–241.
- [20] Thomas, L. & Krebs, C.J. (1997). A review of statistical power analysis software, *Bulletin of the Ecological Society of America* **78**, 126–139.
- [21] Yu, C. (n.d.). Power analysis. Retrieved September 16, 2003 from <http://seamonkey.ed.asu.edu/~alex/teaching/WBI/power.es.html>

N. CLAYTON SILVER

Spearman, Charles Edward

PAT LOVIE

Volume 4, pp. 1886–1887

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Spearman, Charles Edward

Born: September 10, 1863, in London, England.

Died: September 17, 1945, in London, England.

Charles Spearman's entry into academic life was not by any conventional route. On leaving school, he served as an Army officer, mainly in India, for almost 15 years. It seems, however, that military duties were mixed with reading about philosophy and psychology [10]. In 1897, having just completed a two-year course at the Army Staff College, Spearman resigned his commission and, though entirely self-taught, set out for Germany to study experimental psychology in Wundt's laboratory in Leipzig.

Although Spearman eventually obtained a PhD in 1906, his studies had been interrupted by a recall in 1900 to serve as a staff officer during the South African war. And it was during the few months between his release from these duties early in 1902 and returning to Leipzig that he carried out some mental testing studies on schoolchildren, which laid the foundations for his 'Two-Factor Theory' of human ability as well as his pioneering work on correlation and **factor analysis**. These investigations were published in the *American Journal of Psychology* in 1904 ([6] and [7]) – a remarkable achievement for someone with no academic qualifications other than 'passed staff college'.

Spearman's system hypothesized an underlying factor common to all intellectual activity and a second factor specific to the task; later on, these became known as ***g*** and ***s***. Furthermore, whilst individuals were assumed to possess ***g*** (and ***s***) to different degrees, ***g*** would be invoked to different degrees by different tasks. In Spearman's view, once the effects of superfluous variables and observational errors had been reduced to a minimum, the hierarchical pattern of intercorrelations of measurements on a 'hotch-potch' of abilities gave ample evidence of a common factor with which any intellectual activity was 'saturated' to a specific, measurable degree. In obtaining numerical values for these 'saturation', Spearman had carried out a rudimentary factor analysis, which, at last, promised a way of measuring general intelligence, sought after for so long by those in the field of psychological testing.

However, Spearman's two-factor theory was by no means universally acclaimed. Indeed, for the best part of the next 30 years, Spearman would engage in very public battles with critics on both sides of the Atlantic. Some, such as Edward Thorndike and **Louis Thurstone**, doubted that human ability or intelligence could be captured so neatly. Others, especially **Karl Pearson** (whose correlational methods Spearman had adapted), **Godfrey Thomson**, **William Brown** (early in his career), and E.B. Wilson saw grave faults in Spearman's mathematical and statistical arguments (see [4]). Spearman's Herculean effort to establish the two-factor theory as the preeminent model of human intelligence reached its peak in 1927 with the publication of *The Abilities of Man* [9]. But, more sophisticated, multiple factor theories would gradually overshadow this elegantly simple system.

In 1907, Spearman had returned to England to his first post at University College, London (UCL) as part-time Reader in Experimental Psychology. Four years later, he was appointed as the Grote Professor of Mind and Logic and head of psychology and, finally, in 1928, he became Professor of Psychology until his retirement as Emeritus Professor in 1931. He was elected a Fellow of the Royal Society in 1924 and received numerous other honours.

Perhaps Spearman's chief legacy was to put British psychology on the international map by creating the first significant centre of psychological research in the country. The 'London School', as it became known, was renowned for its rigorous pursuit of the scientific and statistical method for studying human abilities – an approach entirely consonant with principles advocated by **Francis Galton** in the previous century.

Nowadays, however, Spearman is remembered almost solely for his correlational work, especially the rank correlation coefficient (see **Spearman's Rho**) (although it is not entirely clear that the version we know today is in fact what Spearman developed [2]), and the so-called Spearman–Brown reliability formula. Although for a long time there were doubts and heated debates (mainly because of claims and stories put about by **Cyril Burt**, a former protégé and his successor as Professor of Psychology at UCL) about who exactly was the originator of factor analysis, Spearman's status as its creator is now firmly established (see [1] and [3]).

And yet, paradoxically, Spearman himself regarded this psychometric and statistical work as secondary

to his far more ambitious mission – establishing fundamental laws of psychology, which would encompass not just the processes inherent in the two-factor theory, but all cognitive activity (see [8]). In spite of his own hopes and claims, Spearman never succeeded in developing this work much beyond an embryonic system. Ironically, though, some of his key ideas have recently reemerged within cognitive psychology.

After retirement, Spearman had continued publishing journal articles and books as well as travelling widely. By the early 1940s, however, he was in failing health and poor spirits. His only son had been killed during the evacuation of Crete in 1941 and he was suffering from blackouts which made working, and life in general, a trial. A bad fall during such an episode in the late summer of 1945 led to a bout of pneumonia. He was admitted to University College Hospital where he took his own life by jumping from a fourth floor window. Spearman believed in the right of individuals to decide when their lives should cease.

Further material about Charles Spearman's life and work can be found in [5] and [10].

References

- [1] Bartholomew, D.J. (1995). Spearman and the origin and development of factor analysis, *British Journal of Mathematical and Statistical Psychology* **48**, 211–220.
- [2] Lovie, A.D. (1995). Who discovered Spearman's rank correlation? *British Journal of Mathematical and Statistical Psychology* **48**, 255–269.
- [3] Lovie, A.D. & Lovie, P. (1993). Charles Spearman, Cyril Burt, and the origins of factor analysis, *Journal of the History of the Behavioral Sciences* **29**, 308–321.
- [4] Lovie, P. & Lovie, A.D. (1995). The cold equations: Spearman and Wilson on factor indeterminacy, *British Journal of Mathematical and Statistical Psychology* **48**, 237–253.
- [5] Lovie, P. & Lovie, A.D. (1996). Charles Edward Spearman, F.R.S. (1863–1945), *Notes and Records of the Royal Society of London* **50**, 1–14.
- [6] Spearman, C. (1904a). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.
- [7] Spearman, C. (1904b). "General intelligence" objectively determined and measured, *American Journal of Psychology* **15**, 202–293.
- [8] Spearman, C. (1923). *The Nature of 'Intelligence' and the Principles of Cognition*, Macmillan Publishing, London.
- [9] Spearman, C. (1927). *The Abilities of Man, their Nature and Measurement*, Macmillan Publishing, London.
- [10] Spearman, C. (1930). C. Spearman, in *A History of Psychology in Autobiography*, Vol. 1, C. Murchison, ed., Clark University Press, Worcester, pp. 299–331.

PAT LOVIE

Spearman's Rho

DIANA KORNBROT

Volume 4, pp. 1887–1888

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Spearman's Rho

Spearman's rho, r_s , is a measure of correlation based on ranks (see **Rank Based Inference**). It is useful when the raw data are ranks, as for example job applicants, or when data are ordinal. Examples of ordinal data include common rating scales based on responses ranging from 'strongly disagree' to 'strongly agree'. Figure 1 shows examples of metric data where r_s is useful. Panel B demonstrates a nonlinear monotonic relation. Panel C demonstrates the effect of an 'outlier'.

Spearman's rho is simply the normal **Pearson product moment correlation** r computed on the ranks of the data rather than the raw data.

Calculation

In order to calculate r_s , it is first necessary to rank both the X and Y variable as shown in Table 1. Then for each pair of ranked values the difference between the ranks, D , is calculated, so that the simple formula in (1) [1] can be used to calculate r_s

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}. \quad (1)$$

Equation (1) overestimates r_s if there are ties. Equations for adjusting for ties exist, but are cumbersome. The simplest method [1] is to use averaged ranks, where tied values are all given the average rank of their positions. Thus, a data set (1, 2, 2, 4) would have ranks (1, 2.5, 2.5, 4).

Table 1 Ranking of data to calculate r_s

X	2	3	5	7	8	40
Y	8	4	5	6	7	1
Rank X	6	5	4	3	2	1
Rank Y	1	5	4	3	2	6
D^2	25	0	0	0	0	25

Hypothesis Testing

For the null hypothesis of no association, that is, $r_s = 0$, and $N > 10$, (2) gives a statistic that is t -distributed with $N - 2$ degrees of freedom [1]

$$t = \frac{r_s \sqrt{N - 2}}{\sqrt{1 - r_s^2}}. \quad (2)$$

Accurate Tables for $N \leq 10$ are provided by Kendall & Gibbons [2].

Confidence limits and hypotheses about values of r_s other than 0 can be obtained by noting that the Fisher transformation gives a statistic z_r that is normally distributed with variance $1/(N - 3)$

$$z_r = \frac{1}{2} \ln \left[\frac{1 + r_s}{1 - r_s} \right]. \quad (3)$$

Comparison with Pearson's r and Kendall's τ

Figure 1 illustrates comparisons of r , r_s and **Kendall's tau**, τ . For normally distributed data with a linear relation, parametric tests based on r are usually more powerful than rank tests based on either r_s or τ . So in panel A: $r = 0.46$, $p = 0.044$; $r_s = 0.48$,

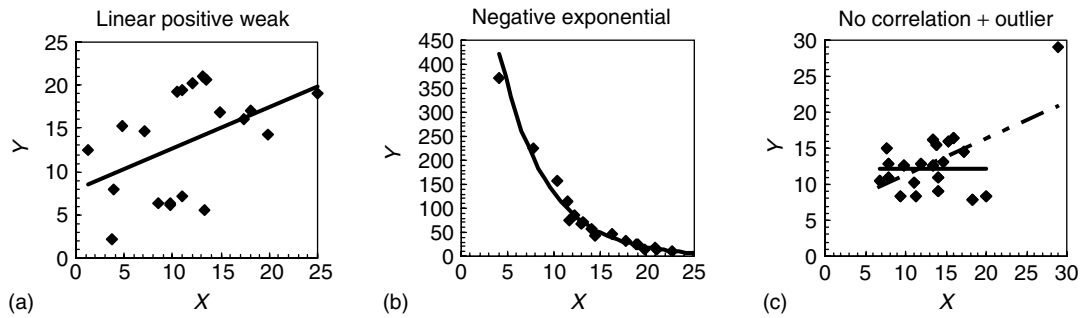


Figure 1 Three relations between Y and X to demonstrate comparisons between r , r_s , and τ

2 Spearman's Rho

$p = 0.032$; $\tau = 0.32$, $p = 0.052$. Panel B shows a nonlinear relation, with the rank coefficients better at detecting a trend: $r = -0.82$; $r_s = -0.99$; $\tau = -0.96$, all of course highly significant. Panel C shows that rank coefficients provide better protection against outliers: $r = 0.56$, $p = 0.010$; $r_s = 0.18$, $p = 0.443$; $\tau = 0.15$, $p = 0.364$. The outlier point (29, 29) no longer causes a spurious significant correlation. Experts [1, 2] recommend τ over r_s as the best rank based procedure, but r_s is far easier to calculate if a computer package is not available.

References

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.
- [2] Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th Edition, Edward Arnold, London & Oxford.

DIANA KORNROT

Sphericity Test

H.J. KESELMAN

Volume 4, pp. 1888–1890

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Sphericity Test

Repeated measures (*see* **Repeated Measures Analysis of Variance**) or longitudinal designs (*see* **Longitudinal Data Analysis**) are used frequently by researchers in the behavioral sciences where **analysis of variance** F tests are typically used to assess treatment effects. However, these tests are sensitive to violations of the assumptions on which they are based, particularly when the design is unbalanced (i.e., group sizes are unequal) [6].

To set the stage for a description of sphericity, consider a hypothetical study described by Maxwell and Delaney [10, p. 534] where 12 subjects are observed in each of four conditions (factor K), for example, at 30, 36, 42, and 48 months of age, and the dependent measure (Y) 'is the child's age-normed general cognitive score on the McCarthy Scales of Children's Abilities'. In this simple repeated measures design, the validity of the within-subjects main effects F test of factor K rests on the assumptions of normality, independence of errors, and homogeneity of the treatment-difference variances (i.e., circularity or sphericity) [3, 14, 15]. Homogeneity of treatment-difference variances means that for all possible differences in scores among the levels of the repeated measures variable [i.e., $Y(30) - Y(36)$, $Y(30) - Y(42)$, $\dots$, $Y(42) - Y(48)$], the population variances of these differences are equal. For designs including a between-subjects grouping factor (J), the validity of the within-subjects main and interaction tests ($F_{(K)}$ and $F_{(JK)}$) rest on two assumptions, in addition to those of normality and independence of errors. First, for each level of the between-subjects factor J , the population treatment-difference variances among the levels of K must be equal. Second, the population covariance matrices (*see* **Correlation and Covariance Matrices**) at each level of J must be equal. Since the data obtained in many disciplines rarely conform to these requirements, researchers using these traditional procedures will erroneously claim treatment effects when none are present, thus filling their literature with false positive claims.

McCall and Appelbaum [12] provide an illustration as to why in many areas of behavioral science research (e.g., developmental psychology, learning psychology), the covariances between the levels of the repeated measures variable will not conform to

the required covariance pattern for a valid univariate F test. They use an example from developmental psychology to illustrate this point. Specifically, adjacent-age assessments typically correlate more highly than developmentally distant assessments (e.g., 'IQ at age three correlates 0.83 with IQ at age four but 0.46 with IQ at age 12'); this type of correlational structure does not correspond to a circular (spherical) covariance structure. That is, for many applications, successive or adjacent measurement occasions are more highly correlated than non-adjacent measurement occasions, with the correlation between these measurements decreasing the farther apart the measurements are in the series. Indeed, as McCall and Appelbaum note 'Most longitudinal studies using age or time as a factor cannot meet these assumptions' (p. 403). McCall and Appelbaum also indicate that the covariance pattern found in learning experiments would also not likely conform to a spherical pattern. As they note, 'experiments in which change in some behavior over short periods of time is compared under several different treatments often cannot meet covariance requirements' (p. 403).

When the assumptions of the conventional F tests have been satisfied, the tests will be valid and will be uniformly most powerful for detecting treatment effects when they are present. These conventional tests are easily obtained with the major statistical packages (e.g., SAS [16] and SPSS [13]; *see* **Software for Statistical Analyses**). Thus, when assumptions are known to be satisfied, behavioral science researchers can adopt the conventional procedures and report the associated P values, since, under these conditions, these values are an accurate reflection of the probability of observing an F value as extreme or more than the observed F statistic.

The result of applying the conventional tests of significance to data that do not conform to the assumptions of (multisample) sphericity will be that too many null hypotheses will be falsely rejected [14]. Furthermore, as the degree of non-sphericity increases, the conventional repeated measures F tests become increasingly liberal [14].

When sphericity/circularity does not exist, the Greenhouse and Geisser [2] and Huynh and Feldt [4] tests are robust alternatives to the traditional tests, provided that the design is balanced or that the covariance matrices across levels of J are equal (*see* **Repeated Measures Analysis of Variance**). The

empirical literature indicates that the Greenhouse and Geisser and Huynh and Feldt adjusted degrees of freedom tests are robust to violations of multisample sphericity as long as group sizes are equal [6]. The P values associated with these statistics will provide an accurate reflection of the probability of obtaining these adjusted statistics by chance under the null hypotheses of no treatment effects. The major statistical packages [SAS, SPSS] provide Greenhouse and Geisser and Huynh and Feldt adjusted P values. However, the Greenhouse and Geisser and Huynh and Feldt tests are not robust when the design is unbalanced [6].

In addition to the Geisser and Greenhouse [2] and Huynh and Feldt [4] corrected degrees of freedom univariate tests, other univariate, multivariate, and hybrid analyses are available to circumvent the restrictive assumption of (multisample) sphericity/circularity. In particular, Johansen's [5] procedure has been found to be robust to violations of multisample sphericity in unbalanced repeated measures designs (see [8]). I refer the reader to [6], [7], and [9].

I conclude by noting that one can assess the (multisample) sphericity/circularity assumption with formal test statistics. For completely within-subjects repeated measures designs, sphericity can be checked with Mauchly's [11] W -test. If the design also contains between-subjects grouping variables, then multisample sphericity is checked in two stages (see [3]). Specifically, one can test whether the population covariance matrices are equal across the between-subjects grouping variable(s) with Box's modified criterion M (see [3]), and if this hypothesis is not rejected, whether sphericity exists (with Mauchly's W -test). However, these tests have been found to be problematic; that is, according to Keselman et al. [8] 'These tests indicate that even when data is obtained from normal populations, the tests for circularity (the M and W criteria) are sensitive to all but the most minute departures from their respective null hypotheses, and consequently the circularity hypothesis is not likely to be found tenable' (p. 481). Thus, it is recommended that researchers adopt alternative procedures, as previously noted, for assessing the effects of repeated measures/longitudinal variables. Lastly, it should be noted that Boik [1] discusses multisample sphericity for repeated measures designs containing multiple dependent variables.

References

- [1] Boik, R.J. (1991). The mixed model for multivariate repeated measures: validity conditions and an approximate test, *Psychometrika* **53**, 469–486.
- [2] Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data, *Psychometrika* **24**, 95–112.
- [3] Huynh, H. & Feldt, L. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F distributions, *Journal of the American Statistical Association* **65**, 1582–1589.
- [4] Huynh, H. & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs, *Journal of Educational Statistics* **1**, 69–82.
- [5] Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression, *Biometrika* **67**, 85–92.
- [6] Keselman, H.J., Algina, A., Boik, R.J. & Wilcox, R.R. (1999). New approaches to the analysis of repeated measurements, in *Advances in Social Science Methodology*, Vol. 5, B. Thompson, ed., JAI Press 251–268.
- [7] Keselman, H.J., Carriere, K.C. & Lix, L.M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous, *Journal of Educational Statistics* **18**, 305–319.
- [8] Keselman, H.J., Rogan, J.C., Mendoza, J.L. & Breen, L.J. (1980). Testing the validity conditions of repeated measures F tests, *Psychological Bulletin* **87**(3), 479–481.
- [9] Keselman, H.J., Wilcox, R.R. & Lix, L.M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs, *Psychophysiology* **40**, 586–596.
- [10] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data. A Model Comparison Perspective*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [11] Mauchly, J.L. (1940). Significance test for sphericity of a normal n -variate distribution, *The Annals of Mathematical Statistics* **29**, 204–209.
- [12] McCall, R.B. & Appelbaum, M.I. (1973). Bias in the analysis of repeated-measures designs: some alternative approaches, *Child Development* **44**, 401–415.
- [13] Norusis, M.J. (2004). *SPSS 12.0 Statistical Procedures Companion*, SPSS Inc., Prentice Hall, Upper Saddle River, NJ.
- [14] Rogan, J.C., Keselman, H.J. & Mendoza, J.L. (1979). Analysis of repeated measurements, *British Journal of Mathematical and Statistical Psychology* **32**, 269–286.
- [15] Rouanet, H. & Lepine, D. (1970). Comparison between treatments in a repeated measures design: ANOVA and multivariate methods, *British Journal of Mathematical and Statistical Psychology* **23**, 147–163.
- [16] SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 8*, 3rd Edition, SAS Institute, Cary.

(See also **Linear Multilevel Models**)

H.J. KESELMAN

Standard Deviation

DAVID CLARK-CARTER

Volume 4, pp. 1891–1891

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Standard Deviation

The standard deviation (SD) is a measure of spread or dispersion. It is defined as the (positive) square root of the variance. Thus, we can find the SD of a population, or of a sample of data, or estimate the SD of a population, by taking the square root of the appropriate version of the calculated variance. The standard deviation for a sample is often represented as S , while for the population, it is denoted by σ .

The standard deviation has an advantage over the variance because it is expressed in the same units as the original measure, whereas the variance is in squared units of the original measure. However, it is still affected by extreme scores.

The SD is a way of putting the mean of a set of values in context. It also facilitates comparison

of the distributions of several samples by showing their relative spread. Moreover, the standard deviation of the distribution of various statistics is also called the **standard error**. The standard error of the mean (SEM), for instance, is important in inferential procedures such as the t Test.

Finally, if a set of data has a normal distribution, then approximately 68% of the population will have a score within the range of one standard deviation below the mean to one standard deviation above the mean. Thus, if IQ were normally distributed and the mean in the population were 100 and the standard deviation were 15, then approximately 68% of people from that population would have an IQ of between 85 and 115 points.

DAVID CLARK-CARTER

Standard Error

DAVID CLARK-CARTER

Volume 4, pp. 1891–1892

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Standard Error

Any statistic is a random variable and, thus, has its own distribution, called a **sampling distribution**. The standard error is the standard deviation of the sampling distribution of a statistic. The most commonly encountered standard error is the standard error of the mean (SEM), which is a measure of the spread of means of samples of the same size from a specific population. Imagine that a sample of a given size is taken randomly from a population and the mean for that sample is calculated and this process is repeated an infinite number of times from the same population. The standard deviation of the distribution of these means is the standard error of the mean. It is found by dividing the standard deviation (SD) for the

population by the square root of the sample size (n):

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}}. \quad (1)$$

Suppose that the population standard deviation of people's recall of words is known to be 4.7 (though usually, of course, we do not know the population SD and must estimate it from the sample), and that we have a sample of six participants, then the standard error of the mean number of words recalled would be $4.7/\sqrt{6} = 1.92$.

The standard error of the mean is a basic element of parametric hypothesis tests on means, such as the z -test and the t Test, and of confidence intervals for means.

DAVID CLARK-CARTER

Standardized Regression Coefficients

RONALD S. LANDIS

Volume 4, pp. 1892–1892

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Standardized Regression Coefficients

Standardized regression coefficients, commonly referred to as beta weights (β), convey the extent to which two standardized variables are linearly related in regression analyses (see **Multiple Linear Regression**). Mathematically, the relationship between unstandardized (b) weights and standardized (β) weights is

$$b = \beta \frac{\sigma_y}{\sigma_x} \quad \text{or} \quad \beta = b \frac{\sigma_x}{\sigma_y} \quad (1)$$

where σ_x and σ_y are the standard deviations of the predictor and criterion variables respectively. Because standardized coefficients reflect the relationship between a predictor and criterion variable after converting both the z-scores, beta weights vary between -1.00 and $+1.00$.

In the simple regression case, a standardized regression coefficient is equal to the correlation (r_{xy}) between the predictor and criterion variable. With multiple predictors, however, the predictor intercorrelations must be controlled for when computing standardized regression coefficients. For example, in situations with two predictor variables, the standardized coefficients (β_1 and β_2) are computed as

$$\begin{aligned} \beta_1 &= \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \\ \beta_2 &= \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \end{aligned} \quad (2)$$

where r_{y1} and r_{y2} are the zero-order correlations between each predictor and the criterion and r_{12}^2 is the squared correlation between the two predictor variables.

Like unstandardized coefficients, standardized coefficients reflect the degree of change in the

criterion variable associated with a unit change in the predictor. Since the standard deviation of a standardized variable is 1, this coefficient is interpreted as the associated standard deviation change in the criterion.

Standardized regression coefficients are useful when a researcher's interest is the estimation of predictor–criterion relationships, independent of the original units of measure. For example, consider two researchers studying the extent to which cognitive ability and conscientiousness accurately predict academic performance. The first researcher measures cognitive ability with a 50-item, multiple-choice test; conscientiousness with a 15-item, self-report measure; and academic performance with college grade point average (GPA). By contrast, the second researcher measures cognitive ability with a battery of tests composed of hundreds of items, conscientiousness through a single-item peer rating, and academic performance through teacher ratings of student performance. Even if the correlations between the three variables are identical across these two situations, the unstandardized regression coefficients will differ, given the variety of measures used by the two researchers. As a result, direct comparisons between the unstandardized coefficients associated with each of the predictors across the two studies cannot be made because of scaling differences. Standardized regression weights, on the other hand, are independent of the original units of measure. Thus, a direct comparison of relationships across the two studies is facilitated by standardized regression weights, much like correlations facilitate generalizations better than covariances. This feature of standardized regression weights is particularly appealing to social scientists who (a) frequently cannot attach substantive meaning to scale scores and (b) wish to compare results across studies that have used different scales to measure specific variables.

RONALD S. LANDIS

Stanine Scores

DAVID CLARK-CARTER

Volume 4, pp. 1893–1893

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Stanine Scores

A stanine score is a type of standardized score. Instead of standardizing the original scores to have a mean of 0 and standard deviation (SD) of 1, as is the case of **z-scores**, the scores are transformed into a nine-point scale; hence, the name stanine as an abbreviation for standard nine. The original scores are generally assumed to be normally distributed or to have been ‘normalized’ by a normalizing transformation. The transformation to stanine scores produces a distribution with a mean of 5 and a standard deviation of 1.96.

Table 1 Percentages of the distribution for each stanine score

Stanine score	1	2	3	4	5	6	7	8	9
Percentage	4	7	12	17	20	17	12	7	4

The percentages of a distribution falling into each stanine score are shown in Table 1.

We see from the table that, for example, 11% of the distribution will have a stanine score of 2 or less; in other words, 11% of a group of people taking the test would achieve a stanine score of either 1 or 2. As with any standardized scoring system, stanine scores allow comparisons of the scores of different people on the same measure or of the same person over different measures.

The development of stanine scoring is usually attributed to the US Air Force during the Second World War [1].

Reference

[1] Anastasi, A. & Urbina, S. (1997). *Psychological Testing*, 7th Edition, Prentice Hall, Upper Saddle River.

DAVID CLARK-CARTER

Star and Profile Plots

BRIAN S. EVERITT

Volume 4, pp. 1893–1894

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Star and Profile Plots

Plotting multivariate data is often a useful step in gathering insights into their possible structure; these insights may be useful in directing later, more formal analyses (see **Multivariate Analysis: Overview**). An excellent review of the many possibilities is given in [1]. For comparing the relative variable values of the observations in small- or moderate-sized multivariate data sets, two similar approaches, the star plot and the profile plot, can, on occasions, be helpful.

In the star plot, each multivariate observation (suitably scaled) is represented by a ‘star’ consisting of a sequence of equiangular spokes called *radii*, with each spoke representing one of the variables. The length of a spoke is proportional to the variable value it represents relative to the maximum magnitude of the variable across all the observations in the sample. A line is drawn connecting the data values for each spoke.

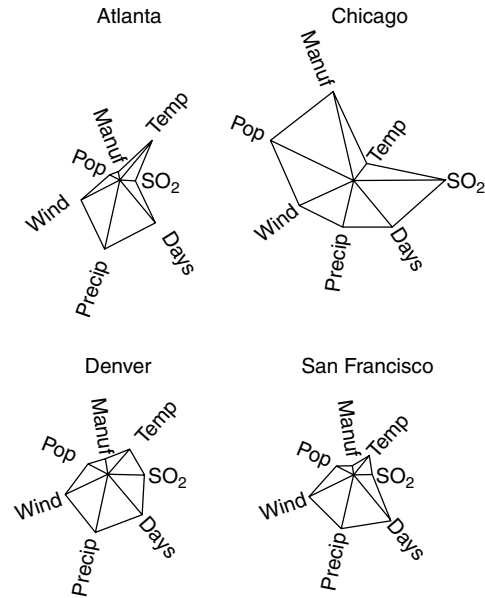


Figure 1 Star plots for air pollution data from four cities in the United States

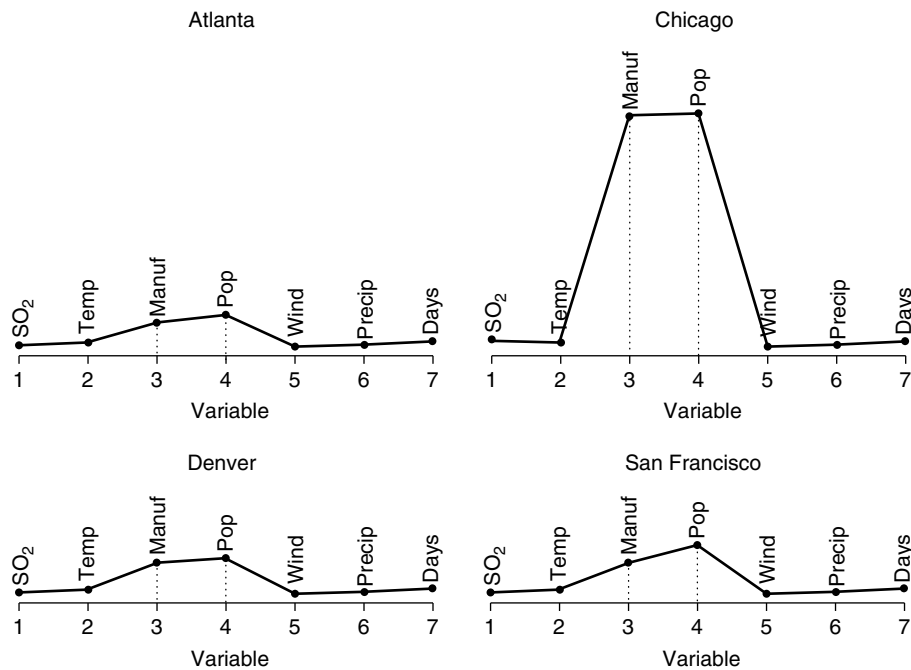


Figure 2 Profile plots for air pollution data from four cities in the United States

2 Star and Profile Plots

Table 1 Air pollution data for four cities in the United States

	SO ₂ (μg/m ³)	Temp (°F)	Manuf	Pop	Wind (miles/h)	Inches (miles/h)	Days
Atlanta	24	61.5	368	497	9.1	48.34	115
Chicago	110	50.6	3344	3369	10.4	34.44	122
Denver	17	51.9	454	515	9.0	12.95	86
San Francisco	12	56.7	453	71.6	8.7	20.66	67

SO₂: Sulphur dioxide content of air,

Temp: Average annual temperature,

Manuf: Number of manufacturing enterprises employing 20 or more workers,

Pop: Population size,

Wind: Average annual wind speed,

Precip: Average annual precipitation,

Days: Average number of days with precipitation per year.

In a profile plot, a sequence of equispaced vertical spikes is used, with each spike representing one of the variables. Again, the length of a given spike is proportional to the magnitude of the variable it represents relative to the maximum magnitude of the variable across all observations.

As an example, consider the data in Table 1 showing the level of air pollution in four cities in the United States along with a number of other climatic and human ecologic variables.

The star plots of the four cities are shown in Figure 1 and the profile plots in Figure 2. In both diagrams, Chicago is clearly identified as being very different from the other cities. In the profile

plot, the remaining three cities appear very similar, but in the star plot, Atlanta is identified as having somewhat different characteristics from the other two.

Star plots are available in some software packages, for example, *S-PLUS*, and profile plots are easily constructed using the command line language of the same package.

Reference

- [1] Carr, D.B. (1998). Multivariate graphics, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester.

BRIAN S. EVERITT

State Dependence

ANDERS SKRONDAL AND SOPHIA RABE-HESKETH

Volume 4, pp. 1895–1895

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

State Dependencezy

It is often observed in the behavioral sciences that the current outcome of a dynamic process depends on prior outcomes, even after controlling or adjusting for covariates. The outcome is often categorical with different categories corresponding to different ‘states’, giving rise to the term *state dependence*.

Examples of state dependence include (a) the elevated risk of committing a crime among previous offenders and (b) the increased probability of experiencing future unemployment among those currently unemployed.

Let y_{it} be the state for unit or subject i at time or occasion t and $\mathbf{x}_{it}$ a vector of observed covariates. For simplicity, we consider dichotomous states ($y_{it} = 1$ or $y_{it} = 0$) and two occasions $t = 1, 2$. State dependence then occurs if

$$\Pr(y_{i2}|\mathbf{x}_{i2}, y_{i1}) \neq \Pr(y_{i2}|\mathbf{x}_{i2}). \quad (1)$$

Statistical models including state dependence are often called *autoregressive models*, *Markov models* (see **Markov Chains**) [4] or *transition models* [1].

James J. Heckman, [2, 3] among others, has stressed the importance of distinguishing between true and spurious state dependence in social science. In the case of *true state dependence*, the increased probability of future unemployment is interpreted as ‘causal’. For instance, a subject having experienced unemployment may be less attractive employers than an identical subject not having experienced unemployment. Alternatively, state dependence may be apparent, which is called *spurious*

state dependence. In this case, past unemployment has no ‘causal’ effect on future unemployment. It is rather unobserved characteristics of the subject (unobserved heterogeneity) not captured by the observed covariates $\mathbf{x}_{it}$ that produce the dependence over time. Some subjects are just more prone to experience unemployment than others, perhaps because they are not ‘suitable’ for the labor market, regardless of their prior unemployment record and observed covariates.

Letting ζ_i denote unobserved heterogeneity for subject i , spurious state dependence occurs if there is state dependence as in the first equation, but the dependence on the previous state disappears when we condition on ζ_i ,

$$\Pr(y_{i2}|\mathbf{x}_{i2}, y_{i1}, \zeta_i) = \Pr(y_{i2}|\mathbf{x}_{i2}, \zeta_i). \quad (2)$$

Referenceszy

- [1] Diggle, P.J., Heagerty, P.J., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford.
- [2] Heckman, J.J. (1981). Heterogeneity and state dependence, in *Studies in Labor Markets*, S. Rosen ed., Chicago University Press, Chicago, pp. 91–139.
- [3] Heckman, J.J. (1991). Identifying the hand of past: distinguishing state dependence from heterogeneity, *American Economic Review* **81**, 75–79.
- [4] Lindsey, J.K. (1999). *Models for Repeated Measurements*, 2nd Edition, Oxford University Press, Oxford.

ANDERS SKRONDAL AND
SOPHIA RABE-HESKETH

Statistical Models

DAVID A. KENNY

Volume 4, pp. 1895–1897

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Statistical Models

Statistical analysis, like breathing, is so routinely performed that we typically do not understand very well exactly what it is that we are doing. Statistical modeling begins with measurements that are attached to units (*see* **Measurement: Overview**). Very often the units are persons, but they may be other animals, objects, or events. It is important to understand that a measurement may refer not to a single entity, but possibly to multiple entities. For example, a measure of aggression refers to at least two people, the victim and the perpetrator, not just one person. Sometimes a measurement refers to multiple levels. So, for instance, children can be in classrooms and classrooms can be in schools. Critical then in statistical analysis is determining the appropriate units and levels.

An essential part of statistical analysis is the development of a statistical model. The model is one or more equations. In each equation, a variable or set of variables are explained. These variables are commonly called *dependent variables* in experimental studies or *outcome variables* in nonexperimental studies. For these variables, a different set of variables are used to explain them and, hence, are called *explanatory variables*.

Some models are causal models (*see* **Linear Statistical Models for Causation: A Critical Review**) for which there is the belief that a change in an explanatory variable changes the outcome variable. Other models are predictive in the sense that only a statistical association between variables is claimed.

Normally, a statistical model has two different parts. In one part, the outcome variables are explained by explanatory variables. Thus, some of the variation of a variable is systematically explained. The second part of the model deals with what is not explained and is the random piece of the model. When these two pieces are placed together, the statistical model can be viewed as [1]

$$\text{DATA} = \text{FIT} + \text{ERROR}. \quad (1)$$

The ‘fit’ part can take on several different functional forms, but very often the set of explanatory variables is assumed to have additive effects. In deciding whether to have a variable in a model, a

judgment of the improvement fit and, correspondingly, reduction of error must be made.

A fundamental difficulty in statistical analysis is how much complexity to add to the model. Some models have too many terms and are too complex, whereas other models are too simple and do not contain enough variables. A related problem is how to build such a model. One can step up and start with no variables and keep adding variables until ‘enough’ variables are added. Alternatively, one can step down and start with a large model and strip out of the model unnecessary variables. In some cases, the most complicated model can be stated. Such a model is commonly referred to as the *saturated model*.

Not all explanatory variables in statistical analysis are the same. Some variables are the essential variables of interest and the central question of the research is their effect. Other variables, commonly called *covariates*, are not of central theoretical interest. Rather, they are included in the model for two fundamentally different reasons. Sometimes they are included because they are presumed to reduce error. At other times, they are included because they are correlated with a key explanatory variable and so their effects need to be controlled.

Decisions about adding and removing a term from a model involves model comparison; that is, the fit of two models are compared: Two models are fit, one of which includes the explanatory variable and the other does not. If the variable is needed in the model, the fit of the latter model would be better than that of the former.

Another issue in statistical analysis is the scaling of variables in statistical models. Sometimes variables have an arbitrary scaling. For instance, gender might be dummy coded (0 = male and 1 = female) or effect coded (−1 = male and 1 = female) coding. The final model should be the very same model, regardless of the coding method used.

The errors in the model are typically assumed to have some sort of distribution. One commonly assumed distribution that is assumed is the normal distribution (*see* **Catalogue of Probability Density Functions**). Very typically, assumptions are made that units are randomly and independently sampled from that distribution.

Classically, in statistical analysis, a distinction is made concerning descriptive versus inferential statistics. Basically, descriptive statistics concerns the estimation of the model and its parameters. For

instance, if a **multiple regression** equation were estimated, the descriptive part of the model concerns the coefficients, the intercept and slopes, and the error variance. Inferential statistics focuses on decision making: Should a variable be kept in the model or are the assumptions made in the model correct?

Model building can either be confirmatory or exploratory [2] (see **Exploratory Data Analysis**). In confirmatory analyses, the steps are planned in advance. For instance, if there are four predictor variables in a model, it might be assumed that three of the variables were important and one was not. So, we might first see if the fourth variable is needed, and then test that the three that were specified are in fact important. In exploratory analyses, researchers go where the data take them. Normally, statistical analyses are a mix of exploratory and confirmatory analyses. While it is often helpful to ask the data a preset set of questions, it is just as important to let the data provide answers to questions that were not asked.

A critical feature of statistical analysis is an understanding of how much the data can tell us. One obvious feature is sample size. More complex models can be estimated with larger sample sizes.

However, other features are important. For instance, the amount of variation, the design of research, and the precision of the measuring instruments are important to understand. All too often, researchers fail to ask enough of their data and so perform too limited statistical analyses. Alternatively, other researchers ask way too much of their data and attempt to estimate models that are too complex for the data. Finding the right balance can be a difficult challenge.

References

- [1] Judd, C.M. & McClelland, G.H. (1989). *Data Analysis: A Model-Comparison Approach*, Harcourt Brace Jovanovich, San Diego.
- [2] Tukey, J.W. (1969). Analyzing data: sanctification or detective work? *American Psychologist* **24**, 83–91.

(See also **Generalized Linear Mixed Models**; **Generalized Linear Models (GLM)**; **Linear Multilevel Models**)

DAVID A. KENNY

Stem and Leaf Plot

SANDY LOVIE

Volume 4, pp. 1897–1898

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Stem and Leaf Plot

Stem and leaf plots, together with **box plots**, form important parts of the graphical portfolio bequeathed us by **John Tukey**, the inventor of **exploratory data analysis** (EDA). These plots trade in on the unavoidable redundancy found in Western number systems in order to produce a more compact display. They can also be thought of as *value-added histograms* in that while both displays yield the same distributional information, stem and leaf plots allow the original sample to be easily reconstituted, thus enabling the analyst to quickly read off such useful measures as the median and the upper and lower **quartiles**.

The stem in the plot is the most significant part of a number, that is, the largest or most left-hand value, while the leaf is the increasingly less significant right-hand parts of a number. So in the number 237, 2 could be the stem, thus making 3 and 7 the leaves. In most instances, a sample will contain relatively few *differently* valued stems, and hence many redundant stems, while the same sample is likely to have much more variation in the value of the leaves, and hence less redundancy in the leaf values. What Tukey did to reduce this redundancy was to create a display made up of a single vertical column of numerically ordered stems, with the appropriate leaves attached to each stem in the form of a horizontal, ordered row. An example using the first 20 readings of 'Pulse After Exercise' (variable Pulse2) from Minitab's *Pulse* dataset (see Figure 1) will illustrate these notions. (Notice that Minitab has decided on a stem width

1	5	8
1	6	
1	6	
4	7	022
10	7	566668
10	8	002444
4	8	8
3	9	4
2	9	6
1	10	
1	10	
1	11	
1	11	8

Figure 1 Stem and leaf plot of 'Pulse After Exercise' data (first 20 readings on variable Pulse2). Column 1 shows the up-and-down cumulated COUNTS; column 2 contains the STEM values, and column 3 the LEAVES

of 5 units, so there are two stems valued 6, two 7s, two 8s, and so on, which are used to represent the numbers 60 to 64 and 65 to 69, 70 to 74 and 75 to 79, 80 to 84 and 85 to 89 respectively).

Reading from the left, the display in Figure 1 consists of three columns, the first containing the cumulative count up to the middle from the lowest value (58), and down to the middle from the highest value (118); the second lists the ordered stems, and the final column the ordered leaves. Thus the very first row of the plot (1 5|8) represents the number 58, which is the smallest number in the sample and hence has a cumulated value of 1 in the first column. The sample contains nothing in the 60s, so there are no leaves for either stem, and the cumulative total remains 1 for these stems. However, there are three readings in the low 70s (70, 72 and 72), hence the fourth row reads as 7|022. This row also has a cumulative total upward from the lowest reading of $(1 + 3) = 4$, thus giving a complete row of (4 7|022). The next row has six numbers (75, four 76s, and a 78), with the row reading 7|566668, that is, a stem of 7 and six leaves. The cumulative total *upward* from the smallest number of 58 in this row is therefore $(1 + 3 + 6) = 10$, which is exactly half the sample size of 20 readings, and hence is the stopping point for this particular count. Finally, the next row reads 10 8|002444, which is the compact version of the numbers 80, 80, 82, 84, 84 and 84, while the 10 in the first position in the row represents the cumulative count from the largest pulse of 118 *down* to the middle of the sample.

Although the major role of the stem and leaf plot is to display the distributional properties of a single sample, it also easy to extract robust measures from it by simply counting up from the lowest value or down from the highest using the cumulative figures in column 1. Thus, in our sample of 20, the lower quartile is the interpolated value lying between the fifth and sixth numbers from the bottom, that is, 75.25, while the upper quartile is a similarly interpolated value of 84, with the median lying midway between the two middle numbers of 78 and 80, that is, 79. Overall, the data looks single peaked and somewhat biased toward low pulse values, a tendency that shows up even more strongly with the full 92 observations.

Stem and leaf plots are, however, less useful for comparing several samples because of their complexity, with the useful upper limit being two samples

2 Stem and Leaf Plot

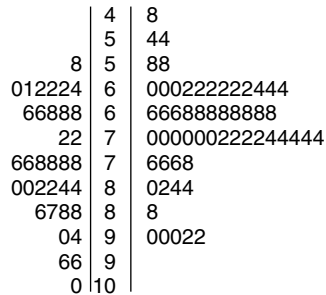


Figure 2 Back-to-back stem and leaf plots of the ‘Pulse after Exercise’ data: males on the left, females on the right

for which the *back-to-back stem and leaf plot* was invented. This consists of the stem and leaf plot for one sample staying orientated as above, with the second one rotated through 180° in the plane and then butted up against the first. The example in Figure 2 again draws again on the Minitab *Pulse* dataset, this time using all 92 ‘pulse before exercise’ readings (variable Pulse1); the left-hand readings are for

males, the right for females – note that the counts columns have been omitted to reduce the chart clutter. Although the female data seems little more peaked than the male, the spreads and medians do not appear to differ much.

Further information on stem and leaf plots can be found in [1–3].

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.
- [2] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1983). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.
- [3] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

SANDY LOVIE

Stephenson, William

JAMES GOOD

Volume 4, pp. 1898–1900

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Stephenson, William

Born: May 14 1902, Chopwell, Co. Durham, UK.

Died: June 14 1989, Columbia, MO.

Initially trained in Physics at the University of Durham (BSc, 1923; MSc, 1925; PhD, 1927), Stephenson also completed his Diploma in the Theory and Practice of Teaching there which brought him into contact with **Godfrey Thomson**, one of the pioneers of **factor analysis**. Inspired by his encounter with Thomson to explore the application of factor analysis in the study of mind, Stephenson moved in 1926 to University College, London to study psychophysics with **Charles Spearman**, and work as Research Assistant to Spearman and to **Cyril Burt**. He also became interested in psychoanalysis and in 1935 began analysis with Melanie Klein.

In 1936, Stephenson accepted an appointment as Assistant Director of the newly established Oxford Institute of Experimental Psychology. War service interrupted his career and he served as a Consultant to the British Armed Forces, rising to the rank of Brigadier General. He became Reader in Experimental Psychology in 1942 and successor to William Brown as Director of the Institute of Experimental Psychology in 1945. Failing to secure the first Oxford Chair in Psychology (filled by George Humphrey in 1947), Stephenson emigrated to the United States, first to the University of Chicago as a Visiting Professor of Psychology and then in 1955, when a permanent academic post at Chicago was not forthcoming, to Greenwich, Connecticut as Research Director of a leading market research firm, Nowland & Co. In 1958, he became Distinguished Research Professor of Advertising at the School of Journalism, University of Missouri, Columbia, where he remained until his retirement in 1972.

Spearman once described Stephenson as the foremost 'creative statistician' of the psychologists of his generation, a view that was endorsed by Egon Brunswik when he wrote that 'Q-technique [was] the most important development in psychological statistics since Spearman's introduction of factor analysis' [1]. Stephenson was a central figure in the development of, and debates about psychometrics and factor analysis. Although the idea of correlating persons rather than traits or test items had been

proposed as early as 1915 by **Cyril Burt** [2], it was Stephenson who saw the potential of this procedure for psychological analysis. He first put forward his ideas about Q-methodology in a letter to *Nature* in 1935 [4]. A detailed exposition together with a challenge to psychology 'to put its house in scientific order' did not appear until 1953 [5]. In his writings, Stephenson employs a distinction (first put forward by Godfrey Thomson) between correlating persons (Q-methodology) and the traditional use of factor analysis in psychometrics to correlate traits or test items (R-methodology). Q-methodology applies to a population of tests or traits, with persons as variables; R-methodology to a population of persons, with tests or traits as variables. Q-methodology provides a technique for assessing a person's subjectivity or point of view, especially concerning matters of value and preference. Stephenson rejected the 'reciprocity principle' promoted by Burt and Cattell that Q and R are simply reciprocal solutions (by rows or by columns) of a single data matrix of scores from objective tests. Q-methodology was seen by Stephenson as involving two separate data matrices, one containing objective scores (R), the other containing data of a subjective kind reflecting perceived representativeness or significance (Q). This was a matter about which Burt and Stephenson eventually agreed to differ in a jointly published paper [3] (*see R & Q Analysis*).

When used with multiple participants, Q-methodology identifies the views that participants have in common and is therefore a technique for the assessment of shared meaning. Stephenson also developed Q for use with a single participant with multiple conditions of instruction [5, 7]. The single case use of Q affords a means of exploring the structure and content of the views that individuals hold about their worlds (for example, the interconnections between a person's view of self, of ideal self, and of self as they imagine they are seen by a variety of significant others).

In developing his ideas about Q-methodology, Stephenson eschewed Cartesian mind-body dualism, thus reflecting an important influence on his thinking of the transactionalism of John Dewey and Arthur Bentley, and the interbehaviorism of Jacob Kantor. His functional and processual theory of self was heavily influenced by Kurt Koffka and Erving Goffman [9]. Building on the work of Johan Huizinga and Wilbur Schramm, Stephenson also developed a theory of communication that focused on the social

and pleasurable aspects of communication as opposed to the exchange of information [6, 8]. Following his retirement, Stephenson devoted much of his time to writing a series of papers on what had been one of his earliest preoccupations, the exploration of the links between quantum theory and subjectivity [10]. Many of Stephenson's central notions are succinctly brought together in a posthumously published monograph [11].

References

- [1] Brunswik, E. (1952). *Letter to Alexander Morin*, Associate Editor, University of Chicago Press, May 20, Western Historical Manuscript Collection, Ellis Library, University of Missouri-Columbia, Columbia, 1.
- [2] Burt, C. (1915). General and specific factors underlying the emotions, *British Association Annual Report* **84**, 694–696.
- [3] Burt, C. & Stephenson, W. (1939). Alternative views on correlations between persons, *Psychometrika* **4**(4), 269–281.
- [4] Stephenson, W. (1935). Technique of factor analysis, *Nature* **136**, 297.
- [5] Stephenson, W. (1953). *The Study of Behavior: Q-Technique and its Methodology*, University of Chicago Press, Chicago.
- [6] Stephenson, W. (1967). *The Play Theory of Mass Communication*, University of Chicago Press, Chicago.
- [7] Stephenson, W. (1974). Methodology of single case studies, *Journal of Operational Psychiatry* **5**(2), 3–16.
- [8] Stephenson, W. (1978). Concourse theory of communication, *Communication* **3**, 21–40.
- [9] Stephenson, W. (1979). The communicability and operancy of the self, *Operant Subjectivity* **3**(1), 2–14.
- [10] Stephenson, W. (1988). Quantum theory of subjectivity, *Integrative Psychiatry* **6**, 180–195.
- [11] Stephenson, W. (1994). *Quantum Theory of Advertising*, School of Journalism, University of Missouri-Columbia, Columbia.

JAMES GOOD

Stevens, S S

ROBERT B. FAUX

Volume 4, pp. 1900–1902

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Stevens, S S

Born: November 4, 1906, in Ogden Utah, USA.

Died: January 18, 1973, in Vail, USA.

S.S. Stevens was a twentieth-century American experimental psychologist who conducted foundational research on sensation and perception, principally in psychoacoustics. However, it is the critical role Stevens played in the development of **measurement** and operationism for which he is probably best known by psychologists and social scientists in general [1, 2].

Upon completion of high school in 1924, Stevens' went to Belgium and France as a Mormon missionary. In 1927, his missionary work completed, Stevens entered the University of Utah where he took advanced courses in the humanities and social sciences and made a failed attempt at algebra [6]. After two years, he transferred to Stanford University where he took a wide variety courses without ever declaring a major, threatening his graduation from that institution. He did graduate in 1931 and was accepted into Harvard Medical School. A \$50.00 fee and the requirement to take organic chemistry during the summer persuaded Stevens that medical school was not for him. He enrolled in Harvard's School of Education, reasoning that the tuition of \$300.00 was the cheapest way to take advantage of the school's resources. He found only one course in education that looked interesting: an advanced statistic course taught by T.L. Kelley. At this time Stevens also took a course in physiology with W.J. Crozier. While exploring Crozier's laboratory one day Stevens encountered B.F. Skinner plotting data. Skinner explained that he was plotting eating curves for rats and that they could be described by power functions. Stevens admitted that he did know what power functions were to which Skinner replied that the best way for him to overcome such inferiority in mathematics was to learn it; advice Stevens was to take seriously [6, 9].

Among the most critical and far-reaching experiences for Stevens at this time was his intellectual relationship with E. G. Boring, the sole professor of psychology at Harvard at this time, as psychology was still affiliated with the philosophy department. It was while he was conducting an experiment for

Boring on color perception Stevens recounts that his scientific career began: He discovered a law-like relationship between color combinations, distance, and perception. This resulted in his first published experiment. Eventually transferring from education to psychology, Stevens defended his dissertation on tonal attributes in May of 1933. Stevens was to remain at Harvard, first as an instructor of psychology, then as Professor of Psychophysics, a title conferred in 1962, until his death [9].

Stevens also audited courses in mathematics as well as in physics, becoming a Research Fellow in Physics for some time. In 1936 he settled on psychology as his profession [6, 9]. After this he spent a year with Hallowell Davis studying electrophysiology at Harvard Medical School. This proved to be another fruitful intellectual relationship, culminating in the book *Hearing* in 1938 [7]. This book was for many years considered a foundational text in psychoacoustics. In addition to this work, Stevens did research on the localization of auditory function. In 1940 he established the Psycho-Acoustic Laboratory at Harvard. Stevens gathered a group of distinguished colleagues to help in this work, among them were G. A. Miller and G. von Békésy. It was during his tenure in Stevens' laboratory that von Békésy won the Nobel Prize for his work on the ear [6, 9]. Interestingly, Stevens was quite uncomfortable in his role as classroom instructor. The only teaching for which Stevens felt affinity was the give and take of laboratory work with apprentices and editorial work with authors. This is reflected in Stevens' ability to attract a group of gifted collaborators.

For many, Stevens' most important achievement was the discovery of the Psychophysical Power Law or Stevens' Law [4]. This law describes the link between the strength of a stimulus, for example, a tone, and the corresponding sensory sensation, in this case loudness. In 1953 Stevens began research in psychophysics that would upend a longstanding psychophysical law that stated that as the strength of a stimulus grew geometrically (as a constant ratio), sensation of that stimulus grew arithmetically – a logarithmic function. This is known as the Weber-Fechner Law. This view of things held sway for many years despite a lack of empirical support. Finding accurate ways to measure experience was a fundamental question in psychophysics since its inception in the nineteenth century. Stevens set about finding a more accurate measure of this relationship. He was

able to demonstrate that the relationship between the subjective intensity of a stimulus and its physical intensity itself was not a logarithmic function but was, rather, a power function. He did this by having observers assign numbers to their subjective impressions of a stimulus, in this case a tone [8, 9]. The results of Stevens' research were better explained as power functions than as logarithmic functions. That is, equal ratios in the stimulus corresponded to equal ratios in one's subjective impressions of the stimulus. Stevens was also able to demonstrate that such a relationship held for different modalities [4, 9].

Stevens' interest in the nature of measurement and operationism stemmed from his research in psychoacoustics. As noted previously, throughout the history of psychophysics more accurate measures of experience were continually being sought. It was Stevens' belief that the nature of measurement needed to be clarified if quantification of sensory attributes was to take place [6]. Throughout the 1930s Stevens set about trying to elucidate the nature of measurement. Harvard University at this time was a hotbed of intellectual debate concerning the nature of science. Stevens was exposed to the ideas of philosopher A.N. Whitehead, physicist P. Bridgman, and mathematician G.D. Birkhoff. Near the end of the decade there was an influx of European philosophers, many from the Vienna Circle, among whom was Rudolf Carnap. At Carnap's suggestion, Stevens organized a club to discuss the Science of Science. Invitations were sent and in late October 1940 the inaugural meeting took place with P.W. Bridgman discussing operationism. Bridgman argued that measurement should be based upon the operations that created it rather than on what was being measured. Throughout the latter half of the 1930s Stevens published several papers on the concept of operationism [6]. For Stevens, as for many psychologists, operationism was a way to reintroduce rigor in the formulation of concepts [6]. Measurement and operationism proved to be quite alluring to psychologists, as it was believed that if psychology was to be taken seriously as a positivistic science it required rigorous measurement procedures akin to those of physics [2, 9].

Stevens recounts that his attempt to explicate the nature of measurement and describe various scales at a Congress for the Unity of Science Meeting in 1939 was unsuccessful. Initially, Stevens identified three scales: ordinal, intensive, and extensive. From feedback given at the Congress Stevens set

about forming various scales and describing the operations he used to form them [6]. Finally, in a 1946 paper Stevens presented his taxonomy of measurement scales: nominal, ordinal, interval, and ratio (*see Scales of Measurement*). Following Bridgman, Stevens defined each of his scales based upon the operations used to create it, leaving the form of the scale invariant, rather than upon what was being measured. Stevens further argued that these operations maintained a hierarchical relationship with each other [3]. Nominal scales have no order, they are used simply to distinguish among entities. For instance: 1 = Tall, 2 = Short, 3 = Large, 4 = Small. Neither arithmetical nor logical operations can be performed on nominal data. Ordinal scales are comprised of rank orderings of events. For example, students relative rankings in a classroom: Student A has achieved a rank of 100; Student B, a rank of 97; Student C, a rank of 83; and so on. Because the intervals between ranks are variable arithmetical operations cannot be carried out; however, logical operations such as 'more than' and 'less than' are possible. Interval scales maintain order and have equal intervals, in other words they have constant units of measurement, as in scales of temperature. The arithmetical operations of addition and subtraction are permitted, as are logical operations. Ratio scales also maintain constant units of measurement and have a true zero point, thus allowing values to be expressed as ratios.

Stevens argued that each scale type was characterized by an allowable transformation that would leave the scale type invariant. For example, nominal scales allow one-to-one substitutions of numbers as they only identify some variable [9]. Stevens used the property of invariance to relate measurement scales to certain allowable statistical procedures. For instance, the correlation coefficient r will retain its value under a linear transformation [5]. This view of measurement scales could be used as a guide to choosing appropriate statistical techniques was challenged from the time of its appearance and continues to be [2]. However, despite its many challenges, Stevens' views on measurement were quickly and widely disseminated to the psychological community. This occurred most notably through Stevens' membership in the Psychological Round Table. This group of experimental psychologists met yearly from 1936 until 1946 to discuss the latest advancements in the discipline [1].

To this day Stevens' views on measurement maintain their influence in psychology [2, 9]. Most students of psychology in their statistics classes become familiar with Stevens' four measurement scales, often without any reference to Stevens himself. Almost from its inception as a distinct academic discipline, the nature of psychology has been questioned. Is it, indeed, a science? Can it be a science? The attraction of Stevens' scales of measurement was that they offered a degree of rigor that lent legitimacy to psychology's claim to be a science. In many ways Stevens continued the tradition begun by **Gustav Fechner** and Wilhelm Wundt, among others in the nineteenth century, of applying the rigors of mathematics and science to psychological questions such as the relationship between a stimulus and the concomitant subjective experience of that stimulus [9].

References

- [1] Benjamin, L.T. (1977). The psychology roundtable: revolution of 1936, *American Psychologist* **32**, 542–549.
- [2] Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept*, Cambridge University Press, Cambridge.
- [3] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.
- [4] Stevens, S.S. (1957). On the psychophysical law, *Psychological Review* **64**, 153–181.
- [5] Stevens, S.S. (1968). Measurement, statistics, and the schemapiric view, *Science* **161**, 849–856.
- [6] Stevens, S.S. (1974). S.S. Stevens, in *A History of Psychology in Autobiography*, Vol. VI, G. Lindzey, ed., Prentice Hall, Englewood Cliffs, pp. 393–420.
- [7] Stevens, S.S. & Davis, H. (1938). *Hearing: Its Psychology and Physiology*, Wiley, New York.
- [8] Still, A. (1997). Stevens, Stanley Smith, in *Biographical Dictionary of Psychology*, N. Sheehy, A.J. Chapman & W.A. Conroy, eds, Routledge, London, pp. 540–543.
- [9] Teghtsoonian, R. (2001). Stevens, Stanley, Smith (1906–73), *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier Science, New York.

ROBERT B. FAUX

Stratification

VANCE W. BERGER AND JIALU ZHANG

Volume 4, pp. 1902–1905

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Stratification

The word ‘stratify’ means ‘to make layers’ in Latin. When a population is composed of several subpopulations and the subpopulations vary considerably on the factors in which we are interested, stratification can reduce variability of the estimates by dividing the population into relatively homogeneous subgroups and treating each subgroup separately. This concept can be applied for stratified sampling (sampling from within each stratum) and stratified random allocation of subjects that need not constitute a random sample, stratified or not (*see Survey Sampling Procedures*). The purpose of stratified sampling is to obtain more precise estimates of variables of interest. Stratified sampling divides the whole population into more homogeneous subpopulations (also called *strata*) in such a way that every observation in the population belongs to one and only one stratum (a partition). From each stratum, a sample is selected independent of other strata. The simplest method is to use a simple random sample in each stratum.

The population estimate is then computed by pooling the information from each stratum together. For example, one may conduct a survey to find out the average home price in the United States. The simplest method may be a simple random sample, by which each home in the United States has an equal chance to be selected into the sample. Then the estimated average home price in the United States would be the average home price in the sample. However, home prices around metropolitan areas, such as New York and Washington, DC, tend to be much higher than those in rural areas. In fact, the variable of interest, rural versus urban, and the influence it exerts on home prices, is not dichotomous, but rather more continuous. One can exploit this knowledge by defining three strata based on the size of the city (rural, intermediate, metropolitan).

The average home price would then be computed in each stratum, and the overall estimated home price in United States would be obtained by pooling the three average home prices with some weight function. This would result in the same estimate as that obtained from simple random sampling if the weight function were derived from the proportions of each type in the sample. That is, letting $X1/n1$, $X2/n2$, and $X3/n3$ denote the average (in the sample) home prices in rural, intermediate, and metropolitan

areas, with $X = X1 + X2 + X3$ and $n = n1 + n2 + n3$, one finds that $X/n = (X1/n1)(n1/n) + (X2/n2)(n2/n) + (X3/n3)(n3/n)$. However, the key to stratified sampling is that the weight functions need not be the observed proportions $n1/n$, $n2/n$, and $n3/n$. In fact, one can use weights derived from external knowledge (such as the number of homes in each type of area), and then the sampling can constrain $n1$, $n2$, and $n3$ so that they are all equal to each other (an equal number of homes would be sampled from each type of area).

This approach, with making use of external weight functions (to reflect the proportion of each stratum in the population, instead of in the sample), results in an estimated average home price obtained with smaller variance compared to that obtained from simple random sampling. This is because each estimate is based on a more homogeneous sample than would otherwise be the case. The drawback of this stratified sampling design is that it adds complexity to the survey, and sometimes the improvement in the estimation, which may not be substantial in some cases, may not be worth the extra complexity that stratified sampling brings to the design [3].

Stratification is also used in random allocation (as opposed to the random sampling just discussed). In the context of a comparative trial (*see Clinical Trials and Intervention Studies*), the best comparative inference takes place when all key covariates are balanced across the treatment groups. When a false positive finding, or a Type I error occurs, this is because of differences across treatment groups in key predictors of the outcome. Such covariate imbalances can also cause Type II errors, or the masking of a true treatment effect.

Stratification according to prognostic factors, such as gender, age, smoking status, or number of children, can guarantee balance with respect to these factors. A separate randomization list, with balance built in, is prepared for each stratum. A consequence of this randomization within strata is that the numbers of patients receiving each treatment are similar not only in an overall sense, but also within each stratum. Generally, the randomization lists across the various strata are not only separate, but also independent. A notable exception is a study of etanercept for children with juvenile rheumatoid arthritis [4], which used blocks within each of two strata, and the corresponding blocks in the two strata were mirror images of each other [1].

The problem with this stratification method is that the number of strata increases quickly as the number of prognostic factors or as the level of a prognostic factor increases. Some strata may not have enough patients to be randomized. Other restricted randomization procedures (also called *stratified randomization procedures*) include, for example, the randomized block design, the maximal procedure [2], and minimization [6].

Within each block, one can consider a variety of techniques for the randomization. The maximal procedure [2] has certain optimality properties in terms of balancing chronological bias and selection bias, but generally, the random allocation rule [5] is used within each block within each stratum. This means that randomization actually occurs within blocks within strata and is conducted without any restrictions besides the blocking itself and the balance it entails at the end of each block. So, for example, if the only stratification factor were gender, then there would be two strata, male and female. This means that there would be one (balanced for treatment groups) randomization for males and another (balanced for treatment groups) randomization for females. If blocking were used within strata, with a fixed block size of four, then the only restriction within the strata would be that each block of four males and each block of four females would have two subjects allocated to each treatment group.

The first four males would constitute a block, as would the first four females, the next four females, the next four males, and so on. There is a limit to the number of strata that can be used to advantage [7]. Each binary stratification factor multiplies the existing number of strata by two; stratification factors with more than two levels multiply the existing number of strata by more than two. For example, gender is binary, and leads to two strata. Add in smoking history (never, ever, current) and there are now $2 \times 3 = 6$ strata. Add in age bracket (classified as 20–30, 30–40, 40–50, 50–60, 60–70) and this six gets multiplied by five, which is the number of age brackets. There are now 30 strata.

If a study of a behavioral intervention has only 100 subjects, then on average there would be slightly more than three subjects per stratum. This situation would defeat the purpose of stratification in that the treatment comparisons within the strata could not be considered robust. Minimization [6] can handle more stratification factors than can a

stratified design. The idea behind minimization is that an imbalance function is minimized to determine the allocation, or at least the allocation that is more likely. That is, a subject to be enrolled is sequentially allocated (provisionally) to each treatment group, and for each such provisional allocation, the resulting imbalance is computed. The treatment group that results in the smallest imbalance will be selected as the favored one. In a deterministic version, the subject would be allocated to this treatment group. In a stochastic version, this treatment group would have the largest allocation probability.

As a simple example, suppose that the trial is underway and 46 subjects have already been enrolled, 23 to each group. Suppose further that the two strongest predictors of outcome are gender and age (over 40 and under 40). Finally, suppose that currently Treatment Group A has four males over 40, five females over 40, seven males under 40, and seven females under 40, while Treatment Group B has eight males over 40, four females over 40, six males under 40, and five females under 40. The 47th subject to be enrolled is a female under 40. Provisionally place this subject in Treatment Group A and compute the marginal female imbalance to be $(5 + 7 + 1 - 4 - 5) = 4$, the marginal age imbalance to be $(7 + 7 + 1 - 6 - 5) = 4$, and the joint female age imbalance to be $(7 + 1 - 5) = 3$.

Now provisionally place this subject in Treatment Group B and compute the marginal female imbalance to be $(5 + 7 - 4 - 5 - 1) = 2$, the marginal age (under 40) imbalance to be $(7 + 7 - 6 - 5 - 1) = 2$, and the joint female age imbalance to be $(7 - 5 - 1) = 1$. Using joint balancing, Treatment Group B would be preferred, as 1 is less than three. Again, the actual allocation may be deterministic, as in simply assign the subject to the group that leads to better balance, B in this case, or it may be stochastic, as in make this assignment with high probability. Using marginal balancing, this subject would still either be allocated to Treatment Group B or have a high probability of being so allocated, as 2 is less than 4. Either way, then, Treatment Group B is favored for this subject. One problem with minimization is that it leads to predictable allocations, and these predictions can lead to strategic subject selection to create an imbalance in a covariate that is not being considered by the imbalance function.

References

- [1] Berger, V. *FDA Product Approval Information – Licensing Action: Statistical Review*, <http://www.fda.gov/cber/products/etanimm052799.htm> 1999, accessed 3/7/02.
- [2] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.
- [3] Lohr, S. (1999). *Sampling: Design and Analysis*, Duxbury Press.
- [4] Lovell, D.J., Giannini, E.H., Reiff, A., Cawkwell, G.D., Silverman, E.D., Nocton, J.J., Stein, L.D., Gedalia, A., Ilowite, N.T., Wallace, C.A., Whitmore, J. & Finck, B.K. (2000). Etanercept in children with polyarticular juvenile rheumatoid arthritis, *The New England Journal of Medicine* **342**(11), 763–769.
- [5] Rosenberger, W.F. & Rukhin, A.L. (2003). Bias properties and nonparametric inference for truncated binomial randomization, *Nonparametric Statistics* **15**, 4–5, 455–465.
- [6] Taves, D.R. (1974). Minimization: a new method of assigning patients to treatment and control groups, *Clinical Pharmacology Therapeutics* **15**, 443–453.
- [7] Therneau, T.M. (1993). How many stratification factors are “too many” to use in a randomization plan? *Controlled Clinical Trials* **14**(2), 98–108.

VANCE W. BERGER AND JIALU ZHANG

Structural Equation Modeling: Categorical Variables

ANDERS SKRONDAL AND SOPHIA RABE-HESKETH

Volume 4, pp. 1905–1910

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Categorical Variables

Introduction

Structural equation models (SEMs) comprise two components, a measurement model and a structural model. The measurement model relates observed responses or ‘indicators’ to **latent variables** and sometimes to observed covariates. The structural model then specifies relations among latent variables and regressions of latent variables on observed variables. When the indicators are categorical, we need to modify the conventional measurement model for continuous indicators. However, the structural model can remain essentially the same as in the continuous case.

We first describe a class of structural equation models also accommodating dichotomous and ordinal responses [5]. Here, a conventional measurement model is specified for multivariate normal ‘latent responses’ or ‘underlying variables’. The latent responses are then linked to observed categorical responses via threshold models yielding probit measurement models.

We then extend the model to generalized latent variable models (e.g., [1], [13]) where, conditional on the latent variables, the measurement models are **generalized linear models** which can be used to model a much wider range of response types.

Next, we briefly discuss different approaches to estimation of the models since estimation is considerably more complex for these models than for conventional structural equation models. Finally, we illustrate the application of structural equation models for categorical data in a simple example.

SEMs for Latent Responses

Structural Model

The structural model can take the same form regardless of response type. Letting j index units or subjects, Muthén [5] specifies the structural model for latent variables η_j as

$$\eta_j = \alpha + \mathbf{B}\eta_j + \mathbf{\Gamma}\mathbf{x}_{1j} + \zeta_j. \quad (1)$$

Here, α is an intercept vector, $\mathbf{B}$ a matrix of structural parameters governing the relations among the latent variables, $\mathbf{\Gamma}$ a regression parameter matrix for regressions of latent variables on observed explanatory variables $\mathbf{x}_{1j}$ and ζ_j a vector of disturbances (typically multivariate normal with zero mean). Note that this model is defined conditional on the observed explanatory variables $\mathbf{x}_{1j}$. Unlike conventional SEMs where all observed variables are treated as responses, we need not make any distributional assumptions regarding $\mathbf{x}_{1j}$.

In the example considered later, there is a single latent variable η_j representing mathematical reasoning or ‘ability’. This latent variable is regressed on observed covariates (gender, race and their interaction),

$$\eta_j = \alpha + \boldsymbol{\gamma}\mathbf{x}_{1j} + \zeta_j, \quad \zeta_j \sim \mathbf{N}(0, \psi), \quad (2)$$

where $\boldsymbol{\gamma}$ is a row-vector of regression parameters.

Measurement Model

The distinguishing feature of the measurement model is that it is specified for *latent* continuous responses $\mathbf{y}_j^*$ in contrast to observed continuous responses $\mathbf{y}_j$ as in conventional SEMs,

$$\mathbf{y}_j^* = \mathbf{v} + \mathbf{\Lambda}\eta_j + \mathbf{K}\mathbf{x}_{2j} + \boldsymbol{\epsilon}_j. \quad (3)$$

Here $\mathbf{v}$ is a vector of intercepts, $\mathbf{\Lambda}$ a factor loading matrix and $\boldsymbol{\epsilon}_j$ a vector of unique factors or ‘measurement errors’. Muthén and Muthén [7] extend the measurement model in Muthén [5] by including the term $\mathbf{K}\mathbf{x}_{2j}$ where $\mathbf{K}$ is a regression parameter matrix for the regression of $\mathbf{y}_j^*$ on observed explanatory variables $\mathbf{x}_{2j}$. As in the structural model, we condition on $\mathbf{x}_{2j}$.

When $\boldsymbol{\epsilon}_j$ is assumed to be multivariate normal (*see Catalogue of Probability Density Functions*), this model, combined with the threshold model described below, is a *probit* model (*see Probits*). The variances of the latent responses are not separately identified and some constraints are therefore imposed. Muthén sets the total variance of the latent responses (given the covariates) to 1.

Threshold Model

Each observed categorical response y_{ij} is related to a latent continuous response y_{ij}^* via a threshold model.

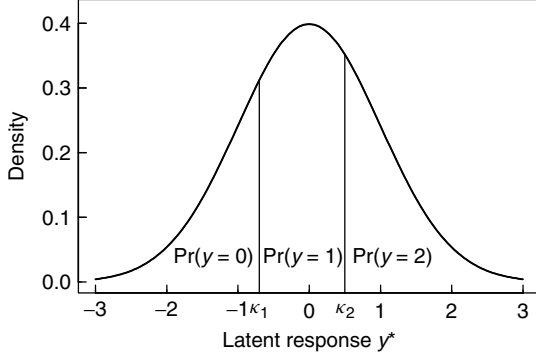


Figure 1 Threshold model for ordinal responses with three categories (This figure has been reproduced from [13] with permission from Chapman and Hall/CRC.)

For ordinal observed responses it is assumed that

$$y_{ij} = \begin{cases} 0 & \text{if } -\infty < y_{ij}^* \leq \kappa_{1i} \\ 1 & \text{if } \kappa_{1i} < y_{ij}^* \leq \kappa_{2i} \\ \vdots & \vdots \\ S & \text{if } \kappa_{Si} < y_{ij}^* \leq \infty. \end{cases} \quad (4)$$

This is illustrated for three categories ($S = 2$) in Figure 1 for normally distributed ϵ_i , where the areas under the curve are the probabilities of the observed responses.

Either the constants $\mathbf{v}$ or the thresholds κ_{1i} are typically set to 0 for identification. Dichotomous observed responses simply arise as the special case where $S = 1$.

Generalized Latent Variable Models

In generalized latent variable models, the measurement model is a generalized linear model of the form

$$\mathbf{g}(\mu_j) = \mathbf{v} + \mathbf{A}\eta_j + \mathbf{K}\mathbf{x}_{2j}, \quad (5)$$

where $\mathbf{g}(\cdot)$ is a vector of link functions which may be of different kinds handling mixed response types (for instance, both continuous and dichotomous observed responses or ‘indicators’). μ_j is a vector of conditional means of the responses given η_j and $\mathbf{x}_{2j}$ and the other quantities are defined as in (3). The conditional models for the observed responses given μ_j are then distributions from the exponential family (see **Generalized Linear Models (GLM)**). Note that there are no explicit unique factors in the model

because the variability of the responses for given values of η_j and $\mathbf{x}_{2j}$ is accommodated by the conditional response distributions. Also note that the responses are implicitly specified as conditionally independent given the latent variables η_j (see **Conditional Independence**).

In the example, we will consider a single latent variable measured by four dichotomous indicators or ‘items’ y_{ij} , $i = 1, \dots, 4$, and use models of the form

$$\text{logit}(\mu_{ij}) \equiv \ln \left(\frac{\Pr(\mu_{ij})}{1 - \Pr(\mu_{ij})} \right) = v_i + \lambda_i \eta_j, \\ v_1 = 0, \lambda_1 = 1. \quad (6)$$

These models are known as two-parameter logistic item response models because two parameters (v_i and λ_i) are used for each item i and the logit link is used (see **Item Response Theory (IRT) Models for Dichotomous Data**). Conditional on the latent variable, the responses are Bernoulli distributed (see **Catalogue of Probability Density Functions**) with expectations $\mu_{ij} = \Pr(y_{ij} = 1 | \eta_j)$. Note that we have set $v_1 = 0$ and $\lambda_1 = 1$ for identification because the mean and variance of η_j are free parameters in (2). Using a probit link in the above model instead of the more commonly used logit would yield a model accommodated by the Muthén framework discussed in the previous section.

Models for counts can be specified using a log link and Poisson distribution (see **Catalogue of Probability Density Functions**). Importantly, many other response types can be handled including ordered and unordered categorical responses, rankings, durations, and mixed responses; see for example, [1, 2, 4, 9, 11, 12 and 13] for theory and applications. A recent book on generalized latent variable modeling [13] extends the models described here to ‘generalized linear latent and mixed models’ (GLLAMMs) [9] which can handle multilevel settings and discrete latent variables.

Estimation and Software

In contrast to the case of multinormally distributed continuous responses, **maximum likelihood estimation** cannot be based on sufficient statistics such as the empirical covariance matrix (and possibly mean vector) of the observed responses. Instead, the likelihood must be obtained by somehow ‘integrating out’ the latent variables η_j . Approaches which

work well but are computationally demanding include adaptive Gaussian quadrature [10] implemented in `gllamm` [8] and Markov Chain Monte Carlo methods (typically with noninformative priors) implemented in BUGS [14] (*see Markov Chain Monte Carlo and Bayesian Statistics*).

For the special case of models with multinormal latent responses (principally probit models), Muthén suggested a computationally efficient limited information estimation approach [6] implemented in `Mplus` [7]. For instance, consider a structural equation model with dichotomous responses and no observed explanatory variables. Estimation then proceeds by first estimating ‘tetrachoric correlations’ (pairwise correlations between the latent responses). Secondly, the asymptotic covariance matrix of the tetrachoric correlations is estimated. Finally, the parameters of the SEM are estimated using weighted least squares (*see Least Squares Estimation*), fitting model-implied to estimated **tetrachoric correlations**. Here, the inverse of the asymptotic covariance matrix of the tetrachoric correlations serves as weight matrix.

Skrondal and Rabe-Hesketh [13] provide an extensive overview of estimation methods for SEMs with noncontinuous responses and related models.

Example

Data

We will analyze data from the Profile of American Youth (US Department of Defense [15]), a survey of the aptitudes of a national probability sample of Americans aged 16 through 23. The responses (1: correct, 0: incorrect) for four items of the arithmetic reasoning test of the Armed Services Vocational Aptitude Battery (Form 8A) are shown in Table 1 for samples of white males and females and black males and females. These data were previously analyzed by Mislevy [3].

Model Specification

The most commonly used measurement model for ability is the two-parameter logistic model in (6) and (2) without covariates.

Item characteristic curves, plots of the probability of a correct response as a function of ability, are

given by

$$\Pr(y_{ij} = 1 | \eta_j) = \frac{\exp(v_i + \lambda_i \eta_j)}{1 + \exp(v_i + \lambda_i \eta_j)}. \quad (7)$$

and shown for this model (using estimates under $\mathcal{M}_1$ in Table 2) in Figure 2.

We then specify a structural model for ability η_j . Considering the covariates

- [Female] F_j , a dummy variable for subject j being female
- [Black] B_j , a dummy variable for subject j being black

we allow the mean abilities to differ between the four groups,

$$\eta_j = \alpha + \gamma_1 F_j + \gamma_2 B_j + \gamma_3 F_j B_j + \zeta_j. \quad (8)$$

This is a MIMIC model where the covariates affect the response via a latent variable only.

A path diagram of the structural equation model is shown in Figure 3. Here, observed variables are represented by rectangles whereas the latent variable is represented by a circle. Arrows represent regressions (not necessary linear) and short arrows residual variability (not necessarily an additive error term). All variables vary between subjects j and therefore the j subscripts are not shown.

We can also investigate if there are direct effects of the covariates on the responses, in addition to the indirect effects via the latent variable. This could be interpreted as ‘item bias’ or ‘differential item

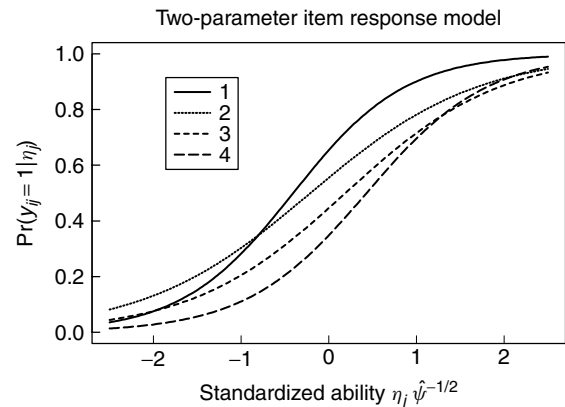


Figure 2 Item characteristic curves for items 1 to 4 (This figure has been reproduced from [13] with permission from Chapman and Hall/CRC.)

4 Structural Equation Modeling: Categorical Variables

Table 1 Arithmetic reasoning data

Item Response				White Males	White Females	Black Males	Black Females
1	2	3	4				
0	0	0	0	23	20	27	29
0	0	0	1	5	8	5	8
0	0	1	0	12	14	15	7
0	0	1	1	2	2	3	3
0	1	0	0	16	20	16	14
0	1	0	1	3	5	5	5
0	1	1	0	6	11	4	6
0	1	1	1	1	7	3	0
1	0	0	0	22	23	15	14
1	0	0	1	6	8	10	10
1	0	1	0	7	9	8	11
1	0	1	1	19	6	1	2
1	1	0	0	21	18	7	19
1	1	0	1	11	15	9	5
1	1	1	0	23	20	10	8
1	1	1	1	86	42	2	4
Total:				263	228	140	145

Source: Mislevy, R.J. Estimation of latent group effects (1985). *Journal of the American Statistical Association* **80**, 993–997 [3].

Table 2 Estimates for ability models

Parameter	$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$	
	Est	(SE)	Est	(SE)	Est	(SE)
Intercepts						
ν_1 [Item1]	0	–	0	–	0	–
ν_2 [Item2]	–0.21	(0.12)	–0.22	(0.12)	–0.13	(0.13)
ν_3 [Item3]	–0.68	(0.14)	–0.73	(0.14)	–0.57	(0.15)
ν_4 [Item4]	–1.22	(0.19)	–1.16	(0.16)	–1.10	(0.18)
ν_5 [Item1] $\times$ [Black] $\times$ [Female]	0	–	0	–	–1.07	(0.69)
Factor loadings						
λ_1 [Item1]	1	–	1	–	1	–
λ_2 [Item2]	0.67	(0.16)	0.69	(0.15)	0.64	(0.17)
λ_3 [Item3]	0.73	(0.18)	0.80	(0.18)	0.65	(0.14)
λ_4 [Item4]	0.93	(0.23)	0.88	(0.18)	0.81	(0.17)
Structural model						
α [Cons]	0.64	(0.12)	1.41	(0.21)	1.46	(0.23)
γ_1 [Female]	0	–	–0.61	(0.20)	–0.67	(0.22)
γ_2 [Black]	0	–	–1.65	(0.31)	–1.80	(0.34)
γ_3 [Black] $\times$ [Female]	0	–	0.66	(0.32)	2.09	(0.86)
ψ	2.47	(0.84)	1.88	(0.59)	2.27	(0.74)
Log-likelihood	–2002.76		–1956.25		–1954.89	
Deviance	204.69		111.68		108.96	
Pearson X^2	190.15		102.69		100.00	

Source: Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton [13].

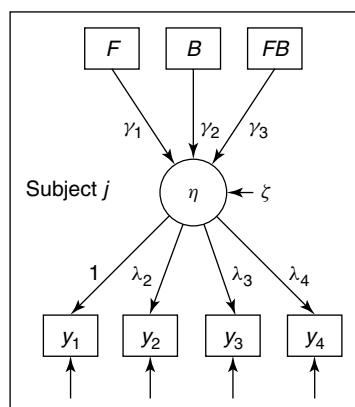


Figure 3 Path diagram of MIMIC model

functioning' (DIF), that is, where the probability of responding correctly to an item differs for instance between black women and others with the same ability (see **Differential Item Functioning**). Such item bias would be a problem since it suggests that candidates cannot be fairly assessed by the test. For instance, if black women perform worse on the first item ($i=1$) we can specify the following model for this item:

$$\text{logit}[\Pr(y_{1j} = 1|\eta_j)] = \beta_1 + \beta_5 F_j B_j + \lambda_1 \eta_j. \quad (9)$$

Results

Table 2 gives maximum likelihood estimates based on 20-point adaptive quadrature estimated using `gllamm` [8]. Estimates for the two-parameter logistic IRT model (without covariates) are given under $\mathcal{M}_1$, for the MIMIC model under $\mathcal{M}_2$ and for the MIMIC model with item bias for black women on the first item under $\mathcal{M}_3$. Deviance and Pearson X^2 statistics are also reported in the table, from which we see that $\mathcal{M}_2$ fits better than $\mathcal{M}_1$. The variance estimate of the disturbance decreases from 2.47 for $\mathcal{M}_1$ to 1.88 for $\mathcal{M}_2$ because some of the variability in ability is 'explained' by the covariates. There is some evidence for a [Female] by [Black] interaction. While being female is associated with lower ability among white people, this is not the case among black people where males and females have similar abilities. Black people have lower mean abilities than both white men and white women. There is little evidence suggesting that item 1 functions differently for black females.

Note that none of the models appear to fit well according to absolute fit criteria (see **Model Fit: Assessment of**). For example, for $\mathcal{M}_2$, the deviance is 111.68 with 53 degrees of freedom, although the Table 1 is perhaps too sparse to rely on the χ^2 distribution.

Conclusion

We have discussed generalized structural equation models for noncontinuous responses. Muthén suggested models for continuous, dichotomous, ordinal and censored (tobit) responses based on multivariate normal latent responses and introduced a limited information estimation approach for his model class.

Recently, considerably more general models have been introduced. These models handle (possibly mixes of) responses such as continuous, dichotomous, ordinal, counts, unordered categorical (polytomous), and rankings. The models can be estimated using maximum likelihood or Bayesian analysis.

References

- [1] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, Arnold Publishing, London.
- [2] Fox, J.P. & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling, *Psychometrika* **66**, 271–288.
- [3] Mislevy, R.J. Estimation of latent group effects, (1985). *Journal of the American Statistical Association* **80**, 993–997.
- [4] Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* **56**, 337–357.
- [5] Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators, *Psychometrika* **49**, 115–132.
- [6] Muthén, B.O. & Satorra, A. (1996). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model, *Psychometrika* **60**, 489–503.
- [7] Muthén, L.K. & Muthén, B.O. (2004). *Mplus User's Guide*, Muthén & Muthén, Los Angeles.
- [8] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004a). GLLAMM Manual. U.C. Berkley Division of Biostatistics Working Paper Series. Working Paper 160. Downloadable from <http://www.bepress.com/ucbbiostat/paper160/>.

6 Structural Equation Modeling: Categorical Variables

- [9] Rabe-Hesketh, S., Skrondal, A., Pickles, A. (2004b). Generalized multilevel structural equation modeling, *Psychometrika* **69**, 167–190.
- [10] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects, *Journal of Econometrics* in press.
- [11] Sammel, M.D., Ryan, L.M. & Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society, Series B* **59**, 667–678.
- [12] Skrondal, A. & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings, *Psychometrika* **68**, 267–287.
- [13] Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton.
- [14] Spiegelhalter, D.J. Thomas, A., Best, N.G. & Gilks, W.R. (1996). *BUGS 0.5 Bayesian Analysis using Gibbs Sampling. Manual (version ii)*, MRC-Biostatistics Unit, Cambridge, Downloadable from <http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml>.
- [15] U. S. Department of Defense. (1982). *Profile of American Youth*, Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, Washington.

ANDERS SKRONDAL AND
SOPHIA RABE-HESKETH

Structural Equation Modeling: Checking Substantive Plausibility

ADI JAFFE AND SCOTT L. HERSHBERGER

Volume 4, pp. 1910–1912

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Checking Substantive Plausibility

There is little doubt that without the process of validation, the undertaking of test development means little. Regardless of the reliability of an instrument (*see Reliability: Definitions and Estimation*), it is impossible for its developers, let alone its users, to be certain that any conclusions drawn from subjects' scores have meaning. In other words, without an understanding of the substantive plausibility of a set of measurement criteria, its implications cannot be confidently stated. It is widely accepted that for a test to be accepted as 'valid', numerous studies using different approaches are often necessary, and even then, these require an ongoing validation process as societal perceptions of relevant constructs change [2].

The classic work regarding the process of construct validation was that of Cronbach and Meehl [6]. Construct validation involves a three-step process; construct definition, construct operationalization, and empirical confirmation. Some of the methods commonly used to develop empirical support for construct validation are Campbell and Fiske's [5] **multitrait-multimethod** matrix method (MTMM) (*see Factor Analysis: Multitrait–Multimethod; Multitrait–Multimethod Analyses*) and **factor analysis**. However, contemporary **structural equation modeling** techniques [4] serve as good methods for evaluating what Cronbach and Meehl termed the *nomological network* (i.e., the relationship between constructs of interests and observable variables, as well as among the constructs themselves).

The process of construct validation usually starts with the theoretical analysis of the relationship between relevant variables or constructs. Sometimes known as the substantive component of construct validation, this step requires the definition of the construct of interest and its theoretical relationship to other constructs. For example, while it is a generally accepted fact that the degree of substance addiction is dependent on the length of substance use (i.e., short-term vs. chronic) [10], research has also demonstrated a relationship between drug

dependence and parental alcohol norms [8], consequences of use [11], and various personality factors [10]. A theoretical relationship could therefore be established between drug dependence – the construct to be validated – and these previously mentioned variables.

The next step in the process of construct validation requires the operationalization of the constructs of interest in terms of measurable, observed variables that relate to each of the specified constructs. Given the above mentioned example, length of substance use could easily be defined as years, months, or weeks of use, while parental alcohol norms could be assessed using either presently available instruments assessing parental alcohol use (e.g., the Alcohol Dependence Scale [12]) or some other measure of parental norms and expectancies regarding the use of alcohol and drugs. The consequences of past alcohol or drug use can be assessed as either number, or severity, of negative consequences associated with drinking. Various personality factors such as fatalism and loneliness can be assessed using available personality scales (e.g., the Social and Emotional Loneliness Scale for Adults [7]). The final structural equation model – nomological network – is shown in Figure 1.

The last step in the process of construct validation, empirical confirmation, requires the use of structural equation modeling to assess and specify the relationships between the observed variables, their related constructs, and the construct of interest. In its broadest sense, structural equation modeling is concerned with testing complex models for the structure of functional relationships between observed variables and constructs, and between the constructs themselves, one of which is the construct to be validated. Often, constructs in structural equation modeling are referred to as **latent variables**. The functional relationships are described by parameters that indicate the magnitude of the effect (direct or indirect) that independent variables have on dependent variables. Thus, a structural equation model can be considered a series of linear regression equations relating dependent variables to independent variables and other dependent variables. Those equations that describe the relations between observed variables and constructs are the *measurement* part of the model; those equations that describe the relations between constructs are the *structural* part of the model (*see All-X Models; All-Y Models*). The coefficients determining the

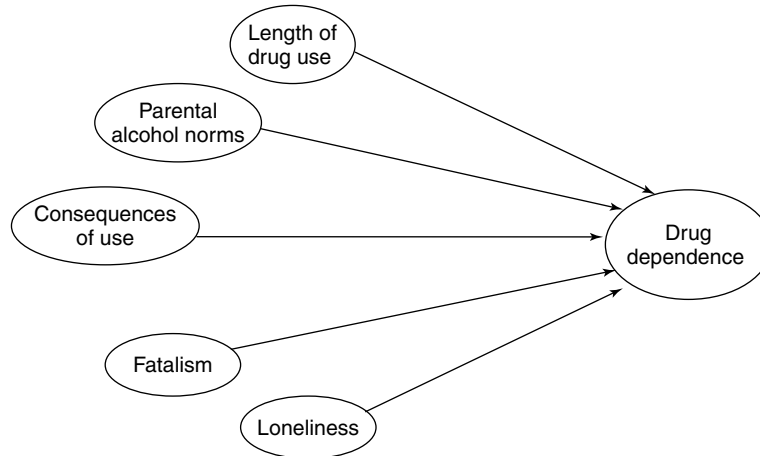


Figure 1 Nomological network for drug dependence

relations are usually the parameters we are interested in solving. By estimating the magnitude and direction of these parameters, one can evaluate the nomological network and hence provide evidence for construct validity. For example, in the model of Figure 1, we are hypothesizing that an observed score on the Alcohol Dependence Scale (not shown) is significantly and positively related to the construct of drug dependence. This relationship is part of the measurement model for drug dependence. Each construct has its own measurement model. To provide another example, we are also hypothesizing that the construct of length of drug use is significantly and positively related to the construct of drug dependence. This relation is part of the structural model linking the construct of interest, drug dependence, to the other constructs.

The most common sequence followed to confirm the nomological network is to first examine the individual measurement models of the constructs and then proceed to examine the structural model relating these constructs [1], although many variations to this sequence have been suggested [3, 9]. However one proceeds to evaluate the different components of the nomological network, ultimately the *global fit* of the nomological network must be evaluated. A great many indices of fit have been developed for this purpose, most of which define global fit in terms of the discrepancy between the observed data and that implied by the model parameters [4]. It is not until each component, both individually and combined, of

the nomological network has been confirmed that one has strong evidence of construct validity.

References

- [1] Anderson, J.C. & Gerbing, D.W. (1988). Structural equation modeling in practice: a review and recommended two-step approach, *Psychological Bulletin* **103**, 411–423.
- [2] Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example, *Educational Measurement: Issues and Practices* **17**, 10–17.
- [3] Bollen, K.A. (1989). *Structural Equation modeling: An Introduction*, Wiley, New York.
- [4] Bollen, K.A. (2000). Modeling strategies: in search of the Holy Grail, *Structural Equation Modeling* **7**, 74–81.
- [5] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
- [6] Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin* **52**, 281–302.
- [7] DiTommaso, E., Brannen, C. & Best, L.A. (2004). Measurement and validity characteristics of the short version of the social and emotional loneliness scale for adults, *Educational and Psychological Measurement* **64**, 99–119.
- [8] Fals-Stewart, W., Kelley, M.L., Cooke, C.G. & Golden, J.C. (2003). Predictors of the psychosocial adjustment of children living in households of parents in which fathers abuse drugs. The effects of postnatal parental exposure, *Addictive Behaviors* **28**, 1013–1031.
- [9] Mulaik, S.A. & Millsap, R.E. (2000). Doing the four-step right, *Structural Equation Modeling* **7**, 36–73.
- [10] Olmstead, R.E., Guy, S.M., O'Mally, P.M. & Bentler, P.M. (1991). Longitudinal assessment of the

- relationship between self-esteem, fatalism, loneliness, and substance use, *Journal of Social Behavior & Personality* **6**, 749–770.
- [11] Robinson, T.E. & Berridge, K.C. (2003). Addiction, *Annual Review of Psychology* **54**, 25–53.
- [12] Skinner, H.A. & Allen, B.A. (1982). Alcohol dependence syndrome: measurement and validation, *Journal of Abnormal Psychology* **91**, 199–209.

ADI JAFFE AND SCOTT L. HERSHBERGER

Structural Equation Modeling: Latent Growth Curve Analysis

JOHN B. WILLETT AND KRISTEN L. BUB

Volume 4, pp. 1912–1922

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Latent Growth Curve Analysis

Learning and development are ubiquitous. When new skills are acquired, when attitudes and interests develop, people change. Measuring change demands a longitudinal perspective (*see Longitudinal Data Analysis*), with multiple waves of data collected on representative people at sensibly spaced intervals (*see Panel Study*). Multiwave data is usually analyzed by individual **growth curve modeling** using a multilevel model for change (*see Generalized Linear Mixed Models* and [6]). Recently, innovative methodologists [2–4] have shown how the multilevel model for change can be mapped onto the general covariance structure model, such as that implemented in *LISREL* (*see Structural Equation Modeling: Software*). This has led to an alternative approach to analyzing change known as *latent growth modeling*. In this chapter, we describe and link these two analogous approaches.

Our presentation uses four waves of data on the reading scores of 1740 Caucasian children from the *Early Childhood Longitudinal Study (ECLS-K; [5])*. Children's reading ability was measured in the Fall and Spring of kindergarten and first grade – we assume that test administrations were six months apart, with time measured from entry into kindergarten. Thus, in our analyses, predictor t – representing time – has values 0, 0.5, 1.0 and 1.5 years. Finally, we know the child's gender (*FEMALE*: boy = 0, girl = 1), which we treat as a time-invariant predictor of change¹.

Introducing Individual Growth Modeling

In Figure 1, we display empirical reading records for ten children selected from the larger dataset. In the top *left* panel is the growth record of child #15013, a boy, with observed reading score on the ordinate and time on the abscissa. Reading scores are represented by a '+' symbol and are connected by a smooth freehand curve summarizing the change trajectory. Clearly, this boy's reading ability improves during kindergarten and first grade. In the top *right* panel, we display similar smoothed change trajectories for

all ten children (dashed trajectories for boys, solid for girls, plotting symbols omitted to reduce clutter). Notice the dramatic changes in children's observed reading scores over time, and how disparate they are from child to child. The complexity of the collection, and because true reading ability is obscured by measurement error, makes it hard to draw defensible conclusions about gender differences. However, perhaps the girls' trajectories do occupy higher elevations than those of the boys, on average.

Another feature present in the reading trajectories in the top two panels of Figure 1 is the apparent acceleration of the observed trajectories between Fall and Spring of first grade. Most children exhibit moderate growth in reading over the first three waves, but their scores increase dramatically over the last time period. The score of child #15013, for instance, rises modestly between waves 1 and 2 (20 to 28 points), modestly again between waves 2 and 3 (28 to 39 points), and then rapidly (to 66 points) by the fourth wave. Because of this nonlinearity, which was also evident in the entire sample, we transformed children's reading scores before further analysis (Singer & Willett [6, Chapter 6] comment on how to choose an appropriate transformation). We used a natural log transformation in order to 'pull down' the top end of the change trajectory disproportionately, thereby linearizing the accelerating raw trajectory.

In the lower panels of Figure 1, we redisplay the data in the newly transformed logarithmic world. The *log-reading* trajectory of child #15013 is now approximately linear in time, with positive slope. To dramatize this, we have superimposed a *linear* trend line on the transformed plot by simply regressing the log-reading scores on time using ordinary least squares regression (OLS) analysis for that child (*see Least Squares Estimation; Multiple Linear Regression*). This trend line has a positive slope, indicating that the log-reading score increases during kindergarten and first grade. In the lower right panel, we display OLS-fitted linear trajectories for all ten children in the subsample and reveal the heterogeneity in change that remains across children (albeit change in log-reading score). In subsequent analyses, we model change in the log-reading scores as a linear function of time.

The individual change trajectory can be described by a 'within-person' or 'level-1' individual growth model ([6], Ch. 3). For instance, here we hypothesize

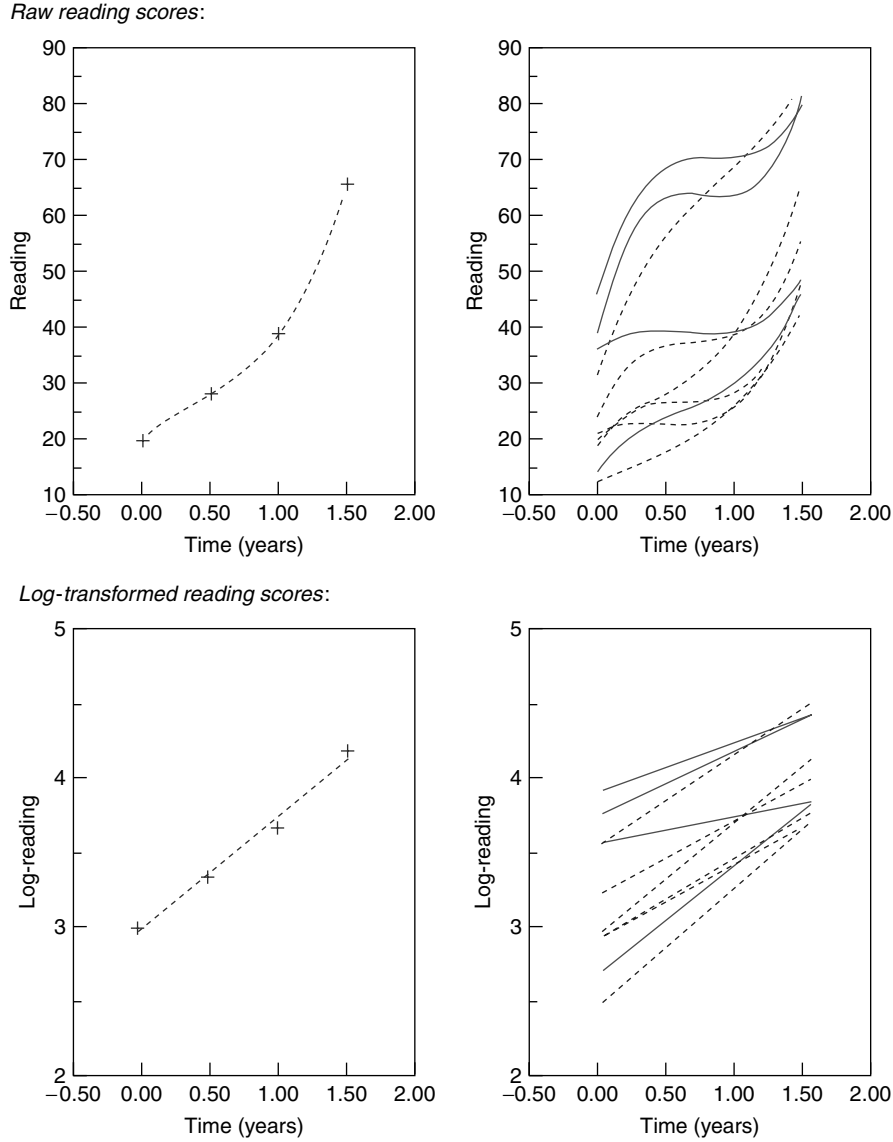


Figure 1 Observed raw and transformed trajectories of reading score over kindergarten and first grade for ten children (boys = dashed; girls = solid). *Top panel:* (a) raw reading score versus time for child #15013, with observed data points (+'s) connected by a smoothed freehand trajectory, (b) smoothed freehand trajectories for all 10 children. *Bottom panel:* (a) log-reading score versus time for child #15013, with an OLS-estimated linear change trajectory, (b) OLS-estimated linear trajectories for all children

that the log-reading score, Y_{ij} , of child i on occasion j is a linear function of time, t :

$$Y_{ij} = \{\pi_{0i} + \pi_{1i}t_j\} + \varepsilon_{ij}, \quad (1)$$

where $i = 1, 2, \dots, 1740$ and $j = 1, 2, 3, 4$ (with, as noted earlier, $t_1 = 0, t_2 = 0.5, t_3 = 1.0$ and $t_4 =$

1.5 years, respectively). We have bracketed the *systematic* part of the model to separate the orderly temporal dependence from the random errors, ε_{ij} , that accrue on each measurement occasion. Within the brackets, you will find the *individual growth parameters*, π_{0i} and π_{1i} :

- π_{0i} is the *intercept* parameter, describing the child's true 'initial' log-reading score on entry into kindergarten (because entry into kindergarten has been defined as the origin of time).
- π_{1i} is the *slope* ('rate of change') parameter, describing the child's true annual rate of change in log-reading score over time. If π_{1i} is positive, true log-reading score increases with time.

If the model is correctly specified, the individual growth parameters capture the defining features of the log-reading trajectory for child i . Of course, in specifying such models, you need not choose a linear specification – many shapes of trajectory are available, and the particular one that you choose should depend on your theory of change and on your inspection of the data [6, Chapter 6].

One assumption built deeply into individual growth modeling is that, while every child's change trajectory has the same functional form (here, linear in time), different children may have different values of the individual growth parameters. Children may differ in intercept (some have low log-reading ability on entry into kindergarten, others are higher) and in slope (some children change more rapidly with time, others less rapidly). Such heterogeneity is evident in the right-hand lower panel of Figure 1.

We have coded the trajectories in the right-hand panels of Figure 1 by child gender. Displays like these help to reveal systematic differences in change trajectory from child to child, and help you to assess whether interindividual variation in change is related to individual characteristics, like gender. Such 'level-2' questions – about the effect of predictors of change – translate into questions about 'between-person' relationships among the individual growth parameters and predictors like gender. Inspecting the right-hand lower panel of Figure 1, for instance, you can ask whether boys and girls differ in their initial scores (do the intercepts differ by gender?) or in the rates at which their scores change (do the slopes differ by gender?).

Analytically, we can handle this notion in a second 'between-person' or 'level-2' statistical model to represent interindividual differences in change. In the level-2 model, we express how we believe the individual growth parameters (standing in place of the full trajectory) depend on predictors of change. For example, we could investigate the impact of child gender on the log-reading trajectory by positing

the following pair of simultaneous level-2 statistical models:

$$\begin{aligned}\pi_{0i} &= \gamma_{00} + \gamma_{01} FEMALE_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} FEMALE_i + \zeta_{1i},\end{aligned}\quad (2)$$

where the level-2 residuals, ζ_{0i} and ζ_{1i} , represent those portions of the individual growth parameters that are 'unexplained' by the selected predictor of change, *FEMALE*. In this model, the γ coefficients are known as the 'fixed effects' and summarize the population relationship between the individual growth parameters and the predictor. They can be interpreted like regular regression coefficients. For instance, if the initial log-reading ability of girls is higher than boys (i.e., if girls have larger values of π_{0i} , on average) then γ_{01} will be positive (since *FEMALE* = 1, for girls). If girls have higher annual rates of change (i.e., if girls have larger values of π_{1i} , on average), then γ_{11} will be positive. Together, the level-1 and level-2 models in (1) and (2) make up the multilevel model for change ([6], Ch. 3).

Researchers investigating change must fit the multilevel model for change to their longitudinal data. Many methods are available for doing this (see [6], Chs. 2 and 3), the simplest of which is exploratory, as in Figure 1. To conduct data-analyses efficiently, the level-1 and level-2 models are usually fitted *simultaneously* using procedures now widely available in major statistical packages. The models can also be fitted using **covariance structure analysis**, as we now describe.

Latent Growth Modeling

Here, we introduce latent growth modeling by showing how the multilevel model for change can be mapped onto the general covariance structure model. Once the mapping is complete, all parameters of the multilevel model for change can be estimated by fitting the companion covariance structure model using standard covariance structure analysis (CSA) software, such as AMOS, LISREL, EQS, MPLUS, etc. (see **Structural Equation Modeling: Software**).

To conduct latent growth analyses, we lay out our data in multivariate format, in which there is a single row in the dataset for each person, with multiple (*multi-*) variables (*-variate*) containing the time-varying information, arrayed horizontally. With four waves of data, multivariate format requires four

4 Structural Equation Modeling: Latent Growth Curve Analysis

columns to record each child's growth record, each column associated with a measurement occasion. Any time-invariant predictor of change, like child gender, also has its own column in the dataset. Multivariate formatting is not typical in longitudinal data analysis (which usually requires a 'person-period' or 'univariate' format), but is required here because of the nature of covariance structure analysis. As its name implies, CSA is an analysis of covariance structure in which, as an initial step, a sample covariance matrix (and mean vector) is estimated to summarize the associations among (and levels of) selected variables, including the multiple measures of the outcome across the several measurement occasions. The data-analyst then specifies statistical models appropriate for the research hypotheses, and the mathematical implications of these hypotheses for the structure of the underlying population covariance matrix and mean vector are evaluated against their sample estimates. Because latent growth analysis compares sample and predicted covariance matrices (and mean vectors), the data must be formatted to support the estimation of covariance matrices (and mean vectors) – in other words, in a multivariate format.

Note, finally, that there is no unique column in the multivariate dataset to record time. In our multivariate format dataset, values in the outcome variable's first column were measured at the start of kindergarten, values in the second column were measured at the beginning of spring in kindergarten, etc. The time values – each corresponding to a particular measurement occasion and to a specific column of outcome values in the dataset – are noted by the analyst and programmed directly into the CSA model. It is therefore more convenient to use latent growth modeling to analyze change when panel data are *time-structured* – when everyone has been measured on an identical set of occasions and possesses complete data. Nevertheless, you can use latent growth modeling to analyze panel datasets with limited violations of time-structuring, by regrouping the full sample into subgroups who share identical time-structured profiles and then analyzing these subgroups simultaneously with CSA multigroup analysis (see **Factor Analysis: Multiple Groups**).

Mapping the Level-1 Model for Individual Change onto the CSA Y-measurement Model

In (1), we specified that the child's log-reading score, Y_{ij} , depended linearly on time, measured from

kindergarten entry. Here, for clarity, we retain symbols t_1 through t_4 to represent the measurement timing but you should remember that each of these time symbols has a known value (0, 0.5, 1.0, and 1.5 years, respectively) that is used when the model is fitted. By substituting into the individual growth model, we can create equations for the value of the outcome on each occasion for child i :

$$\begin{aligned} Y_{i1} &= \pi_{0i} + \pi_{1i}t_1 + \varepsilon_{i1} \\ Y_{i2} &= \pi_{0i} + \pi_{1i}t_2 + \varepsilon_{i2} \\ Y_{i3} &= \pi_{0i} + \pi_{1i}t_3 + \varepsilon_{i3} \\ Y_{i4} &= \pi_{0i} + \pi_{1i}t_4 + \varepsilon_{i4} \end{aligned} \quad (3)$$

that can easily be rewritten in simple matrix form, as follows:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{bmatrix} \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}. \quad (4)$$

While this matrix equation is unlike the representation in (1), it says exactly the same thing – that observed values of the outcome, Y , are related to the times (t_1, t_2, t_3 , and t_4), to the individual growth parameters (π_{0i} and π_{1i}), and to the measurement errors ($\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}$, and ε_{i4}). The only difference between (4) and (1) is that all values of the outcome and of time, and all parameters and time-specific residuals, are arrayed neatly as vectors and matrices. (Don't be diverted by the strange vector of zeros introduced immediately to the right of the equals sign – it makes no difference to the *meaning* of the equation, but it will help our subsequent mapping of the multilevel model for change onto the general CSA model).

In fact, the new growth model representation in (4) maps straightforwardly onto the CSA *Y-Measurement Model*, which, in standard *LISREL* notation, is

$$\mathbf{Y} = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (5)$$

where $\mathbf{Y}$ is a vector of observed scores, $\boldsymbol{\tau}_y$ is a vector intended to contain the population means of $\mathbf{Y}$, $\boldsymbol{\Lambda}_y$ is a matrix of factor loadings, $\boldsymbol{\eta}$ is a vector of latent (endogenous) constructs, and $\boldsymbol{\varepsilon}$ is a vector of residuals². Notice that the new matrix representation of the individual growth model in (4) matches the

CSA Y-Measurement Model in (5) providing that the observed and latent score vectors are set to:

$$\mathbf{Y} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \quad (6)$$

and providing that parameter vector $\boldsymbol{\tau}_y$ and loading matrix $\mathbf{\Lambda}_y$ are specified as containing the following constants and known times:

$$\boldsymbol{\tau}_y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{\Lambda}_y = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{bmatrix}. \quad (7)$$

Check this by substituting from (6) and (7) into (5) and multiplying out – you will conclude that, with this specification of score vectors and parameter matrices, the CSA Y-Measurement Model can act as, or contain, the individual growth trajectory from the multilevel model for change.

Notice that (1), (3), (4), (5), and (6) all permit the measurement errors to participate in the individual growth process. They state that level-1 residual ε_{i1} disturbs the true status of the i th child on the first measurement occasion, ε_{i2} on the second occasion, ε_{i3} on the third, and so on. However, so far, we have made no claims about the underlying distribution from which the errors are drawn. Are the errors normally distributed, homoscedastic, and independent over time within-person? Are they heteroscedastic or auto-correlated? Now that the individual change trajectory is embedded in the Y-Measurement Model, we can easily account for level-1 error covariance structure because, under the usual CSA assumption of a multivariate normal distribution for the errors, we can specify the CSA parameter matrix $\boldsymbol{\Theta}_\varepsilon$ to contain hypotheses about the covariance matrix of $\boldsymbol{\varepsilon}$. In an analysis of change, we usually compare nested models with alternative error structures to identify which error structure is optimal. Here, we assume that level-1 errors are distributed normally, independently, and heteroscedastically over time within-person:³

$$\boldsymbol{\Theta}_\varepsilon = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\varepsilon_3}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon_4}^2 \end{bmatrix}. \quad (8)$$

Ultimately, we estimate all level-1 variance components on the diagonal of $\boldsymbol{\Theta}_\varepsilon$ and reveal the

action of measurement error on reading score on each occasion.

The key point is that judicious specification of CSA score and parameter matrices forces the level-1 individual change trajectory into the Y-Measurement Model in a companion covariance structure analysis. Notice that, unlike more typical covariance structure analyses, for example, confirmatory factor analysis (see **Factor Analysis: Confirmatory**) the $\mathbf{\Lambda}_y$ matrix in (7) is entirely specified as a set of known constants and times rather than as unknown latent factor loadings to be estimated. Using the Y-Measurement Model to represent individual change in this way ‘forces’ the individual-level growth parameters, π_{0i} and π_{1i} , into the endogenous construct vector $\boldsymbol{\eta}$, creating what is known as the *latent growth vector*, $\boldsymbol{\eta}$. This notion – that the CSA $\boldsymbol{\eta}$ -vector can be *forced* to contain the individual growth parameters – is critical in latent growth modeling, because it suggests that level-2 interindividual variation in change can be modeled in the CSA Structural Model, as we now show.

Mapping the Level-2 Model for Interindividual Differences in Change onto the CSA Structural Model

In an analysis of change, at level-2, we ask whether interindividual heterogeneity in change can be predicted by other variables, such as features of the individual’s background and treatment. For instance, in our data-example, we can ask whether between-person heterogeneity in the log-reading trajectories depends on the child’s gender. Within the growth-modeling framework, this means that we must check whether the individual growth parameters – the true intercept and slope standing in place of the log-reading trajectories – are related to gender. Our analysis therefore asks: Does initial log-reading ability differ for boys and girls? Does the annual rate at which log-reading ability changes depend upon gender? In latent growth modeling, level-2 questions like these, which concern the distribution of the individual growth parameters across individuals and their relationship to predictors, are addressed by specifying a *CSA Structural Model*. Why? Because it is in the CSA structural model that the vector of unknown endogenous constructs $\boldsymbol{\eta}$ – which now contains the all-important individual growth parameters, π_{0i} and π_{1i} – is hypothesized to vary across people.

Recall that the CSA Structural Model stipulates that endogenous construct vector η is potentially related to both itself and to exogenous constructs ξ by the following model:

$$\eta = \alpha + \Gamma\xi + \mathbf{B}\eta + \zeta, \quad (9)$$

where α is a vector of intercept parameters, Γ is a matrix of regression coefficients that relate exogenous predictors ξ to outcomes η , $\mathbf{B}$ is a matrix of regression coefficients that permit elements of the endogenous construct vector η to predict each other, and ζ is a vector of residuals. In a covariance structure analysis, we fit this model to data, simultaneously with the earlier measurement model, and estimate parameters α , Γ and $\mathbf{B}$. The rationale behind latent growth modeling argues that, by structuring (9) appropriately, we can force parameter matrices α , Γ and $\mathbf{B}$ to contain the fixed effects central to the multilevel modeling of change.

So, what to do? Inspection of the model for systematic interindividual differences in change in (2) suggests that the level-2 component of the multilevel model for change can be reformatted in matrix form, as follows:

$$\begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix} [FEMALE_i] + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} + \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix}, \quad (10)$$

which is identical to the CSA Structural Model in (9), providing that we force the elements of the CSA $\mathbf{B}$ parameter matrix to be zero throughout:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (11)$$

and that we permit the α vector and the Γ matrix to be free to contain the fixed-effects parameters from the multilevel model for change:

$$\alpha = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix} \quad (12)$$

and providing we can force the potential predictor of change – child gender – into the CSA exogenous construct vector, ξ . In this new level-2 specification of the structural model, the latent intercept vector, α , contains the level-2 fixed-effects parameters γ_{00} and γ_{10} , defined earlier as the population intercept and slope of the long-reading trajectory for boys (when

FEMALE = 0). The Γ matrix contains the level-2 fixed-effects parameters γ_{01} and γ_{11} , representing increments to the population average intercept and slope for girls, respectively. By fitting this CSA model to data, we can estimate all four fixed effects.

When a time-invariant predictor like *FEMALE* is present in the structural model, the elements of the latent residual vector ζ in (10) represent those portions of true intercept and true slope that are *unrelated* to the predictor of change – the ‘adjusted’ values of true intercept and slope, with the linear effect of child gender partialled out. In a covariance structure analysis of the multilevel model for change, we assume that latent residual vector ζ is distributed normally with zero mean vector and covariance matrix Ψ ,

$$\Psi = \text{Cov}\{\zeta\} = \begin{bmatrix} \sigma_{\zeta_0}^2 & \sigma_{\zeta_{01}} \\ \sigma_{\zeta_{10}} & \sigma_{\zeta_1}^2 \end{bmatrix}, \quad (13)$$

which contains the residual (partial) variances and covariance of true intercept and slope, controlling for the predictor of change, *FEMALE*. We estimate these level-2 variance components in any analysis of change.

But there is one missing link that needs resolving before we can proceed. How is the hypothesized predictor of change, *FEMALE*, loaded into the exogenous construct vector, ξ ? This is easily achieved via the so-far-unused CSA *X-Measurement Model*. And, in the current analysis, the process is disarmingly simple because there is only a single infallible predictor of change, child gender. So, in this case, while it may seem a little weird, the specification of the X-Measurement Model derives from a tautology:

$$FEMALE_i = (0) + (1)(FEMALE_i) + (0). \quad (14)$$

This, while not affecting predictor *FEMALE*, facilitates comparison with the CSA X-Measurement Model:

$$\mathbf{X} = \tau_x + \Lambda_x \xi + \delta. \quad (15)$$

By comparing (14) and (15), you can see that the gender predictor can be incorporated into the analysis by specifying an X-Measurement Model in which:

- Exogenous score vector $\mathbf{X}$ contains one element, the gender predictor, *FEMALE*, itself.
- The X-measurement error vector, δ , contains a single element whose value is fixed at zero,

embodying the assumption that gender is measured infallibly (with ‘zero’ error).

- The τ_x mean vector contains a single element whose value is fixed at zero. This forces the mean of *FEMALE* (which would reside in τ_x if the latter were not fixed to zero) into the CSA latent mean vector, κ , which contains the mean of the exogenous construct, ξ , in the general CSA model.
- The matrix of exogenous latent factor loadings Λ_x contains a single element whose value is fixed at 1. This forces the metrics of the exogenous construct and its indicator to be identical.

Thus, by specifying a CSA X-Measurement Model in which the score vectors are

$$\mathbf{X} = [FEMALE_i], \quad \delta = [0] \quad (16)$$

and the parameter matrices are fixed at:

$$\tau_x = [0], \quad \Lambda_x = [1], \quad (17)$$

we can make the CSA exogenous construct ξ represent child gender. And, since we know that exogenous construct ξ is a predictor in the CSA Structural Model, we have succeeded in inserting the predictor of change, child gender, into the model for interindividual differences in change. As a final consequence of (14) through (17), the population mean of the predictor of change appears as the sole element of the CSA exogenous construct mean vector, κ :

$$\kappa = \text{Mean}\{\xi\} = [\mu_{FEMALE}] \quad (18)$$

and the population variance of the predictor of change appears as the sole element of CSA exogenous construct covariance matrix Φ :

$$\Phi = \text{Cov}\{\xi\} = [\sigma_{FEMALE}^2]. \quad (19)$$

Both of these parameter matrices are estimated when the model is fitted to data. And, although we do not demonstrate it here, the X-Measurement Model in (14) through (19) can be reconfigured to accommodate multiple time-invariant predictors of change, and even several indicators of each predictor construct if available. This is achieved by expanding the exogenous indicator and constructing score vectors to include sufficient elements to contain the new indicators and constructs, and the parameter matrix

Table 1 Trajectory of change in the logarithm of children’s reading score over four measurement occasions during kindergarten and first grade, by child gender. Parameter estimates, approximate *P* values, and goodness of fit statistics from the multilevel model for change, obtained with latent growth modeling (n = 1740)

Effect	Parameter	Estimate
<i>Fixed effects:</i>		
	γ_{00}	3.1700***
	γ_{10}	0.5828***
	γ_{01}	0.0732***
	γ_{11}	−0.0096
<i>Variance components:</i>		
	$\sigma_{\varepsilon_1}^2$	0.0219***
	$\sigma_{\varepsilon_2}^2$	0.0228***
	$\sigma_{\varepsilon_3}^2$	0.0208***
	$\sigma_{\varepsilon_4}^2$	0.0077***
	$\sigma_{\zeta_0}^2$	0.0896***
	$\sigma_{\zeta_1}^2$	0.0140***
	$\sigma_{\zeta_0\zeta_1}$	−0.0223***
<i>Goodness of fit:</i>		
	χ^2	1414.25***
	<i>Df</i>	7

~ = $p < .10$, * = $p < .05$, ** = $p < .01$, *** = $p < .00$.

Λ_x is expanded to include suitable loadings (Willett and Singer [6; Chapter 8] give an example with multiple predictors).

So, the CSA version of the multilevel model for change – now called the *latent growth model* – is complete. It consists of the CSA X-Measurement, Y-Measurement, and Structural Models, defined in (14) through (19), (4) through (8), and (9) through (13), respectively and is displayed as a path model in Figure 2. In the figure, by fixing the loadings associated with the outcome measurements to their constant and temporal values, we emphasize how the endogenous constructs were forced to become the individual growth parameters, which are then available for prediction by the hypothesized predictor of change. We fitted the latent growth model in (4) through (14) to our reading data on the full sample of 1740 children using LISREL (see Appendix I). Table 1 presents full maximum-likelihood (FML) estimates of all relevant parameters from latent regression-weight matrix Γ and parameter matrices Φ , α , and Ψ .

The estimated level-2 fixed effects are in the first four rows of Table 1. The first and second rows contain estimates of parameters γ_{00} and γ_{10} , representing true initial log-reading ability ($\hat{\gamma}_{00} =$

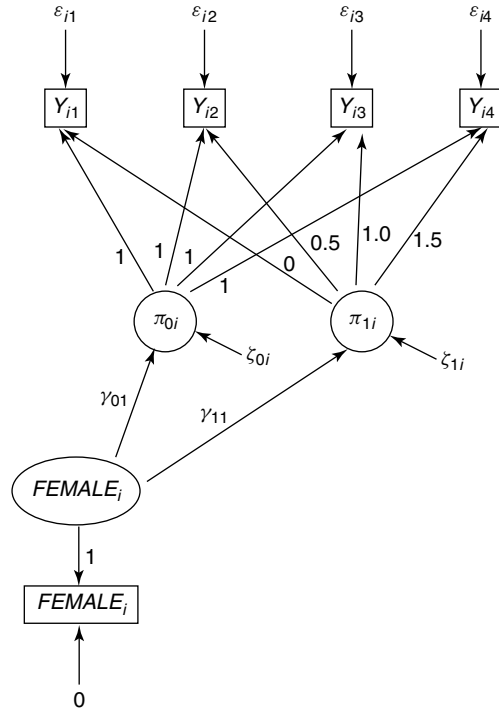
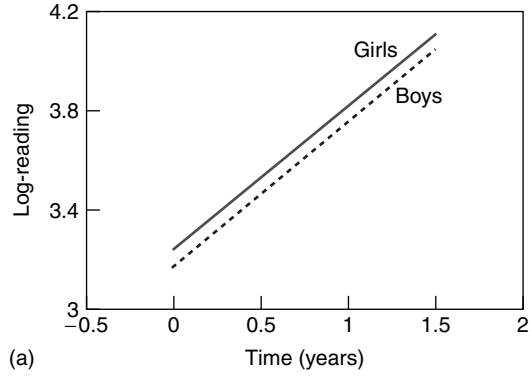


Figure 2 Path diagram for the hypothesized latent growth in reading score. Rectangles represent observed indicators, circles represent latent constructs, and arrows and their associated parameters indicate hypothesized relationships

3.170, $p < .001$) and true annual rate of change in log-reading ability ($\hat{\gamma}_{10} = 0.583$, $p < .001$) for boys (for whom $FEMALE = 0$). Anti-logging tells us that, on average, boys: (a) begin kindergarten with an average reading ability of 23.8 ($= e^{3.1700}$), and (b) increase their reading ability by 79% ($= 100(e^{0.5828} - 1)$) per year. The third and fourth rows contain the estimated latent regression coefficients γ_{01} (0.073, $p < .001$) and γ_{11} (-0.010 , $p > .10$), which capture differences in change trajectories between boys and girls. Girls have a higher initial level of 3.243 ($= 3.170 + 0.073$) of log-reading ability, whose anti-log is 25.6 and a statistically significant couple of points higher than the boys. However, we cannot reject the null hypothesis associated with γ_{11} (-0.010 , $p > .10$) so, although the estimated annual rate of increase in log-reading ability for girls is 0.572 ($= 0.5828 - 0.0096$), a little smaller than boys, this difference is not statistically significant. Nonetheless, anti-logging, we find that girls' reading

Fitted log-reading scores



Fitted reading scores



Figure 3 Fitted log-reading and reading trajectories over kindergarten and first grade for prototypical Caucasian children, by gender

ability increases by about 78% ($= 100(e^{0.5828} - 1)$) per year, on average. We display fitted log-reading and reading trajectories for prototypical boys and girls in Figure 3 – once de-transformed, the trajectories are curvilinear and display the acceleration we noted earlier in the raw data.

Next, examine the random effects. The fifth through eighth rows of Table 1 contain estimated level-1 error variances, one per occasion, describing the measurement fallibility in log-reading score over time. Their estimated values are 0.022, 0.023, 0.021, and 0.008, respectively, showing considerable homoscedasticity over the first three occasions but measurement error variance decreases markedly in the spring of first grade. The tenth through twelfth rows of Table 1 contain the estimated level-2 variance components, representing estimated partial (residual)

variances and partial covariance of true initial status and rate of change, after controlling for child gender (see **Partial Correlation Coefficients**). We reject the null hypothesis associated with each variance component, and conclude that there is predictable true variation remaining in both initial status and rate of change.

Conclusion: Extending the Latent Growth-Modeling Framework

In this chapter, we have shown how a latent growth-modeling approach to analyzing change is created by mapping the multilevel model for change onto the general CSA model. The basic latent growth modeling approach that we have described can be extended in many important ways:

- **You can include any number of waves of longitudinal data**, by simply increasing the number of rows in the relevant score vectors. Including more waves generally leads to greater precision for the estimation of the individual growth trajectories and greater reliability for measuring change.
- **You need not space the occasions of measurement equally**, although it is most convenient if everyone in the sample is measured on the *same set* of irregularly spaced occasions. However, if they are not, then latent growth modeling can still be conducted by first dividing the sample into subgroups of individuals with identical temporal profiles and using multigroup analysis to fit the multilevel model for change simultaneously in all subgroups.
- **You can specify curvilinear individual change**. Latent growth modeling can accommodate polynomial individual change of any order (provided sufficient waves of data are available), or any other curvilinear change trajectory in which individual status is linear in the growth parameters.
- **You can model the covariance structure of the level-1 measurement errors explicitly**. You need not accept the independence and homoscedasticity assumptions of classical analysis unchecked. Here, we permitted level-1 measurement errors to be heteroscedastic, but other, more general, error covariance structures can be hypothesized and tested.
- **You can model change in several domains simultaneously**, including both exogenous and

endogenous domains. You simply extend the empirical growth record and the measurement models to include rows for each wave of data available, in each domain.

- **You can model *intervening effects***, whereby an exogenous predictor may act directly on endogenous change and also indirectly via the influence of intervening factors, each of which may be time-invariant or time varying.

In the end, you must choose your analytic strategy to suit the problems you face. Studies of change can be designed in enormous variety and the multilevel model for change can be specified to account for all manner of trajectories and error structures. But, it is always wise to have more than one way to deal with data – latent growth modeling often offers a flexible alternative to more traditional approaches.

Notes

1. The dataset is available at <http://gseacademic.harvard.edu/~willettjo/>.
2. Readers unfamiliar with the general CSA model should consult Bollen [1].
3. Supplementary analyses suggested that this was reasonable.

Appendix I Specimen LISREL Program

```
/*Specify the number of variables
(indicators) to be read from
the external data-file of
raw data*/
data ni=6
/*Identify the location of the
external data-file*/
raw fi = C:\Data\ECLS.dat
/*Label the input variables and
select those to be analyzed*/
label
id Y1 Y2 Y3 Y4 FEMALE
select
2 3 4 5 6 /
/*Specify the hypothesized covariance
structure model*/
model ny=4 ne=2 ty=ze ly=fu,
fi te=di,fi c
nx=1 nk=1 lx=fu,fi tx=fr
```

```
        td=ze ph=sy,fr          c
        al=fr ga=fu,fr be=ze
        ps=sy,fr
/*Label the individual growth
parameters as endogenous
constructs (eta's)*/
le
pi0 pil
/*Label the predictor of change as
an exogenous construct (ksi) */
lk
FEMALE
/*Enter the required ``1's'' and
measurement times into the
Lambda-Y matrix*/
va 1 ly(1,1) ly(2,1) ly(3,1)
    ly(4,1)
va 0.0 ly(1,2)
va 0.5 ly(2,2)
va 1.0 ly(3,2)
va 1.5 ly(4,2)
/*Enter the required scaling factor
``1'' into the Lambda-X matrix*/
va 1.0 lx(1,1)
/*Free up the level-1 residual
variances to be estimated*/
fr te(1,1) te(2,2) te(3,3) te(4,4)
/*Request data-analytic output to 5
decimal places*/
ou nd=5
```

References

- [1] Bollen, K. (1989). *Structural Equations with Latent Variables*, Wiley, New York.
- [2] McArdle, J.J. (1986). Dynamic but structural equation modeling of repeated measures data, in *Handbook of Multivariate Experimental Psychology*, Vol. 2. J.R. Nesselrode & R.B. Cattell, eds, Plenum Press, New York, pp. 561–614.
- [3] Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.
- [4] Muthen, B.O. (1991). Analysis of longitudinal data using latent variable models with varying parameters, in *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions*, L.M. Collins & J.L. Horn, eds, American Psychological Association, Washington, pp. 1–17.
- [5] NCES (2002). *Early Childhood Longitudinal Study*. National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education, Washington.
- [6] Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, New York.

JOHN B. WILLETT AND KRISTEN L. BUB

Structural Equation Modeling: Mixture Models

JEROEN K. VERMUNT AND JAY MAGIDSON

Volume 4, pp. 1922–1927

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Mixture Models

Introduction

This article discusses a modeling framework that links two well-known statistical methods: **structural equation modeling** (SEM) and **latent class** or **finite mixture modeling**. This hybrid approach was proposed independently by Arminger and Stein [1], Dolan and Van der Maas [4], and Jedidi, Jagpal and DeSarbo [5]. Here, we refer to this approach as mixture SEM or latent class SEM.

There are two different ways to view mixture SEM. One way is as a refinement of multivariate normal (MVN) mixtures (*see* **Finite Mixture Distributions**), where the within-class covariance matrices are smoothed according to a postulated SEM structure. MVN mixtures have become a popular tool for **cluster analysis** [6, 10], where each cluster corresponds to a latent (unobservable) class. Names that are used when referring to such a use of mixture models are latent profile analysis, mixture-model clustering, model-based clustering, probabilistic clustering, Bayesian classification, and latent class clustering. Mixture SEM restricts the form of such latent class clustering, by subjecting the class-specific mean vectors and covariance matrices to a postulated SEM structure such as a one-factor, a latent-growth, or an autoregressive model. This results in MVN mixtures that are more parsimonious and stable than models with unrestricted covariance structures.

The other way to look at mixture SEM is as an extension to standard SEM, similar to multiple group analysis. However, an important difference between this and standard multiple group analysis, is that in mixture SEM group membership is not observed. By incorporating latent classes into an SEM model, various forms of unobserved heterogeneity can be detected. For example, groups that have identical (unstandardized) factor loadings but different error variances on the items in a factor analysis, or groups that show different patterns of change over time. Dolan and Van der Maas [4] describe a nice application from developmental psychology, in which (as a result of the existence of qualitative development

stages) children who do not master certain types of tasks have a mean and covariance structure that differs from the one for children who master the tasks.

Below, we first introduce standard MVN mixtures. Then, we show how the SEM framework can be used to restrict the means and covariances. Subsequently, we discuss parameter estimation, model testing, and software. We end with an empirical example.

Multivariate Normal Mixtures

Let $\mathbf{y}_i$ denote a P -dimensional vector containing the scores for individual i on a set of P observed continuous random variables. Moreover, let K be the number of mixture components, latent classes, or clusters, and π_k the prior probability of belonging to latent class or cluster k or, equivalently, the size of cluster k , where $1 \leq k \leq K$. In a mixture model, it is assumed that the density of $\mathbf{y}_i$, $f(\mathbf{y}_i|\boldsymbol{\theta})$, is a mixture or a weighted sum of K class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ [4, 10]. That is,

$$f(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i|\boldsymbol{\theta}_k). \quad (1)$$

Here, $\boldsymbol{\theta}$ denotes the vector containing all unknown parameters, and $\boldsymbol{\theta}_k$ the vector of the unknown parameters of cluster k .

The most common specification for the class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is multivariate normal (*see* **Catalogue of Probability Density Functions**), which means that the observed variables are assumed to be normally distributed within latent classes, possibly after applying an appropriate nonlinear transformation. Denoting the class-specific mean vector by $\boldsymbol{\mu}_k$, and the class-specific covariance matrix by $\boldsymbol{\Sigma}_k$, we obtain the following class-specific densities:

$$f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-P/2} |\boldsymbol{\Sigma}_k|^{-1/2} \times \exp \left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right\}. \quad (2)$$

In the most general specification, no restrictions are imposed on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ parameters; that is, the model-based clustering problem involves estimating a separate set of means, variances, and covariances for each latent class. Although in most clustering applications, the main objective is finding classes that differ with respect to their means or locations, in the

2 Structural Equation Modeling: Mixture Models

MVN mixture model, clusters may also have different shapes.

An unrestricted MVN mixture model with K latent classes contains $(K - 1)$ unknown class sizes, $K \cdot P$ class-specific means, $K \cdot P$ class-specific variances and $K \cdot P \cdot (P - 1)/2$ class-specific covariances. As the number of indicators and/or the number of latent classes increases, the number of parameters to be estimated may become quite large, especially the number of free parameters in Σ_k . Thus, to obtain more parsimony and stability, it is not surprising that restrictions are typically imposed on the class-specific covariance matrices.

Prior to using SEM models to restrict the covariances, a standard approach to reduce the number of parameters is to assume local independence. Local independence means that all within-cluster covariances are equal to zero, or, equivalently, that the covariance matrices, Σ_k , are diagonal matrices. Models that are less restrictive than the local independence model can be obtained by fixing some but not all covariances to zero, or, equivalently, by assuming certain pairs of y 's to be mutually dependent within latent classes.

Another approach to reduce the number of parameters, is to assume the equality or homogeneity of variance–covariance matrices across latent classes; that is, $\Sigma_k = \Sigma$. Such a homogeneous or class-independent error structure yields clusters having the same forms but different locations. This type of constraint is equivalent to the restrictions applied to the covariances in linear discriminant analysis. Note that this between-class equality constraint can be applied in combination with any structure for Σ .

Banfield and Raftery [2] proposed reparameterizing the class-specific covariance matrices by an eigenvalue decomposition:

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k'. \quad (3)$$

The parameter λ_k is a scalar, $\mathbf{D}_k$ is a matrix with eigenvectors, and $\mathbf{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k . More precisely, $\lambda_k = |\Sigma_k|^{1/d}$, where d is the number of observed variables, and $\mathbf{A}_k$ is scaled such that $|\mathbf{A}_k| = 1$.

A nice feature of the above decomposition is that each of the three sets of parameters has a geometrical interpretation: λ_k indicates what can be called the volume of cluster k , $\mathbf{D}_k$ its orientation, and $\mathbf{A}_k$ its shape. If we think of a cluster as a

clutter of points in a multidimensional space, the volume is the size of the clutter, while the orientation and shape parameters indicate whether the clutter is spherical or ellipsoidal. Thus, restrictions imposed on these matrices can directly be interpreted in terms of the geometrical form of the clusters. Typical restrictions are to assume matrices to be equal across classes, or to have the forms of diagonal or identity matrices [3].

Mixture SEM

As an alternative to simplifying the Σ_k matrices using the eigenvalue decomposition, the mixture SEM approach assumes a **covariance-structure model**. Several authors [1, 4, 5] have proposed using such a mixture specification for dealing with unobserved heterogeneity in SEM. As explained in the introduction, this is equivalent to restricting the within-class mean vectors and covariance matrices by an SEM. One interesting SEM structure for Σ_k that is closely related to the eigenvalue decomposition described above is a factor-analytic model (*see Factor Analysis: Exploratory*) [6, 11]. Under the factor-analytic structure, the within-class covariances are given by:

$$\Sigma_k = \Lambda_k \Phi_k \Lambda_k' + \Theta_k. \quad (4)$$

Assuming that there are Q factors, Λ_k is a $P \times Q$ matrix with factor loadings, Φ_k is a $Q \times Q$ matrix containing the variances of, and the covariances between, the factors, and Θ_k is a $P \times P$ diagonal matrix containing the unique variances. Restricted covariance structures are obtained by setting $Q < P$ (for instance, $Q = 1$), equating factor loadings across indicators, or fixing some factor loading to zero. Such specifications make it possible to describe the covariances between the y variables within clusters by means of a small number of parameters.

Alternative formulations can be used to define more general types of SEM models. Here, we use the Lisrel submodel that was also used by Dolan and Van der Maas [4]. Other alternatives are the full Lisrel [5], the RAM [8], or the conditional mean and covariance structure [1] formulations.

In our Lisrel submodel formulation, the SEM for class k consists of the following two (sets of) equations:

$$\mathbf{y}_i = \mathbf{v}_k + \Lambda_k \boldsymbol{\eta}_{ik} + \boldsymbol{\epsilon}_{ik} \quad (5)$$

$$\eta_{ik} = \alpha_k + \mathbf{B}_k \eta_{ik} + \zeta_{ik}. \quad (6)$$

The first equation concerns the measurement part of the model, in which the observed variables are regressed on the latent factors η_{ik} . Here, $\mathbf{v}_k$ is a vector of intercepts, $\mathbf{\Lambda}_k$ a matrix with factor loadings and $\boldsymbol{\varepsilon}_{ik}$ a vector with residuals. The second equation is the structural part of the model, the path model for the factors. Vector α_k contains the intercepts, matrix $\mathbf{B}_k$ the path coefficients and vector ζ_{ik} the residuals. The implied mean and covariance structures for latent class k are

$$\mu_k = \mathbf{v}_k + \mathbf{\Lambda}_k (\mathbf{I} - \mathbf{B}_k)^{-1} \alpha_k \quad (7)$$

$$\Sigma_k = \mathbf{\Lambda}_k (\mathbf{I} - \mathbf{B}_k)^{-1} \Phi_k (\mathbf{I} - \mathbf{B}_k')^{-1} \mathbf{\Lambda}_k' + \Theta_k, \quad (8)$$

where Θ_k and Φ_k denote the covariance matrices of the residuals $\boldsymbol{\varepsilon}_{ik}$ and ζ_{ik} . These equations show the connection between the SEM parameters, and the parameters of the MVN mixture model.

Covariates

An important extension of the mixture SEM described above is obtained by including covariates to predict class membership, with possible direct effects on the item means. Conceptually, it makes sense to distinguish (endogenous) variables that are used to identify the latent classes, from (exogenous) variables that are used to predict to which cluster an individual belongs.

Using the same basic structure as in 1, this yields the following mixture model:

$$f(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(\mathbf{z}_i) f_k(\mathbf{y}_i | \boldsymbol{\theta}_k). \quad (9)$$

Here, $\mathbf{z}_i$ denotes person i 's covariate values. Alternative terms for the $\mathbf{z}$'s are concomitant variables, grouping variables, external variables, exogenous variables, and inputs. To reduce the number of parameters, the probability of belonging to class k given covariate values $\mathbf{z}_i$, $\pi_k(\mathbf{z}_i)$, will generally be restricted by a multinomial logit model; that is, a logit model with 'linear effects' and no higher order interactions.

An even more general specification is obtained by allowing covariates to have direct effects on the indicators, which yields

$$f(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k(\mathbf{z}_i) f_k(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\theta}_k). \quad (10)$$

The conditional means of the $\mathbf{y}$ variables are now directly related to the covariates, as proposed by Arminger and Stein [1]. This makes it possible to relax the implicit assumption in the previous specification, that the influence of the $\mathbf{z}$'s on the $\mathbf{y}$'s goes completely via the latent classes (see, for example, [9]).

Estimation, Testing, and Software

Estimation

The two main estimation methods in mixture SEM and other types of MVN mixture modeling are **maximum likelihood (ML)** and maximum posterior (MAP). The log-likelihood function required in ML and MAP approaches can be derived from the probability density function defining the model. Bayesian MAP estimation involves maximizing the log-posterior distribution, which is the sum of the log-likelihood function and the logs of the priors for the parameters (*see Bayesian Statistics*).

Although generally, there is not much difference between ML and MAP estimates, an important advantage of the latter method is that it prevents the occurrence of boundary or terminal solutions: probabilities and variances cannot become zero. With a very small amount of prior information, the parameter estimates are forced to stay within the interior of the parameter space. Typical priors are Dirichlet priors for the latent class probabilities, and inverted-Wishart priors for the covariance matrices. For more details on these priors, see [9].

Most mixture modeling software packages use the EM algorithm, or some modification of it, to find the ML or MAP estimates. In our opinion, the ideal algorithm starts with a number of EM iterations, and when close enough to the final solution, switches to Newton–Raphson. This is a way to combine the advantages of both algorithms – the stability of EM even when far away from the optimum, and the speed of Newton–Raphson when close to the optimum (*see Optimization Methods*).

A well-known problem in mixture modeling analysis is the occurrence of local solutions. The best way to prevent ending with a local solution is to use multiple sets of starting values. Some computer programs for mixture modeling have automated the search for good starting values using several sets of random starting values.

4 Structural Equation Modeling: Mixture Models

When using mixture SEM for clustering, we are not only interested in the estimation of the model parameters, but also in the classification of individual into clusters. This can be based on the posterior class membership probabilities

$$\pi_k(\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\pi_k(\mathbf{z}_i) f_k(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\theta}_k)}{\sum_{k=1}^K \pi_k(\mathbf{z}_i) f_k(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\theta}_k)}. \quad (11)$$

The standard classification method is modal allocation, which amounts to assigning each object to the class with the highest posterior probability.

Model Selection

The model selection issue is one of the main research topics in mixture-model clustering. Actually, there are two issues involved: the first concerns the decision about the number of clusters, the second concerns the form of the model, given the number of clusters. For an extended overview on these topics, see [6].

Assumptions with respect to the forms of the clusters, given their number, can be tested using standard likelihood-ratio tests between nested models, for instance, between a model with an unrestricted covariance matrix and a model with a restricted covariance matrix. Wald tests and Lagrange multiplier tests can be used to assess the significance of certain included or excluded terms, respectively. However, these kinds of chi-squared tests cannot be used to determine the number of clusters.

The approach most often used for model selection in mixture modeling is to use information criteria, such as AIC, BIC, and CAIC (**Akaike**, Bayesian, and Consistent Akaike Information Criterion). The most recent development is the use of computationally intensive techniques like parametric bootstrapping [6] and Markov Chain Monte Carlo methods [3] to determine the number of clusters, as well as their forms.

Another approach for evaluating mixture models is based on the uncertainty of classification, or, equivalently, the separation of the clusters. Besides the estimated total number of misclassifications, Goodman–Kruskal lambda, Goodman–Kruskal tau, or entropy-based measures can be used to indicate how well the indicators predict class membership.

Software

Several computer programs are available for estimating the various types of mixture models discussed in this paper. Mplus [7] and Mx [8] are syntax-based programs that can deal with a very general class of mixture SEMs. Mx is somewhat more general in terms of model possible constraints. Latent GOLD [9] is a fully Windows-based program for estimating MVN mixtures with covariates. It can be used to specify restricted covariance structures, including a number of SEM structures such as a one-factor model within blocks of variables, and a compound-symmetry (or random-effects) structure.

An Empirical Example

To illustrate mixture SEM, we use a longitudinal data set made available by Patrick J. Curran at <http://www.duke.edu/~curran/>. The variable of interest is a child's reading recognition skill measured at four two-year intervals using the Peabody Individual Achievement Test (PIAT) Reading Recognition subtest. The research question of interest is whether a one-class model with its implicit assumption that a single pattern of reading development holds universally is correct, or whether there are different types of reading recognition trajectories among different latent groups. Besides information on reading recognition, we have information on the child's gender, the mother's age, the child's age, the child's cognitive stimulation at home, and the child's emotional support at home. These variables will be used as covariates. The total sample size is 405, but only 233 children were measured at all assessments. We use all 405 cases in our analysis, assuming that the missing data is missing at random (MAR) (see **Dropouts in Longitudinal Data**). For parameter estimation, we used the Latent GOLD and Mx programs.

One-class to three-class models (without covariates) were estimated under five types of SEM structures fitted to the within-class covariance matrices. These SEM structures are local independence (LI), saturated (SA), random effects (RE), autoregressive (AR), and one factor (FA). The BIC values reported in Table 1 indicate that two classes are needed when using a SA, AR, or FA structure.¹ As is typically the case, working with a misspecified covariance structure (here, LI or RE), yields an overestimation of

Table 1 Test results for the child's reading recognition example

Model	log-likelihood	# parameters	BIC
A1. 1-class LI	-1977	8	4003
A2. 2-class LI	-1694	17	3490
A3. 3-class LI	-1587	26	3330
B1. 1-class SA	-1595	14	3274
B2. 2-class SA	-1489	29	3151
B3. 3-class SA	-1459	44	3182
C1. 1-class RE	-1667	9	3375
C2. 2-class RE	-1561	19	3237
C3. 3-class RE	-1518	29	3211
D1. 1-class AR	-1611	9	3277
D2. 2-class AR	-1502	19	3118
D3. 3-class AR	-1477	29	3130
E1. 1-class FA	-1611	12	3294
E2. 2-class FA	-1497	25	3144
E3. 3-class FA	-1464	38	3157
F. D2 + covariates	-1401	27	2964

the number of classes. Based on the BIC criterion, the two-class AR model (Model D2) is the model that is preferred. Note that this model captures the dependence between the time-specific measures with a single path coefficient, since the coefficients associated with the autoregressive component of the model are assumed to be equal for each pair of adjacent time points.

Subsequently, we included the covariates in the model. Child's age was assumed to directly affect the indicators, in order to assure that the encountered trajectories are independent of the child's age at the first occasion. Child's gender, mother's age, child's cognitive stimulation, and child's emotional support were assumed to affect class membership. According to the BIC criterion, this model (Model F) is much better than the model without covariates (Model D2).

According to Model F, Class 1 contains 61% and class 2 39% of the children. The estimated means for class 1 are 2.21, 3.59, 4.51, and 5.22, and for class 2, 3.00, 4.80, 5.81, and 6.67. These results show that class 2 starts at a higher level and grows somewhat faster than class 1. The estimates of the class-specific variances are 0.15, 0.62, 0.90 and 1.31 for class 1, and 0.87, 0.79, 0.94, and 0.76 for class 2. This indicates that the within-class heterogeneity increases dramatically within class 1, while it is quite stable within class 2. The

estimated values of the class-specific path coefficients are 1.05 and 0.43, respectively, indicating that even with the incrementing variance, the autocorrelation is larger in latent class 1 than in latent class 2.²

The age effects on the indicators are highly significant. As far as the covariate effects on the log-odds of belonging to class 2 instead of class 1 are concerned, only the mother's age is significant. The older the mother, the higher the probability of belonging to latent class 2.

Notes

1. BIC is defined as minus twice the log-likelihood plus $\ln(N)$ times the number of parameters, where N is the sample size (here 450).
2. The autocorrelation is a standardized path coefficient that can be obtained as the product of the unstandardized coefficient and the ratio of the standard deviations of the independent and the dependent variable in the equation concerned. For example, the class 1 autocorrelation between time points 1 and 2 equals $1.05(\sqrt{0.15}/\sqrt{0.62})$.

References

- [1] Arminger, G. & Stein, P. (1997). Finite mixture of covariance structure models with regressors: loglikelihood function, distance estimation, fit indices, and a complex example, *Sociological Methods and Research* **26**, 148–182.
- [2] Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**, 803–821.
- [3] Bensmail, H., Celeux, G., Raftery, A.E. & Robert, C.P. (1997). Inference in model based clustering, *Statistics and Computing* **7**, 1–10.
- [4] Dolan, C.V. & Van der Maas, H.L.J. (1997). Fitting multivariate normal finite mixtures subject to structural equation modeling, *Psychometrika* **63**, 227–253.
- [5] Jedidi, K., Jagpal, H.S. & DeSarbo, W.S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity, *Marketing Science* **16**, 39–59.
- [6] McLachlan, G.J. & Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.
- [7] Muthén, B. & Muthén, L. (1998). *Mplus: User's Manual*, Muthén & Muthén, Los Angeles.
- [8] Neale, M.C., Boker, S.M., Xie, G. & Maes, H.H. (2002). *Mx: Statistical Modeling*, VCU, Department of Psychiatry, Richmond.
- [9] Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD's User's Guide*, Statistical Innovations, Boston.

6 Structural Equation Modeling: Mixture Models

- [10] Vermunt, J.K. & Magidson, J. (2002). Latent class cluster analysis, in *Applied Latent Class Analysis*, J.A. Hage-naars & A.L. McCutcheon, eds, Cambridge University Press, Cambridge, pp. 89–106.
- [11] Yung, Y.F. (1997). Finite mixtures in confirmatory factor-analysis models, *Psychometrika* **62**, 297–330.

JEROEN K. VERMUNT AND JAY MAGIDSON

Structural Equation Modeling: Multilevel

FIONA STEELE

Volume 4, pp. 1927–1931

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Multilevel

Multilevel Factor Models

Suppose that for each individual i we observe a set of R continuous responses $\{y_{ri} : r = 1, \dots, R\}$. In standard **factor analysis**, we assume that the pairwise correlations between the responses are wholly explained by their mutual dependency on one or more underlying factors, also called **latent variables**. If there is only one such factor, the factor model may be written:

$$y_{ri} = \alpha_r + \lambda_r \eta_i + e_{ri}, \quad (1)$$

where α_r is the grand mean for response r , η_i is a factor with loading λ_r for response r , and e_{ri} is a residual. We assume that both the factor and the residuals are normally distributed. In addition, we assume that the residuals are uncorrelated, which follows from the assumption that the correlation between the responses is due to their dependency on the factor; conditional on this factor, the responses are independent.

A further assumption of model (1) is that the y 's are independent across individuals. Often, however, individuals will be clustered in some way, for example, in areas or institutions, and responses for individuals in the same cluster are potentially correlated. In the context of regression analysis, multilevel or **hierarchical models** have been developed to account for within-cluster correlation and to explore between-cluster variation. Multilevel models include cluster-level residuals or random effects that represent unobserved variables operating at the cluster level; conditional on the random effects, individuals' responses are assumed to be independent (see **Generalized Linear Mixed Models**). The factor model has also been extended to handle clustering, leading to a multilevel factor model (see [3, 5]). In the multilevel factor model, in addition to having residuals at both the individual and cluster level as in a multilevel regression model, there may be factors at both levels. Suppose, for example, that academic ability is assessed using a series of tests. An individual's score on each of these tests is likely to depend on his overall ability (represented by an individual-level factor) and the ability of children in the same

school (a school-level factor). A multilevel extension of (1) with a single factor at the cluster level may be written:

$$y_{rij} = \alpha_r + \lambda_r^{(1)} \eta_{ij}^{(1)} + \lambda_r^{(2)} \eta_j^{(2)} + u_{rj} + e_{rij}, \quad (2)$$

where y_{rij} is response r for individual i ($i = 1, \dots, n_j$) in cluster j ($j = 1, \dots, J$), $\eta_{ij}^{(1)}$ and $\eta_j^{(2)}$ are factors at the individual and cluster levels (levels 1 and 2 respectively) with loadings $\lambda_r^{(1)}$ and $\lambda_r^{(2)}$, and u_{rj} and e_{rij} are residuals at the individual and cluster levels. We assume $\eta_{ij}^{(1)} \sim N(0, \sigma_{\eta^{(1)}}^2)$, $\eta_j^{(2)} \sim N(0, \sigma_{\eta^{(2)}}^2)$, $u_{rj} \sim N(0, \sigma_{ur}^2)$ and $e_{rij} \sim N(0, \sigma_{er}^2)$.

As is usual in factor analysis, constraints on either the factor loadings or the factor variances are required in order to fix the scale of the factors. In the example that follows, we will constrain the first loading of each factor to one, which constrains each factor to have the same scale as the first response. An alternative is to constrain the factor variances to one. If both the factors and responses are standardized to have unit variance, factor loadings can be interpreted as correlations between a factor and the responses. If responses are standardized and constraints are placed on factor loadings rather than factor variances, we can compute standardized loadings for a level k factor (omitting subscripts) as $\lambda_r^{(k)*} = \lambda_r^{(k)} \sigma_{\eta^{(k)}}$.

The model in (2) may be extended in a number of ways. Goldstein and Browne [2] propose a general factor model with multiple factors at each level, correlations between factors at the same level, and covariate effects on the responses.

An Example of Multilevel Factor Analysis

We will illustrate the application of multilevel factor analysis using a dataset of science scores for 2439 students in 99 Hungarian schools. The data consist of scores on four test booklets: a core booklet with components in earth science, physics and biology, two biology booklets and one in physics. Therefore, there are six possible test scores (one earth science, three biology, and two physics). Each student responds to a maximum of five tests, the three tests in the core booklet, plus a randomly selected pair of tests from the other booklets. Each test is marked out of ten. A detailed description of the data is given in [1]. All analysis presented here is based on standardized test scores.

2 Structural Equation Modeling: Multilevel

Table 1 Pairwise correlations (variances on diagonal) at school and student levels

	E. Sc. core	Biol. core	Biol. R3	Biol. R4	Phys. core	Phys. R2
<i>School level</i>						
E. Sc. core	0.16					
Biol. core	0.68	0.22				
Biol. R3	0.51	0.68	0.07			
Biol. R4	0.46	0.67	0.46	0.23		
Phys. core	0.57	0.89	0.76	0.62	0.24	
Phys. R2	0.56	0.77	0.58	0.64	0.77	0.16
<i>Student level</i>						
E. Sc. core	0.84					
Biol. core	0.27	0.78				
Biol. R3	0.12	0.13	0.93			
Biol. R4	0.14	0.27	0.19	0.77		
Phys. core	0.26	0.42	0.10	0.29	0.77	
Phys. R2	0.22	0.34	0.15	0.39	0.43	0.83

Table 2 Estimates from two-level factor model with one factor at each level

	Student level			School level		
	$\lambda_r^{(1)}$ (SE)	$\lambda_r^{(1)*}$	σ_{er}^2 (SE)	$\lambda_r^{(2)}$ (SE)	$\lambda_r^{(2)*}$	σ_{ur}^2 (SE)
E. Sc. core	1 ^a	0.24	0.712 (0.023)	1 ^a	0.36	0.098 (0.021)
Biol. core	1.546 (0.113)	0.37	0.484 (0.022)	2.093 (0.593)	0.75	0.015 (0.011)
Biol. R3	0.583 (0.103)	0.14	0.892 (0.039)	0.886 (0.304)	0.32	0.033 (0.017)
Biol. R4	1.110 (0.115)	0.27	0.615 (0.030)	1.498 (0.466)	0.53	0.129 (0.029)
Phys. core	1.665 (0.128)	0.40	0.422 (0.022)	2.054 (0.600)	0.73	0.036 (0.013)
Phys. R2	1.558 (0.133)	0.37	0.526 (0.030)	1.508 (0.453)	0.54	0.057 (0.018)

^aConstrained parameter.

Table 1 shows the correlations between each pair of standardized test scores, estimated from a multivariate multilevel model with random effects at the school and student levels. Also shown are the variances at each level; as is usual with attainment data, the within-school (between-student) variance is substantially larger than the between-school variance. The correlations at the student level are fairly low. Goldstein [1] suggests that this is due to the fact that there are few items in each test so the student-level reliability is low. Correlations at the school level (i.e., between the school means) are moderate to high, suggesting that the correlations at this level might be well explained by a single factor.

The results from a two-level factor model, with a single factor at each level, are presented in Table 2. The factor variances at the student and school levels are estimated as 0.127(SE = 0.016) and 0.057(SE = 0.024) respectively. As at each level the standardized

loadings have the same sign, we interpret the factors as student-level and school-level measures of overall attainment in science. We note, however, that the student-level loadings are low, which is a result of the weak correlations between the test scores at this level (Table 1). Biology R3 has a particularly low loading; the poor fit for this test is reflected in a residual variance estimate (0.89), which is close to the estimate obtained from the multivariate model (0.93). Thus, only a small amount of the variance in the scores for this biology test is explained by the student-level factor, that is, the test has a low communality. At the school level the factor loadings are higher, and the school-level residual variances from the factor model are substantially lower than those from the multivariate model. The weak correlations between the test scores and the student-level factor suggest that another factor at this level may improve the model. When a second student-level factor is added to the

model, however, the loadings on one factor have very large standard errors (results not shown). Goldstein and Browne [2] consider an alternative specification with two correlated student-level factors, but we do not pursue this here. In their model, the loadings for physics on one factor are constrained to zero, and on the other factor, zero constraints are placed on the loadings for all tests except physics.

Multilevel Structural Equation Models

While it is possible to include covariates in a multilevel factor model (see [2]), we are often interested in the effects of covariates on the underlying factors rather than on each response. We may also wish to allow a factor to depend on other factors. A **structural equation model** (SEM) consists of two components: (a) a measurement model in which the multivariate responses are assumed to depend on factors (and possibly covariates), and (b) a structural model in which factors depend on covariates and possibly other factors.

A simple multilevel SEM is:

$$y_{rij} = \alpha_r + \lambda_r^{(1)} \eta_{ij}^{(1)} + u_{rj} + e_{rij} \quad (3)$$

$$\eta_{ij}^{(1)} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}. \quad (4)$$

In the structural model (4), the individual-level factor is assumed to be a linear function of a covariate x_{ij} and cluster-level random effects u_j , which are assumed to be normally distributed with variance σ_u^2 . Individual-level residuals e_{ij} are assumed to be normally distributed with variance σ_e^2 . To fix the scale

of $\eta_{ij}^{(1)}$, the intercept in the structural model, β_0 , is constrained to zero and one of the factor loadings is constrained to one.

The model defined by (3) and (4) is a multilevel version of what is commonly referred to as a multiple indicators multiple causes (MIMIC) model. If we substitute (4) in (3), we obtain a special case of the multilevel factor model with covariates, in which u_j is a cluster-level factor with loadings equal to the individual-level factor loadings. If we believe that the factor structure differs across levels, a cluster-level factor can be added to the measurement model (3) and u_j removed from (4). A further equation could then be added to the structural model to allow the cluster-level factor to depend on cluster-level covariates. Another possible extension is to allow for dependency between factors, either at the same or different levels.

In the MIMIC model x is assumed to have an indirect effect on the y 's through the factor. The effect of x on y_r is $\lambda_r^{(1)} \beta_1$. If instead we believe that x has a direct effect on the y 's, we can include x as an explanatory variable in the measurement model. A model in which the same covariate affects both the y 's and a factor is not identified (see [4] for a demonstration of this for a single-level SEM).

An Example of Multilevel Structural Equation Modeling

We illustrate the use of multilevel SEM by applying the MIMIC model of (3) and (4) to the Hungarian

Table 3 Estimates from a two-level MIMIC model

Measurement model	$\lambda_r^{(1)}$ (SE)	σ_{er}^2 (SE)	σ_{ur}^2 (SE)
E. Sc. core	1 ^a	0.717 (0.023)	0.095 (0.019)
Biol. core	1.584 (0.095)	0.490 (0.021)	0.023 (0.011)
Biol. R3	0.628 (0.083)	0.892 (0.039)	0.033 (0.017)
Biol. R4	1.150 (0.100)	0.613 (0.030)	0.128 (0.028)
Phys. core	1.727 (0.109)	0.406 (0.021)	0.033 (0.012)
Phys. R2	1.491 (0.105)	0.537 (0.029)	0.053 (0.018)
Structural model	Estimate (SE)		
β_1	-0.133 (0.018)		
σ_e^2	0.117 (0.013)		
σ_u^2	0.077 (0.015)		

^aConstrained parameter.

science data. We consider one covariate, gender, which is coded 1 for girls and 0 for boys. The results are shown in Table 3. The parameter estimates for the measurement model are similar to those obtained for the factor model (Table 2). The results from the structural model show that girls have significantly lower overall science attainment (i.e., lower values on $\eta_{ij}^{(1)}$) than boys. A standardized coefficient may be calculated as $\beta_1^* = \beta_1/\sigma_e$. In our example $\hat{\beta}_1^* = -0.39$, which is interpreted as the difference in standard deviation units between girls' and boys' attainment, after adjusting for school effects (u_j). We may also compute the proportion of the residual variance in overall attainment that is due to differences between schools, which in this case is $0.077/(0.077 + 0.117) = 0.40$.

Estimation and Software

A multilevel factor model may be estimated in several ways using various software packages. A simple estimation procedure, described in [1], involves fitting a multivariate multilevel model to the responses to obtain estimates of the within-cluster and between-cluster covariance matrices, possibly adjusting for covariate effects. These matrices are then analyzed using any SEM software (see **Structural Equation Modeling: Software**). Muthén [6] describes another two-stage method, implemented in MPlus (www.statmodel.com), which involves analyzing the within-cluster and between-cluster covariances simultaneously using procedures for multigroup analysis. Alternatively, estimation may be carried out in a single step using Markov chain Monte Carlo (MCMC) methods (see [2]), in MLwiN (www.mlwin.com) or WinBUGS (www.mrc-bsu.cam.ac.uk/bugs/).

Multilevel structural equation models may also be estimated using Muthén's two-stage method. Simultaneous analysis of the within-covariance and between-covariance matrices allows cross-level constraints to be introduced. For example, as described above, the MIMIC model is a special case of a general factor model with factor loadings constrained to be equal across levels. For very general models, however, a one-stage approach may be required; examples include random coefficient models and models for mixed response types where multivariate normality cannot be assumed (see [7] for further discussion of

the limitations of two-stage procedures). For these and other general SEMs, Rabe-Hesketh et al. [7] propose **maximum likelihood estimation**, which has been implemented in gllamm (www.gllamm.org), a set of Stata programs. An alternative is to use Monte Carlo Markov Chain methods, which are available in WinBUGS (see **Markov Chain Monte Carlo and Bayesian Statistics**).

A common problem in the analysis of multivariate responses is **missing data**. For example, in the Hungarian study, responses are missing by design as each student was tested on the core booklet plus two randomly selected booklets from the remaining three. The estimation procedures and software for multilevel SEMs described above can handle missing data, under a 'missing at random' assumption.

In the analysis of the Hungarian science data, we used MLwiN to estimate the multilevel factor model and WinBUGS to estimate the multilevel MIMIC model. The parameter estimates and standard errors presented in Tables 2 and 3 are the means and standard errors from 20 000 samples, with a burn-in of 2000 samples.

Further Topics

We have illustrated the application and interpretation of simple multilevel factor models and SEMs in analyses of the Hungarian science data. In [2], extensions to more than one factor at the same level and correlated factors are considered. Other generalizations include allowing for additional levels, which may be hierarchical or cross-classified with another level, and random coefficients. For instance, in the science attainment example, schools may be nested in areas and the effect of gender on attainment may vary across schools and/or areas.

We have restricted our focus to models for multilevel continuous responses. In many applications, however, responses will be categorical or a mixture of different types. For a discussion of multilevel SEMs for binary, polytomous, or mixed responses, see [7] and **Structural Equation Modeling: Categorical Variables**.

We have not considered multilevel structures that arise from longitudinal studies, where the level one units are repeated measures nested within individuals at level two. (See **Multilevel and SEM Approaches to Growth Curve Modeling** for a discussion of multilevel SEMs for longitudinal data.)

References

- [1] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold, London.
- [2] Goldstein, H. & Browne, W.J. (2002). Multilevel factor analysis modelling using Markov Chain Monte Carlo (MCMC) estimation, in *Latent Variable and Latent Structure Models*, G. Marcoulides & I. Moustaki, eds, Lawrence Erlbaum, pp. 225–243.
- [3] McDonald, R.P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data, *British Journal of Mathematical and Statistical Psychology* **42**, 215–232.
- [4] Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* **69**, 337–357.
- [5] Muthén, B.O. (1989). Latent variable modelling in heterogeneous populations, *Psychometrika* **54**, 557–585.
- [6] Muthén, B.O. (1994). Multilevel covariance structure analysis, *Sociological Methods and Research* **22**, 376–398.
- [7] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). Generalized multilevel structural equation modelling, *Psychometrika* **69**, 183–206.

FIONA STEELE

Structural Equation Modeling: Nonstandard Cases

GEORGE A. MARCOULIDES

Volume 4, pp. 1931–1934

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Nonstandard Cases

A statistical model is only valid when certain assumptions are met. The assumptions of **structural equation models (SEM)** can be roughly divided into two types: structural and distributional [22]. Structural assumptions demand that no intended (observed or theoretical) variables are omitted from the model under consideration, and that no misspecifications are made in the equations underlying the proposed model. Distributional assumptions include linearity of relationships, completeness of data, multivariate normality (*see Multivariate Normality Tests*), and adequate sample size. With real data obtained under typical data gathering situations, violations of these distributional assumptions are often inevitable. So, two general questions can be asked about these distributional assumptions: (a) What are the consequences of violating them? (b) What strategies should be used to cope with them? In this article, each of these questions is addressed.

Linearity of Relationships

Not all relationships examined in the social and behavioral sciences are linear. Fortunately, various procedures are available to test such nonlinearities. Kenny and Judd [11] formulated the first nonlinear SEM model. Their approach used the product of observed variables to define **interaction effects** in **latent variables**. The equation $y = \alpha + \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1\xi_2 + \zeta$ was used to describe both the direct effects of ξ_1 and ξ_2 on y and the interactive effect $\xi_1\xi_2$. To model the interaction of ξ_1 and ξ_2 , multiplicative values of the interactive effect were created. Jöreskog and Yang [10] expanded the approach and illustrated how, even with more elaborate models, a one-product term variable is sufficient to identify all the parameters in the model. But even this approach is difficult to apply in practice due to the complicated nonlinear constraints that must be specified, and the need for large samples. If the interacting variable is discrete (e.g., gender), or can be made so by forming some data groupings, a multisample

approach can be easily applied (*see Factor Analysis: Multiple Groups*). Based on the multisample approach, the interaction effects become apparent as differences in the parameter estimates when the same model is applied to the grouped sets of data created. Jöreskog [8] recently introduced the latent variable score approach to test interaction effects. The factor score approach does not require the creation of product variables or the sorting of the data based on a categorization of the potential interacting variables. The approach can also be easily implemented using the PRELIS2 and SIMPLIS programs [9, 25] (*see Structural Equation Modeling: Software*). Various chapters in Schumacker and Marcoulides [26] discuss both the technical issues and the different methods of estimation available for dealing with nonlinear relationships.

Multivariate Normality

Four estimation methods are available in most SEM programs: Unweighted Least Squares (ULS), Generalized Least Squares (GLS), Asymptotically Distribution Free (ADF) – also called *Weighted Least Squares* (WLS) (*see Least Squares Estimation*), and **Maximum Likelihood (ML)**. ML and GLS are used when data are normally distributed. The simplest way to examine normality is to consider univariate **skewness** and **kurtosis**. A measure of multivariate skewness and kurtosis is called *Mardia's coefficient* and its normalized estimate [4]. For normally distributed data, Mardia's coefficient will be close to zero, and its normalized estimate nonsignificant. Another method for judging bivariate normality is based on a plot of the χ^2 percentiles and the mean distance measure of individual observations. If the distribution is normal, the plot of the χ^2 percentiles and the mean distance measure should resemble a straight line [13, 19].

Research has shown that the ML and GLS methods can be used even with minor deviations from normality [20] – the parameter estimates generally remain valid, although the standard errors may not. With more serious deviations the ADF method can be used as long as the sample size is large. With smaller sample sizes, the Satorra-Bentler robust method of parameter estimation (a special type of ADF method) should be used.

Another alternative to handling nonnormal data is to make the data more 'normal-looking' by applying a transformation on the raw data. Numerous

transformation methods have been proposed in the literature. The most popular are square root transformations, power transformations, reciprocal transformations, and logarithmic transformations.

The presence of categorical variables may also cause nonnormality. Muthén [14] developed a categorical and continuous variable methodology (implemented in *Mplus* (see **Structural Equation Modeling: Software**) [16]), which basically permits the analysis of any combination of dichotomous, ordered polytomous, and measured variables. With data stemming from designs with only a few possible response categories (e.g., ‘Very Satisfied’, ‘Somewhat Satisfied’, and ‘Not Satisfied’), the ADF method can also be used with polychoric (for assessing the degree of association between ordinal variables) or polyserial (for assessing the degree of association between an ordinal variable and a continuous variable) correlations. Ignoring the categorical attributes of data obtained from such items can lead to biased results. Fortunately, research has shown that when there are five or more response categories (and the distribution of data is normal) the problems from disregarding the categorical nature of responses are likely to be minimized.

Missing Data

Data sets with missing values are commonly encountered. **Missing data** are often dealt with by ad hoc procedures (e.g., listwise deletion, pairwise deletion, mean substitution) that have no theoretical justification, and can lead to biased estimates. But there are a number of alternative theory-based procedures that offer a wide range of good data analytic options. In particular, three ML estimation algorithms are available: (a) the multigroup approach [1, 15], which can be implemented in most existing SEM software; (b) full information maximum likelihood (FIML) estimation, which is available in LISREL [9], EQS6 [5], AMOS [2], *Mplus* [16], and *Mx* [18]; and (c) the Expectation Maximization (EM) algorithm, which is available in SPSS Missing Values Analysis, EMCOV [6], and NORM [24]. All the available procedures assume that the missing values are either missing completely at random (MCAR) or missing at random (MAR) (see **Dropouts in Longitudinal Data; Dropouts in Longitudinal Studies: Methods of Analysis**). Under MCAR, the probability of a missing response is independent of both

the observed and unobserved data. Under MAR, the probability of a missing response depends only on the observed responses. The MCAR and MAR conditions are also often described as ignorable nonresponses. But sometimes respondents and nonrespondents with the same observed data values differ systematically with respect to the missing values for nonrespondents. Such cases are often called *nonignorable nonresponses* or data missing not at random (MNAR), and, although there is usually no correction method available, sometimes, a case-specific general modeling approach might work.

To date, the FIML method has been shown to yield unbiased estimates under both MCAR and MAR scenarios, including data with mild deviations from multivariate normality. This suggests that FIML can be used in a variety of empirical settings with missing data. **Multiple imputation** is also becoming a popular method to deal with missing data. Multiple data sets are created with plausible values replacing the missing data, and a complete data set is subsequently analyzed [7]. For dealing with nonnormal missing data, the likelihood-based approaches developed by Arminger and Sobel [3], and Yuan and Bentler [27] may work better.

Outliers

Real data sets almost always include **outliers**. Sometimes, outliers are harmless and do not change the results, regardless of whether they are included or deleted from the analyses. But, sometimes, they have much influence on the results, particularly on parameter estimates. Assuming outliers can be correctly identified, deleting them is often the preferred approach. Another approach is to fit the model to the data without the outliers and inspect the results to examine the impact of including them in the analysis.

Power and Sample Size Requirements

A common modeling concern involves using the appropriate sample size. Although various sample size rules-of-thumb have been proposed in the literature (e.g., 5–10 observations per parameter, 50 observations per variable, 10 times the number of free model parameters), no one rule can be applied to all situations encountered. This is because the sample size needed for a study depends on many factors

including the complexity of the model, distribution of the variables, amount of missing data, reliability of the variables, and strength of the relationships among the variables. Standard errors are also particularly sensitive to sample size issues. For example, if the standard errors in a proposed model are overestimated, significant effects can easily be missed. In contrast, if standard errors are underestimated, significant effects can be overemphasized. As a consequence, proactive Monte Carlo analyses should be used to help determine the sample size needed to achieve accurate Type I error control and parameter estimation precision. The methods introduced by Satorra and Saris [21, 23], and MacCallum, Brown, and Sugawara [12] can be used to assess the sample size in terms of power of the goodness of fit of the model. Recently, Muthén and Muthén [17] illustrated how *Mplus* can be used to conduct a Monte Carlo study to help decide on sample size and determine power.

References

- [1] Allison, P.D. (1987). Estimation of linear models with incomplete data, in *Sociological Methodology*, C.C. Clogg, ed., Jossey-Bass, San Francisco pp. 71–103.
- [2] Arbuckle, J.L. & Wothke, W. (1999). *AMOS 4.0 User's Guide*, Smallwaters, Chicago.
- [3] Arminger, G. & Sober, M.E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data, *Journal of the American Statistical Association* **85**, 195–2003.
- [4] Bentler, P.M. (1995). *EQS Structural Equations Program Manual*, Multivariate Software, Encino.
- [5] Bentler, P.M. (2002). *EQS6 Structural Equations Program Manual*, Multivariate Software, Encino.
- [6] Graham, J.W. & Hofer, S.M. (1993). *EMCOV.EXE Users Guide*, Unpublished manuscript, University of Southern California.
- [7] Graham, J.W. & Hofer, S.M. (2000). Multiple imputation in multivariate research, in *Modeling Longitudinal and Multilevel Data*, T. Little, K.U. Schnabel & J. Baumert, eds., Lawrence Erlbaum Associates, Mahwah.
- [8] Jöreskog, K.G. (2000). *Latent Variable Scores and their Uses*, Scientific Software International, Inc, Lincolnwood, (Acrobat PDF file: <http://www.sscentral.com/>).
- [9] Jöreskog, K.G., Sörbom, D., Du Toit, S. & Du Toit, M. (1999). *LISREL 8: New Statistical Features*, Scientific Software International, Chicago.
- [10] Jöreskog, K.G. & Yang, F. (1996). Nonlinear structural equation models: the Kenny-Judd model with interaction effects, in *Advanced Structural Equation Modeling: Issues and Techniques*, G.A. Marcoulides & R.E. Schumacker, eds., Lawrence Erlbaum Associates, Mahwah, pp. 57–89.
- [11] Kenny, D.A. & Judd, C.M. (1984). Estimating the non-linear and interactive effects of latent variables, *Psychological Bulletin* **96**, 201–210.
- [12] MacCallum, R.C., Browne, M.W. & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure models, *Psychological Methods* **1**, 130–149.
- [13] Marcoulides, G.A. & Hershberger, S.L. (1997). *Multivariate Statistical Methods: A First Course*, Lawrence Erlbaum Associates, Mahwah.
- [14] Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.
- [15] Muthén, B.O., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random, *Psychometrika* **52**, 431–462.
- [16] Muthén, L.K. & Muthén, B.O. (2002a). *Mplus User's Guide*, Muthén & Muthén, Los Angeles.
- [17] Muthén, L.K. & Muthén, B.O. (2002b). How to use a Monte Carlo study to decide on sample size and determine power, *Structural Equation Modeling* **9**, 599–620.
- [18] Neale, M.C. (1994). *Mx: Statistical Modeling*, Department of Psychiatry, Medical College of Virginia, Richmond.
- [19] Raykov, T. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.
- [20] Raykov, T. & Widaman, K.F. (1995). Issues in applied structural equation modeling research, *Structural Equation Modeling* **2**, 289–318.
- [21] Saris, W.E. & Satorra, A. (1993). Power evaluations in structural equation models, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage Publications, Newbury Park pp. 181–204.
- [22] Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments, *Quality & Quantity* **24**, 367–386.
- [23] Satorra, A. & Saris, W.E. (1985). The power of the likelihood ratio test in covariance structure analysis, *Psychometrika* **50**, 83–90.
- [24] Schafer, J.L. (1998). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.
- [25] Schumacker, R.E. (2002). Latent variable interaction modeling, *Structural Equation Modeling* **9**, 40–54.
- [26] Schumacker, R.E. & Marcoulides, G.A., eds (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.
- [27] Yuan, K.H. & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data, *Sociological Methodology* **30**, 165–200.

GEORGE A. MARCOULIDES

Structural Equation Modeling: Nontraditional Alternatives

EDWARD E. RIGDON

Volume 4, pp. 1934–1941

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Nontraditional Alternatives

Introduction

Within the broad family of multivariate analytical techniques (*see* **Multivariate Analysis: Overview**), the boundaries that associate similar techniques or separate nominally different techniques are not always clear. **Structural equation modeling** (SEM) can be considered a generalization of a wide class of multivariate modeling techniques. Nevertheless, there are a variety of techniques which, in comparison to conventional SEM, are similar enough on some dimensions but different enough in procedure to be labeled variant SEM methods. (Not all of the creators of these methods would appreciate this characterization.) Each of these methods addresses one or more of the problems which can stymie or complicate conventional SEM analysis, so SEM users ought to be familiar with these alternative techniques. The aim here is to briefly introduce some of these techniques and give SEM users some sense of when these methods might be employed instead of, or in addition to, conventional SEM techniques.

Partial Least Squares

Partial Least Squares (PLS) is now the name for a family of related methods. For the most part, the essential distinction between these methods and conventional SEM is the same as the distinction between principal component analysis and factor analysis. Indeed, PLS can be thought of as ‘a constrained form of component modeling’ [5], as opposed to constrained modeling of common factors. The strengths and weaknesses of PLS versus SEM largely follow from this distinction.

Origin

PLS was invented by Herman Wold, in the 1960s, under the inspiration of Karl Jöreskog’s pioneering work in SEM. Wold, a renowned econometrician (he invented the term, ‘recursive model’, for example)

was Jöreskog’s mentor. Wold’s intent was to develop the same kind of structural analysis technique but starting from the basis of principal component analysis. In particular, reacting to SEM’s requirements for large samples, multivariate normality, and substantial prior theory development, Wold aimed to develop a structural modeling technique that was compatible with small sample size, arbitrary distributions, and what he termed *weak theory*. Like many terms associated with PLS, the meaning of ‘weak theory’ is not precisely clear. McDonald [12] strongly criticized the methodology for the *ad hoc* nature of its methods.

Goals

Unlike conventional SEM, PLS does not aim to test a model in the sense of evaluating discrepancies between empirical and model-implied covariance matrices. Eschewing assumptions about data distributions or even about sample size, PLS does not produce an overall test statistic like conventional SEM’s χ^2 . Instead, the stated aim of PLS is to maximize prediction of (that is, to minimize the unexplained portion of) dependent variables, especially dependent observed variables. Indeed, PLS analysis will show smaller residual variances, but larger residual covariances, than conventional SEM analysis of the same data. PLS analysis ‘develops by a dialog between the investigator and the computer’ [20] – users estimate models and make revisions until they are satisfied with the result.

Estimation of model parameters is purely a byproduct of this optimization effort. Unlike conventional SEM, PLS parameter estimates are not generally consistent, in the sense of converging on population values as sample size approaches infinity. Instead, PLS claims the property of ‘consistency at large’ [20]. PLS parameter estimates converge on population values as both sample size and the number of observed variables per construct both approach infinity. In application, PLS tends to overestimate structural parameters (known here as ‘inner relations’) and underestimate measurement parameters (known here as ‘outer relations’) [20]. The literature argues that this is a reasonable trade, giving up the ability to accurately estimate essentially hypothetical model parameters [7] for an improved ability to predict the data; in other words, it is a focus on the ends rather than the means.

Inputs

At a minimum, Partial Least Squares can be conducted on the basis of a correlation matrix and a structure for relations between the variables. Ideally, the user will have the raw data. Since PLS does not rely on distributional assumptions and asymptotic properties, standard errors for model parameters are estimated via **jackknifing** (randomly omitting data points and reestimating model parameters, and observing the empirical variation in the parameter estimates), while model quality is evaluated, in part, through a blindfolding technique (repeatedly estimating the model parameters with random data points omitted, each time using those parameter estimates to predict the values of the missing data points, and observing the accuracy of these estimates) known as the Stone–Geisser test [7]. Wold [20] specified that the data could be ‘scalar, ordinal, or interval’.

Because PLS proceeds in a stepwise fashion through a series of regressions, rather than attempting to estimate all parameters simultaneously, and because it does not rely on distributional assumptions, sample size requirements for PLS are said to be substantially lower than those for conventional SEM [3], but there is little specific guidance. On the one hand, Wold [20] asserted that PLS comes into its own in situations that are ‘data-rich but theory-primitive’, and the ‘consistency at large’ property suggests that larger sample size will improve results. Chin [3], borrowing a regression heuristic, suggests finding the larger of either (a) the largest number of arrows from observed variables pointing to any one block variable (see below), or (b) the largest number of arrows from other block variables pointing to any one block variable, and multiplying by 10.

While PLS does not focus on model testing in the same way as conventional SEM, PLS users are still required to specify an initial model or structure for the variables. PLS is explicitly not a model-finding tool like exploratory factor analysis [5] (see **Factor Analysis: Exploratory**). Each observed variable is uniquely associated with one block variable. These block variables serve the same role in a PLS model as common factor **latent variables** do in SEM. However, block variables in PLS are not latent variables [12] – they are weighted composites of the associated observed variables, and hence observable themselves. This means that the block variables share in all aspects of the observed variables, including random error, hence the bias in parameter estimation as

compared to conventional SEM. But it also means that the PLS user can always assign an unambiguous score for each block variable for each case. In SEM, because the latent variables are not composites of the observed variables, empirical ‘factor scores’ (expressions of a latent variable in terms of the observed variables) can never be more than approximate, nor is there one clearly best method for deriving these scores.

Besides specifying which observed variables are associated with which block, the user must also specify how each set of observed variables is associated with the block variable. The user can choose to regress the observed variables on the block variable or to regress the block variable on the observed variables. If the first choice is made for all blocks, this is known as ‘Mode A’. If the second choice is made for all blocks, this is known as ‘Mode B’, while making different choices for different blocks is known as ‘Mode C’ (see Figure 1). Users cannot make different choices for different variables within a block, but it is expected that users will try estimating the model with different specifications in search of a satisfactory outcome.

The best-known implementations of PLS are limited to recursive structural models. Relations between block variables may not include reciprocal relations, feedback loops, or correlated errors between block

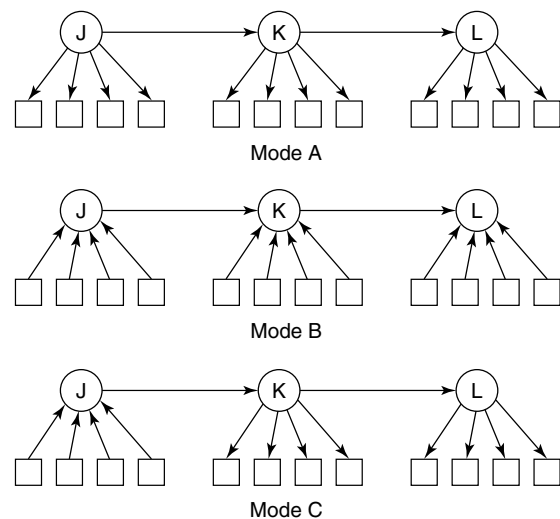


Figure 1 Three modes for estimating relations between observed variables and block variables in partial least squares

variables. Hui [8] developed a procedure and program for PLS modeling of nonrecursive models (*see Recursive Models*), but it seems to have been little used.

Execution

PLS parameter estimation proceeds iteratively. Each iteration involves three main steps [20]. The first step establishes or updates the value of each block variable as a weighted sum of the measures in the block. Weights are standardized to yield a block variable with a variance of 1. The second step updates estimates of the ‘inner relations’ and ‘outer relations’ parameters. The ‘inner relations’ path weights are updated through least squares regressions, as indicated by the user’s model. For the ‘outer relations’ part of this step, each block variable is replaced by a weighted sum of all other block variables to which it is directly connected. Wold [20] specified that these sums were weighted simply by the sign of the correlation between the two blocks – hence each block variable is replaced by a ‘sign-weighted sum’. For example, the *J* block variable in Figure 1 would be replaced by the *K* block variable, weighted by the sign of the correlation between *J* and *K*. The *K* block variable would be replaced by a sum of the *J* and *L* block variables, each weighted by the signs of the (*J*, *K*) and (*K*, *L*) correlations. From here, the procedure for estimating the ‘outer relations’ weights depends on the choice of mode. For Mode A blocks, each observed variable in a given block is regressed on the sign-weighted composite for its block, in a set of independent bivariate regressions (*see Multivariate Multiple Regression*). For Mode B blocks, the sign-weighted composite is regressed on all the observed variables in the block, in one multiple regression. Then the next iteration begins by once again computing the values of the block variables as weighted sums of their observed variables. Estimation iterates until the change in values becomes smaller than some convergence criterion. Chin [3] notes that while different PLS incarnations have used slightly different weighting schemes, the impact of those differences has never been substantial.

Output and Evaluation

PLS produces estimates of path weights, plus R^2 calculations for dependent variables and a block

correlation matrix. For Mode A blocks, PLS also produces loadings, which are approximations to the loadings one would derive from a true factor model. Given raw data, PLS will also produce jackknifed standard errors and a value for the Stone–Geisser test of ‘predictive relevance’.

Falk and Miller [5] cautiously offer a number of rules of thumb for evaluating the quality of the model resulting from a PLS analysis. At the end, after possible deletions, there should still be three measures per block. Loadings, where present, should be greater than 0.55, so that the communalities (loadings squared) should be greater than 0.30. Every predictor should explain at least 1.5% of the variance of every variable that it predicts. R^2 values should be above 0.10, and the average R^2 values across all dependent block variables should be much higher.

PLS Variants and PLS Software

Again, PLS describes a family of techniques rather than just one. Each step of the PLS framework is simple in concept and execution, so the approach invites refinement and experimentation. For example, Svante Wold, Herman Wold’s son, applied a form of PLS to chemometrics (loosely, the application of statistics to chemistry), as a tool for understanding the components of physical substances. As a result, the Proc PLS procedure currently included in the SAS package is designed for this chemometrics form of PLS, rather than for Herman Wold’s PLS. There are undoubtedly many proprietary variations of the algorithm and many proprietary software packages being used commercially. The most widely used PLS software must surely be Lohmöller’s [11] LVPLS. The program is not especially user-friendly, as it has not benefited from regular updates. Lohmöller’s program and a user manual are distributed, free, by the Jefferson Psychometrics Laboratory at the University of Virginia (<http://kiptron.psyc.virginia.edu/disclaimer.html>). Also in circulation is a beta version of Chin’s PLSGraph [4], a PLS program with a graphical user interface.

Tetrad

In many ways, Tetrad represents the polar opposite of PLS. There is a single Tetrad methodology, although that methodology incorporates many tools

and algorithms. The rationale behind the methodology is fully specified and logically impeccable, and the methodology aims for optimality in well-understood terms. Going beyond PLS, which was designed to enable analysis in cases where theory is weak, the creators of Tetrad regard theory as largely irrelevant: ‘In the social sciences, there is a great deal of talk about the importance of “theory” in constructing causal explanations of bodies of data. . . . In many of these cases the necessity of theory is badly exaggerated.’ [17]. Tetrad is a tool for searching out plausible causal inferences from correlational data, within constraints. Tetrad and related tools have found increasing application by organizations facing a surplus of data and a limited supply of analysts. These tools and their application fall into a field known as ‘knowledge discovery in databases’ or ‘data mining’ [6]. Academic researchers may turn to such tools when they gain access to secondary commercial data and want to learn about structures underlying that data. Extensive online resources are available from the Tetrad Project at Carnegie Mellon University (<http://www.phil.cmu.edu/projects/tetrad/>) and from Pearl (<http://bayes.cs.ucla.edu/jp-home.html> and <http://bayes.cs.ucla.edu/BOOK-2K/index.html>).

Origin

Tetrad was developed primarily by Peter Spirtes, Clark Glymour, and Richard Scheines [17], a group of philosophers at Carnegie-Mellon University who wanted to understand how human beings develop causal reasoning about events. Prevailing philosophy lays out rigorous conditions which must be met before one can defensibly infer that A causes B. Human beings make causal inferences every day without bothering to check these conditions – sometimes erroneously, but often correctly. From this basis, the philosophers moved on to study the conditions under which it was possible to make sound causal inferences, and the kinds of procedures that would tend to produce the most plausible causal inferences from a given body of data. Their research, in parallel with contributions by Judea Pearl [14] and others, produced algorithms which codified procedures for quickly and automatically uncovering possible causal structures that are consistent with a given data set. The Tetrad program gives researchers access to these algorithms.

Goals

As noted above, the explicit aim of Tetrad is to uncover plausible inferences about the causal structure that defines relationships among a set of variables. Tetrad’s creators deal with causality in a straightforward way. By contrast, conventional SEM deals gingerly with causal claims, having hardly recovered from Ling’s [10] caustic review of Kenny’s [9] early SEM text, *Correlation and Causality*. Today, any mention of the phrase, ‘causal modeling’ (an early, alternate name for structural equation modeling), is almost sure to be followed by a disclaimer about the near impossibility of making causal inferences from correlational data alone. SEM users, as well as the researchers associated with the Tetrad program, are well aware of the typically nonexperimental nature of their data, and of the various potential threats to the validity of causal inference from such data. Nevertheless, the aim of Tetrad is to determine which causal inferences are consistent with a given data set. As Pearl [14] notes, however, the defining attribute of the resulting inferences is not ‘truth’ but ‘plausibility’: the approach ‘identifies the mechanisms we can plausibly infer from non-experimental data; moreover, it guarantees that any alternative mechanism will be less trustworthy than the one inferred because the alternative would require more contrived, hindsightful adjustment of parameters (i.e., functions) to fit the data’.

Inputs

Tetrad proceeds by analysis of an empirical correlation matrix. Sampling is a key issue for researchers in this area. Both [14] and [18] devote considerable attention to the problem of heterogeneity – of a given data set including representative of different populations. Even if all of the represented populations conform to the same structural equation model, but with different population values for some parameters, the combined data set may fit poorly to that same model, with the parameters freely estimated [13]. This phenomenon, discussed in this literature under the title, ‘Simpson’s Paradox’ (see **Paradoxes**), explains why researchers must take care to ensure that data are sampled only from homogeneous populations. As noted below, Tetrad does not actually estimate model parameters, so sample size need only be large enough to stably estimate the correlation matrix. Still, larger samples do improve precision.

Unlike conventional SEM, however, users of Tetrad are not expected to have a full theoretical model, or any prior knowledge about the causal structure. Still, Tetrad does exploit prior knowledge. Researchers can impose constraints on the relations between variables – requiring either that a certain relation must exist, or that it must not exist – whether those constraints arise from broad theory or from logical considerations related to such factors as time order. Thus, one might allow ‘parent’s occupation’ to have a causal effect on ‘child’s occupation’, but might disallow a relationship in the opposite direction. Tetrad does not accommodate nonrecursive relationships – the authors view such structures as incompatible with common sense notions of cause and effect.

Far more important than broad ‘theory’, for the Tetrad user, are logical analysis and the set of assumptions which they are willing to make. One key assumption involves ‘causal sufficiency’ [17]. A set of observed variables is causally sufficient if the causal structure of those variables can be explained purely by relations among those variables themselves. If a set of observed variables are causally sufficient, then there literally are no latent or unobserved variables involved in a Tetrad analysis. If the researcher does not believe that the set of variables is causally sufficient, they can employ Tetrad algorithms that search for latent variables – for variables outside the data set that explain relations between the observed variables.

Thus, besides providing the correlation matrix, the researcher faces two key choices. First, will they assume causal sufficiency, or not? The assumption greatly simplifies and speeds execution, but a false assumption here invalidates the analysis. In addition, the researcher must select an alpha or Type I error probability. Two key tools for model selection are the tetrad and the partial correlation. A tetrad is a function of the covariances or correlations of four variables:

$$\tau_{ABCD} = \sigma_{AB}\sigma_{CD} - \sigma_{AC}\sigma_{BD} \quad (1)$$

A tetrad that is equal to zero is called a *vanishing tetrad*. Different model structures may or may not imply different sets of vanishing tetrads. For example, if four observed variables are all reflective measures of the same common factor (see Figure 2), then all tetrads involving the four factors vanish. But even if one of the four variables actually measures a different factor, all tetrads still vanish [17].

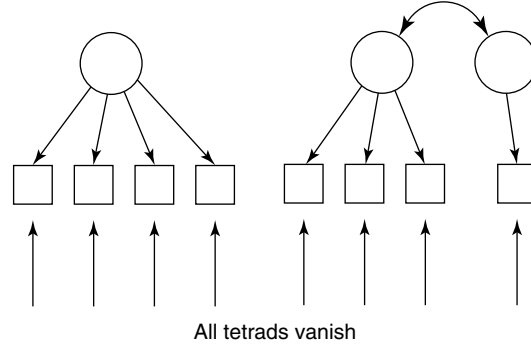


Figure 2 Two measurement models with different substantive implications, yet both implying that all tetrads vanish

The partial correlation of variables B and C , given A , is:

$$\rho_{BC.A} = \frac{\rho_{BC} - \rho_{AB} \times \rho_{AC}}{\sqrt{1 - \rho_{AB}^2} \sqrt{1 - \rho_{AC}^2}} \quad (2)$$

A zero partial correlation suggests no direct relationship between two variables. For example, for three mutually correlated variables A , B , and C , $\rho_{BC.A} = 0$ suggests either that A is direct or indirect predictor of both B and C or that A mediates the relationship between B and C . By contrast, if A and B predict C , then $\rho_{AB.C} \neq 0$, even if A and B are uncorrelated [17].

When the program is determining whether tetrads and partial correlations are equal to 0, it must take account of random sampling error. A smaller value for alpha, such as .05, will mean wider confidence intervals when Tetrad is determining whether a certain quantity is equal to 0. As a result, the program will identify more potential restrictions which can be imposed, and will produce a more parsimonious and determinate model. A larger value, such as .10 or greater, will lead to fewer ‘nonzero’ results, producing a less determinate and more saturated model, reflecting greater uncertainty for any given sample size.

Execution

Tetrad’s method of analysis is based on a rigorous logic which is detailed at length in [14] and [17], and elsewhere. In essence, these authors argue that if a certain causal structure actually underlies a data

set, then the vanishing (and nonvanishing) tetrads and zero (and nonzero) partial correlations and other correlational patterns which are logically implied by that structure will be present in the data, within random sampling error. Only those patterns that are implied by the true underlying structure should be present. Certainly, there may be multiple causal structures that are consistent with a certain data set, but some of these may be ruled out by the constraints imposed by the researcher. (The Tetrad program is designed to identify all causal structures that are compatible with both the data and the prior constraints.) Therefore, if a researcher determines which correlational patterns are present, and which causal structures are consistent with the empirical evidence and with prior constraints, then, under assumptions, the most plausible causal inference is that those compatible causal structures are, in fact, the structures that underlie the data.

Tetrad uses algorithms to comb through all of the available information – correlations, partial correlations, tetrads, and prior constraints – in its search for plausible causal structures. These algorithms have emerged from the authors' extensive research. Different algorithms are employed if the researcher does not embrace the causal sufficiency assumption. Some of these algorithms begin with an unordered set of variables, where the initial assumption is that the set of variables are merely correlated, while others require a prior ordering of variables. The ideal end-state is a model of relations between the variables that is fully directed, where the observed data are explained entirely by the causal effects of some variables upon other variables. In any given case, however, it may not be possible to achieve such a result given the limitations of the data and the prior constraints. The algorithms proceed by changing mere correlations into causal paths, or by deleting direct relationships between pairs of variables, until the system is as fully ordered as possible.

Outputs

The chief output of Tetrad analysis is information about what sorts of constraints on relations between variables can be imposed based on the data. Put another way, Tetrad identifies those permitted causal structures that are most plausible. If causal sufficiency is not assumed, Tetrad also indicates what sorts of latent variables are implied by its analysis, and which

observed variables are affected. Tetrad also identifies situations where it cannot resolve the direction of causal influence between two variables.

Unlike conventional SEM or even PLS, Tetrad produces no parameter estimates at all. Computation of tetrads and partial correlations does not require parameter estimates, so Tetrad does not require them. Users who want parameter estimates, standard errors, and fit indices, and so forth might estimate the model(s) recommended by Tetrad analysis using some conventional SEM package. By contrast, researchers who encounter poor fit when testing a model using a conventional SEM package might consider using Tetrad to determine which structures are actually consistent with their data. Simulation studies [18] suggest researchers are more likely to recover a true causal structure by using Tetrad's search algorithms, which may return multiple plausible models, than they will by using the model modification tools in conventional SEM software.

Confirmatory Tetrad Analysis

While Partial Least Squares and Tetrad are probably the two best-known methodologies in this class, others are certainly worth mentioning. Bollen and Ting ([1, 2, 19]) describe a technique called *confirmatory tetrad analysis* (CTA). Unlike Tetrad, CTA aims to test whether a prespecified model is consistent with a data set. The test proceeds by determining, from the structure of the model, which tetrads ought to vanish, and then empirically determining whether or not those tetrads are actually zero. As compared to SEM, CTA has two special virtues. First, as noted above, tetrads are computed directly from the correlation matrix, without estimating model parameters. This is a virtue in situations where the model in question is not statistically identified [1]. Being not identified means that there is no one unique best set of parameter estimates. SEM users have the freedom to specify models that are not statistically identified, but the optimization procedures in conventional SEM packages generally fail when they encounter such a model. Identification is often a problem for models that include causal or formative indicators [2], for example. Such indicators are observed variables that are predictors of, rather than being predicted by, latent variables, as in Mode B estimation in PLS (see Figure 1). In conventional SEM, when a latent

variable has only formative indicators, identification problems are quite likely. For such a model, CTA can formally test the model, by way of a χ^2 test statistic [1].

A second virtue of CTA is that two competing structural equation models which are not nested in the conventional sense may be nested in terms of the vanishing tetrads that they each imply [1]. When two models are nested in conventional SEM terms, it means, in essence, that the free parameters of one model are a strict subset of those in the other model. Competing nested models can be evaluated using a χ^2 difference test, but if the models are not nested, then the χ^2 difference test cannot be interpreted. Other model comparison procedures exist within conventional SEM, but they do not readily support hypothesis testing. If the vanishing tetrads implied by one model are a strict subset of the vanishing tetrads implied by the other, then the two models are nested in tetrad terms, and a comparison of such models yields a χ^2 difference test. Models may be nested in tetrad terms even if they are not nested in conventional SEM terms, so conventional SEM users may find value in CTA, in such cases. In addition, avoiding parameter estimation may enable model testing at lower sample sizes than are required by conventional SEM.

The largest hurdle for CTA appears to be the problem of redundant tetrads ([1], [19]). For example, if $\tau_{ABCD} = \rho_{AB}\rho_{CD} - \rho_{AC}\rho_{BD} = 0$, then it necessarily follows that $\tau_{ACBD} = \rho_{AC}\rho_{BD} - \rho_{AB}\rho_{CD}$ is also 0. Thus, CTA involves determining not only which vanishing tetrads are implied by a model, but also which vanishing tetrads are redundant. Testing the model, or comparing the competing models, proceeds from that point.

HyBlock

William Rozeboom's HyBall ([15, 16]) is a comprehensive and flexible package for conducting exploratory factor analysis (EFA). One module in this package, known as HyBlock, performs a kind of structured exploratory factor analysis that may be a useful alternative to the sparse measurement models of conventional SEM. In using HyBlock, a researcher must first allocate the variables in a data set into blocks of related variables. The user must also specify how many common factors are to be extracted

from each block and the structural linkages between blocks. Like PLS and Tetrad, HyBlock is limited to recursive structural models. HyBlock conducts the blockwise factor analysis and estimates the between-block structural model, generating both parameter estimates and familiar EFA diagnostics. Thus, one might view HyBlock as an alternative to PLS that is firmly grounded in factor analysis. Alternatively, one might view HyBlock as an alternative to conventional SEM for researchers who believe that their measures are factorially complex.

Conclusion

Conventional SEM is distinguished by large-sample empirical testing of a prespecified, theory-based model involving observed variables which are linked to latent variables through a sparse measurement model. Methodologies which vary from these design criteria will continue to have a place in the multivariate toolkit.

References

- [1] Bollen, K.A. & Ting, K.-F. (1993). Confirmatory tetrad analysis, in *Sociological Methodology*, P.V. Marsden, ed., American Sociological Association, Washington, 147–175.
- [2] Bollen, K.A. & Ting, K.-F. (2000). A tetrad test for causal indicators, *Psychological Methods* 5, 3–22.
- [3] Chin, W. (1998). The partial least squares approach for structural equation modeling, in G.A. Marcoulides, ed., *Modern Business Research Methods*, Lawrence Erlbaum Associates, Mahwah, 295–336.
- [4] Chin, W. (2001). *PLS-Graph User's Guide, Version 3.0*, Houston: Soft Modeling, Inc.
- [5] Falk, R.F. & Miller, N.B. (1992). *A Primer for Soft Modeling*, University of Akron, Akron.
- [6] Frawley, W.J., Piatetsky-Shapiro, G. & Matheus, C.J. (1991). Knowledge discovery in databases: an overview, in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro & W.J. Frawley, eds, AAAI Press, Menlo Park, pp. 1–27.
- [7] Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* 70, 320–328.
- [8] Hui, B.S. (1982). On building partial least squares with interdependent inner relations, in K.G. Joreskog & H. Wold, eds, *Systems Under Indirect Observation, Part II*, North-Holland, 249–272.
- [9] Kenny, D.A. (1979). *Correlation and Causality*, Wiley, New York.

8 Structural Equation Modeling: Nontraditional Alternatives

- [10] Ling, R.F. (1982). Review of 'Correlation and Causation [sic]', *Journal of the American Statistical Association* **77**, 489–491.
- [11] Lohmöller, J.-B. (1984). *LVPS Program Manual: Latent Variables Path Analysis with Partial Least Squares Estimation*, Zentralarchiv für empirische Sozialforschung, Universität zu Köln, Köln.
- [12] McDonald, R.P. (1996). Path analysis with composite variables, *Multivariate Behavioral Research* **31**, 239–270.
- [13] Muthén, B. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika* **54**, 557–585.
- [14] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.
- [15] Rozeboom, W.W. (1991). HYBALL: a method for subspace-constrained oblique factor rotation, *Multivariate Behavioral Research* **26**, 163–177.
- [16] Rozeboom, W.W. (1998). *HYBLOCK: A routine for exploratory factoring of block-structured data. Working paper*, University of Alberta, Edmonton.
- [17] Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction and Search*, 2nd Edition, MIT Press, Cambridge.
- [18] Spirtes, P., Scheines, R. & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL programs, *Sociological Methods & Research* **19**, 3–66.
- [19] Ting, K.-F. (1995). Confirmatory tetrad analysis in SAS, *Structural Equation Modeling* **2**, 163–171.
- [20] Wold, H. (1985). Partial least squares, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 581–591.

EDWARD E. RIGDON

Structural Equation Modeling: Overview

RANDALL E. SCHUMACKER

Volume 4, pp. 1941–1947

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Overview

Structural equation modeling (SEM) has been historically referred to as linear structural relationships, **covariance structure analysis**, or latent variable modeling. SEM has traditionally tested hypothesized theoretical models that incorporate a correlation methodology with correction for unreliability of measurement in the observed variables. SEM models have currently included most statistical applications using either observed variables and/or **latent variables**, for example, **multiple linear regression**, **path analysis**, **factor analysis**, latent growth curves (see **Structural Equation Modeling: Latent Growth Curve Analysis**), multilevel, and interaction models (see **Generalized Linear Mixed Models**). Six basic steps are involved in structural equation modeling: model specification, model identification, model estimation, model testing, model modification, and model validation.

Model Specification

A researcher uses all relevant theory and related research to develop a theoretical model, that is, specifies a theoretical model. A theoretical model establishes how latent variables are related. The researcher wants the specific theoretical model to be confirmed by the observed sample variance–covariance data. A researcher must decide which observed variables to include in the theoretical model and how these observed variables measure latent variables. *Model specification* implies that the researcher specifies observed and latent variable relationships in a theoretical model and designates which parameters in the model are important. A model is properly specified when the true population model is consistent with the theoretical model being tested, that is, the sample variance–covariance matrix is sufficiently reproduced by the theoretical model. The goal of SEM is to determine whether the theoretical model generated the sample variance–covariance matrix. The sample variance–covariance matrix therefore implies some underlying theoretical model (covariance structure).

The researcher can determine the extent to which the theoretical model reproduces the sample variance–covariance matrix. The theoretical model produces an implied variance–covariance matrix whose elements are subtracted from the original sample variance–covariance matrix to produce a residual variance–covariance matrix. If the residual values are larger than expected, then the theoretical model is *misspecified*. The theoretical model misspecification can be due to errors in either not including an important variable or in including an unimportant variable. A misspecified theoretical model results in bias parameter estimates for variables that may be different from the true population model. This bias is known as *specification error* and indicates that the theoretical model may not fit the sample variance–covariance data and be statistically acceptable.

Model Identification

A researcher must resolve the model identification problem prior to the estimation of parameters for observed and latent variables in the theoretical model. *Model identification* is whether a unique set of parameter estimates can be computed, given the theoretical model and sample variance–covariance data. If many different parameter estimates are possible for the theoretical model, then the model is not identified, that is, indeterminacy exists in the model. The sample variance–covariance data may also fit more than one theoretical model equally well. Model indeterminacy occurs when there are not enough constraints in the theoretical model to obtain unique parameter estimates for the variables. Model identification problems are solved by imposing additional constraints in the theoretical model, for example, specifying latent variable variance.

Observed and latent variable parameters in a theoretical model must be specified as a *free* parameter, a *fixed* parameter, or a *constrained* parameter. A *free* parameter is a parameter that is unknown and a researcher wants to estimate it. A *fixed* parameter is a parameter that is not free, rather fixed to a specific value, for example, 0 or 1. A *constrained* parameter is a parameter that is unknown, but set to equal one or more other parameters. *Model identification* therefore involves setting variable parameters as fixed, free, or constrained (see **Identification**).

2 Structural Equation Modeling: Overview

There have been traditionally three types of model identification distinctions. The three types are distinguished by whether the sample variance–covariance matrix can uniquely estimate the variable parameters in the theoretical model. A theoretical model is *underidentified* when one or more variable parameters cannot be uniquely determined, *just-identified* when all of the variable parameters are uniquely determined, and *overidentified* when there is more than one way of estimating a variable parameter(s). The just- or overidentified model distinctions are considered model identified, while the underidentified distinction yields unstable parameter estimates, and the degrees of freedom for the theoretical model are zero or negative. An underidentified model can become identified when additional constraints are imposed, that is, the model degrees of freedom equal one or greater.

There are several conditions for model identification. A necessary, but not sufficient, condition for model identification, is the *order condition*, under which the number of free variable parameters to be estimated must be less than or equal to the number of distinct values in the sample variance–covariance matrix, that is, only the diagonal variances and one set of off-diagonal covariances are counted. The number of distinct values in the sample variance–covariance matrix is equal to $p(p + 1)/2$, where p is the number of observed variables. A saturated model (all variables are related in the model) with p variables has $p(p + 3)/2$ free variable parameters. For a sample variance–covariance matrix S with 3 observed variables, there are 6 distinct values [$3(3 + 1)/2 = 6$] and 9 free (independent) parameters [$3(3 + 3)/2$] that can be estimated. Consequently, the number of free parameters estimated in any theoretical model must be less than or equal to the number of distinct values in the variance–covariance matrix. While the *order condition* is necessary, other sufficient conditions are required, for example, the *rank condition*. The *rank condition* requires an algebraic determination of whether each parameter in the model can be estimated from the sample variance–covariance matrix and is related to the determinant of the matrix.

Several different solutions for avoiding model identification problems are available to the researcher. The first solution involves a decision about which observed variables measure each latent variable. A fixed parameter of one for an observed variable or a latent variable will set the measurement scale for that

latent variable, preventing scale *indeterminacy* in the theoretical model for the latent variable. The second solution involves a decision about specifying the theoretical model as *recursive* or *nonrecursive*. A *recursive* model is when all of the variable relationships are unidirectional, that is, no bidirectional paths exist between two latent variables (see **Recursive Models**). A *nonrecursive* model is when a bidirectional relationship (reciprocal path) between two latent variables is indicated in the theoretical model. In *nonrecursive* models, the correlation of the latent variable errors should be included to correctly estimate the parameter estimates for the two latent variables. The third solution is to begin with a less complex theoretical model that has fewer parameters to estimate. The less complex model would only include variables that are absolutely necessary and only if this model is identified would you develop a more complex theoretical model. A fourth solution is to save the model-implied variance–covariance matrix and use it as the original sample variance–covariance matrix. If the theoretical model is identified, then the parameter estimates from both analyses should be identical. A final solution is to use different starting values in separate analyses. If the model is identified, then the estimates should be identical. A researcher can check model identification by examining the degrees of freedom for the theoretical model, the rank test, or the inverse of the information matrix.

Model Estimation

A researcher can designate initial parameter estimates in a theoretical model, but more commonly, the SEM software automatically provides initial default estimates or start values. The *initial default estimates* are computed using a noniterative two-stage least squares estimation method. These initial estimates are consistent and rather efficient relative to other iterative methods. After initial start values are selected, SEM software uses one of several different estimation methods available to calculate the final observed and latent variable parameter estimates in the theoretical model, that is, estimates of the population parameters in the theoretical model (see **Structural Equation Modeling: Software**).

Our goal is to obtain parameter estimates in the theoretical model that compute an implied variance–covariance matrix Σ , which is as close as

possible to the sample variance–covariance matrix S . If all residual variance–covariance matrix elements are zero, then $S - \Sigma = 0$ and $\chi^2 = 0$, that is, a perfect fit of the theoretical model to the sample variance–covariance data. Model estimation therefore involves the selection of a *fitting function* to minimize the difference between Σ and S . Several fitting functions or estimation methods are currently available: unweighted or ordinary least squares (ULS or OLS) (see **Least Squares Estimation**), generalized least squares (GLS), **maximum likelihood (ML)**, weighted least squares (WLS) (see **Least Squares Estimation**), and asymptotic distribution free (ADF). Another goal in model estimation is to use the correct fit function to obtain parameter estimates that are *unbiased, consistent, sufficient, and efficient*, that is, *robust* (see **Estimation**).

The ULS or OLS parameter estimates are consistent, but have no distributional assumptions or associated statistical tests, and are scale-dependent, that is, changes in the observed variable measurement scale yield different sets of parameter estimates. The GLS and ML parameter estimates are not scale-dependent so any transformed observed variables will yield parameter estimates that are related. The GLS and ML parameter estimates have desirable asymptotic properties (large sample properties) that yield minimum error variance and unbiased estimates. The GLS and ML estimation methods assume multivariate normality (see **Catalogue of Probability Density Functions**) of the observed variables (the sufficient conditions are that the observations are independent, identically distributed, and kurtosis is zero). The WLS and ADF estimation methods generally require a large sample size and do not require observed variables to be normally distributed.

Model estimation with binary and ordinal scaled observed variables introduces a parameter estimation problem in structural equation modeling (see **Structural Equation Modeling: Categorical Variables**). If observed variables are ordinal scaled or nonnormally distributed, then GLS and ML estimation methods yield parameter estimates, standard errors, and test statistics that are not robust. SEM software uses different techniques to resolve this problem: a categorical variable matrix (CVM) that does not use Pearson product-moment correlations or an asymptotic variance–covariance matrix based on polychoric correlations of two ordinal variables, polyserial correlations of an ordinal

and an interval variable, and Pearson product-moment correlations of two interval variables. All three types of correlations (Pearson, polychoric, and polyserial) are then used to create an asymptotic covariance matrix for analysis in the SEM software. A researcher *should not* use mixed types of correlation matrices or variance–covariance matrices in SEM software, rather create a CVM or asymptotic variance–covariance matrix.

The type of estimation method to use with different theoretical models is still under investigation. The following recommendations, however, define current practice. A theoretical model with interval scaled multivariate normal observed variables should use the ULS or OLS estimation method. A theoretical model with interval scaled multivariate nonnormal observed variables should use GLS, WLS, or ADF estimation methods. A theoretical model with ordinal scaled observed variables should use the CVM approach or an asymptotic variance–covariance matrix with WLS or ADF estimation methods.

Model Testing

Model testing involves determining whether the sample data fits the theoretical model once the final parameter estimates have been computed, that is, to what extent is the theoretical model supported by the sample variance–covariance matrix. Model testing can be determined using an omnibus global test of the entire theoretical model or by examining the individual parameter estimates in the theoretical model.

An omnibus global test of the theoretical model can be determined by interpreting several different model fit criteria. The various model fit criteria are based on a comparison of the theoretical model variance–covariance matrix Σ to the sample variance–covariance matrix S . If Σ and S are similar, then the sample data fit the theoretical model. If Σ and S are quite different, then the sample data do not fit the theoretical model. The model fit criteria are computed on the basis of knowledge of the saturated model (all variable relationships defined), independence model (no variable relationships defined), sample size, degrees of freedom and/or the chi-square value, and range in value from 0 (no fit) to 1 (perfect fit) for several subjective model fit criteria.

The model fit criteria are categorized according to model fit, model parsimony, and model comparison. Model fit is interpreted using the chi-square

(χ^2), goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), or root-mean-square residual (RMR). The model fit criteria are based on a difference between the sample variance–covariance matrix (S) and the theoretical model–reproduced variance–covariance matrix (Σ). Model comparison is interpreted using the Tucker–Lewis index (TLI), Bentler–Bonett Non-Normed fit index (NNFI), Bentler–Bonett Normed fit index (NFI), or the Bentler comparative fit index (CFI) (*see Goodness of Fit*). The model comparison criteria compare a theoretical model to an independence model (no variable relationships defined). Model parsimony is interpreted using the normed chi-square (NC), parsimonious fit index (PNFI or PCFI), or **Akaike information criterion (AIC)**. The model parsimony criteria are determined by the number of estimated parameters required to achieve a given value for chi-square, that is, an overidentified model is compared with a restricted model.

Model testing can also involve interpreting the individual parameter estimates in a theoretical model for statistical significance, magnitude, and direction. Statistical significance is determined by testing whether a free parameter is different from zero, that is, parameter estimates are divided by their respective standard errors to yield a test statistic. Another interpretation is whether the sign of the parameter estimate agrees with expectations in the theoretical model. For example, if the expectation is that more education will yield a higher income level, then an estimate with a positive sign would support that expectation. A third interpretation is whether the parameter estimates are within an expected range of values. For example, variances should not have negative values and correlations should not exceed one. The interpretation of parameter estimates therefore considers whether parameter estimates are statistically significant, are in the expected direction, and fall within an expected range of acceptable values; thus, parameter estimates should have a practical and meaningful interpretation.

Model Modification

Model modification involves adding or dropping variable relationships in a theoretical model. This is typically done when sample data do not fit the theoretical model, that is, parameter estimates and/or

model test criteria are not reasonable. Basically, the initial theoretical model is modified and the new modified model is subsequently evaluated.

There are a number of procedures available for model modification or what has been termed a *specification search*. An intuitive way to consider modifying a theoretical model is to examine the statistical significance of each parameter estimated in the model. If a parameter is not statistically significant, then drop the variable from the model, which essentially sets the parameter estimate to zero in the modified model. Another intuitive method is to examine the residual matrix, that is, the differences between the sample variance–covariance matrix S and the theoretical model reproduced variance–covariance matrix Σ elements. The residual values in the matrix should be small in magnitude and similar across variables. Large residual values overall indicate that the theoretical model was not correctly specified, while a large residual value for a single variable indicates a problem with that variable only. A large standardized residual value for a single variable indicates that the variable's covariance is not well defined in the theoretical model. The theoretical model would be examined to determine how this particular covariance could be explained, for example, by estimating parameters of other variables.

SEM software currently provides model modification indices for variables whose parameters were not estimated in the theoretical model. The modification index for a particular variable indicates how much the omnibus global chi-square value is expected to decrease in the modified model if a parameter is estimated for that variable. A modification index of 50 for a particular variable suggests that the omnibus global chi-square value for the modified model would be decreased by 50. Large modification indices for variables offer suggestions on how to modify the theoretical model by adding variable relationships in the modified model to yield a better fitting model.

Other model modification indices in SEM software are the *expected parameter change*, *Lagrange multiplier*, and *Wald statistics*. The expected parameter change (EPC) statistic indicates the estimated change in the magnitude and direction of variable parameters if they were to be estimated in a modified model, in contrast to the expected decrease in the omnibus global chi-square value. The EPC is especially informative when the sign of a variable parameter is not in the expected direction, that is, positive instead

of negative. The EPC in this situation would suggest that the parameter for the variable be fixed. The lagrange multiplier (LM) statistic is used to evaluate the effect of freeing a set of fixed parameters in a theoretical model. The Lagrange multiplier statistic can consider a set of variable parameters and is therefore considered the multivariate analogue of the modification index. The Wald (W) statistic is used to evaluate whether variables in a theoretical model should be dropped. The Wald (W) statistic can consider a set of variable parameters and is therefore considered the multivariate analogue of the *individual variable critical values*.

Empirical research suggests that model modification is most successful when the modified model is similar to the underlying population model that reflects the sample variance–covariance data. Theoretically, many different models might fit the sample variance–covariance data. Consequently, new specification search procedures generate all possible models and list the best fitting models based on certain model fit criteria, for example, chi-square, AIC, BIC. For example, a multiple regression equation with 17 independent variables predicting a dependent variable would yield 2^{17} or 131,072 regression models, not all of which would be theoretically meaningful. SEM software permits the formulation of all possible models; however, the outcome of any specification search should still be guided by theory and practical considerations, for example, the time and cost of acquiring the data.

Model Validation

Model validation involves checking the stability and accuracy of a theoretical model. Different methods are available using SEM software: replication, cross-validation, simulation, **bootstrap**, **jackknife**, and specification search. A researcher should ideally seek model validation using additional random samples of data (replication), that is, multiple sample analysis with the same theoretical model. The other validation methods are used in the absence of replication to provide evidence of model validity, that is, the stability and accuracy of the theoretical model.

Cross-validation involves randomly splitting a large sample data set into two smaller data sets. The theoretical model is analyzed using each data set to compare parameter estimates and model fit statistics.

In the simulation approach, a theoretical model is compared to a known population model; hence, a population data set is simulated using a random number generator with a known population model. The bootstrap technique is used to determine the stability of parameter estimates by using a random sample data set as a pseudo-population data set to repeatedly create randomly sampled data sets with replacement. The theoretical model is analyzed using all of the bootstrap data sets and results are compared. The jackknife technique is used to determine the impact of outliers on parameter estimates and fit statistics by creating sample data sets where a different data value is excluded each time. The exclusion of a single data value from each sample data set identifies whether an outlier data value is influencing the results. Specification search examines all possible models and selects the best model on the basis of a set of model fit criteria. This approach permits a comparison of the initial theoretical model to other plausible models that are supported by the sample data.

SEM Software

Several structural equation modeling software packages are available to analyze theoretical models. A theoretical model is typically drawn using squares or rectangles to identify observed variables, small circles or ellipses to identify observed variable measurement error, larger circles or ellipses to identify latent variables, curved arrows to depict variable correlation, and straight arrows to depict prediction of dependent variables by independent variables. A graphical display of the theoretical model therefore indicates direct and indirect effects amongst variables in the model.

Four popular SEM packages are Amos, EQS, Mplus, and LISREL. Amos employs a graphical user interface with drawing icons to create a theoretical model and link the model to a data set, for example, SPSS save file. EQS incorporates a statistics package, model diagrammer, and program syntax to analyze theoretical models. EQS can use either a graphical display or program syntax to analyze models. Mplus integrates random effect, factor, and latent class analysis in both cross-sectional and longitudinal settings for single-level and multilevel designs. LISREL is a matrix command language that specifies the type of matrices to use for a specific theoretical model including which parameters are free and

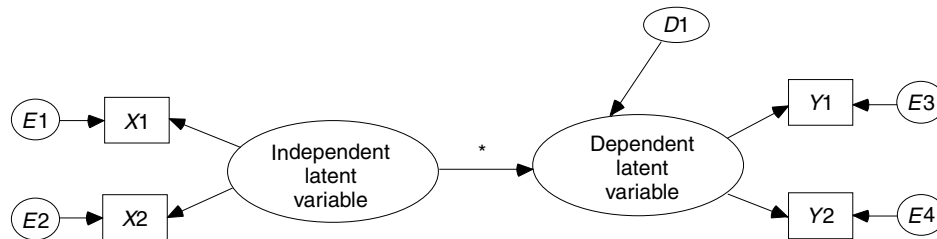


Figure 1 Basic Structural Equation Model

fixed. LISREL includes a data set preprocessor program, PRELIS, to edit and analyze sample data. LISREL also includes a program with simple language commands, SIMPLIS, to input data and specify a theoretical model (*see Structural Equation Modeling: Software*).

All four SEM software packages include excellent documentation, tutorials, and data set examples to illustrate how to analyze different types of theoretical models. The SEM software packages are available at their respective Internet web sites and include student versions:

Amos: <http://www.spss.com/amos>
 EQS: <http://www.mvsoft.com/>
 Mplus <http://www.statmodel.com/>
 LISREL: <http://www.ssicentral.com/>

SEM Example

The theoretical model (Figure 1) depicts a single independent latent variable predicting a single dependent latent variable. The independent latent variable is defined by two observed variables, X_1 and X_2 , with corresponding measurement error designated as E_1 and E_2 . The dependent latent variable is defined by two observed variables, Y_1 and Y_2 , with corresponding measurement error designated as E_3 and E_4 . The parameter estimate or structure coefficient of interest to be estimated is indicated by the asterisk (*) on the straight arrow from the independent latent variable to the dependent latent variable with D_1 representing the error in prediction. The computer output will

also yield an R-squared value that indicates how well the independent latent variable predicts the dependent latent variable.

Further Reading

- Geoffrey, M. (1997). *Basics of Structural Equation Modeling*, Sage Publications, Thousand Oaks, ISBN: 0-8039-7408-6; 0-8039-7409-4 (pbk).
- Marcoulides, G. & Schumacker, R.E., eds (1996). *Advanced Structural Equation Modeling: Issues and Techniques*, Lawrence Erlbaum Associates, Mahwah.
- Marcoulides, G.A. & Schumacker, R.E. (2001). *Advanced Structural Equation Modeling: New Developments and Techniques*, Lawrence Erlbaum Associates, Mahwah.
- Schumacker, R.E. & Marcoulides, G.A., eds (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.
- Schumacker, R.E. & Lomax, R.G. (2004). *A Beginner's Guide to Structural Equation Modeling*, 2nd Edition, Lawrence Erlbaum Associates, Mahwah.
- Structural Equation Modeling: A Multidisciplinary Journal*. Lawrence Erlbaum Associates, Inc, Mahwah.
- Tenko, R. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah, ISBN: 0-8058-3568-7; 0-8058-3569-5 (pbk).

(*See also Linear Statistical Models for Causation: A Critical Review; Residuals in Structural Equation, Factor Analysis, and Path Analysis Models; Structural Equation Modeling: Checking Substantive Plausibility; Structural Equation Modeling: Nontraditional Alternatives*)

RANDALL E. SCHUMACKER

Structural Equation Modeling: Software

EDWARD E. RIGDON

Volume 4, pp. 1947–1951

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling: Software

Introduction

The first widely distributed special-purpose software for estimating **structural equation models** (SEM) appeared more than twenty years ago. The earliest packages were written with the Fortran programming language and designed to run on mainframe computers, in an age when computing time was a scarce commodity and users punched instructions on stacks of cards. Features and terminology reminiscent of that earlier time still survive in the latest versions of some older SEM packages. While SEM software has become much more sophisticated, professional and user-friendly over the years, SEM software still is not what it could be. There is still substantial room for improvement.

Today, researchers can choose from a variety of packages that differ markedly in terms of their intellectual heritage, interface, statistical sophistication, flexibility, integration with other statistical packages, and price. Researchers may find that a few key criteria will substantially narrow their list of possible choices. Then again, packages regularly add capabilities, often mimicking their competitors, so any purchase or use decision should be based on the latest information.

This overview includes only packages that are primarily designed to estimate conventional structural equation models, with extensions. It excludes packages which are designed for SEM variants such as Tetrad and **Partial Least Squares** (*see Structural Equation Modeling: Nontraditional Alternatives*). It also excludes sophisticated modeling packages such as aML (<http://www.applied-ml.com/>) or GLLAMM (<http://www.gllamm.org/>), which are powerful and flexible tools but which lack many of the features, such as a broad array of fit indices and other fit diagnostics from the SEM literature, that users would expect in a SEM package.

Choices

At least a dozen packages are available whose primary or major purpose is the estimation of structural equation models. Typically, each package began as a

tool created by leading SEM researchers to facilitate their own analyses, only gradually being adapted to the needs of a larger body of customers. This origin as a tool for a SEM expert may partly explain why these packages seem to offer so little help to the average or novice user. This article starts with a brief overview of most of the known packages, in alphabetical order.

The Amos package, written by James Arbuckle and distributed by SPSS (<http://www.spss.com/amos/>) was unusual when it first appeared. It was perhaps the first SEM package to be designed for a graphical computing environment, like Microsoft Windows®. Taking advantage of the environment's capabilities, Amos allowed users to specify models by drawing them, and offered users a set of drawing tools for the purpose. (Amos also includes a command language called Amos Basic.) Amos was also an early leader in implementing advanced missing data techniques (*see Missing Data*). While these capabilities have been copied, to a certain degree, by other leading programs, Amos retains a reputation for being easy to use, and the package has continued to add innovations in this area.

Proc Calis, written by Wolfgang Hartmann, is a procedure within the SAS package (<http://www.sas.com/>). In the early 1980s, Proc Calis was arguably the most sophisticated SEM package available. Its ability to specify constraints on model parameters as nonlinear functions of other parameters was instrumental in allowing researchers to model quadratic effects and multiplicative interactions between latent variables [2] (*see Structural Equation Modeling: Nonstandard Cases*). Over the intervening years, however, Proc Calis has not added features and extended capabilities to keep pace with developments.

EQS, written by Peter Bentler, has long been one of the leading SEM packages, thanks to the extensive contributions of its author, both to the program and to the field of SEM (<http://www.mvsoft.com/>). EQS was long distinguished by special features for dealing with nonnormal data, such as a **kurtosis-adjusted χ^2** statistic [4], and superior procedures for modeling ordinal data (*see Ordinal Regression Models*).

LISREL, written by Karl Jöreskog and Dag Sörbom, pioneered the field of commercial SEM software, and it still may be the single most widely used and well-known package (<http://www.>

ssicentral.com/). Writers have regularly blurred the distinction between LISREL as software and SEM as statistical method. Along with EQS, LISREL was an early leader in offering procedures for modeling ordinal data. Despite their program's advantage in terms of name recognition, however, the pressure of commercial and academic competition has forced the authors to continue updating their package. LISREL's long history is reflected both in its clear Fortran legacy and its enormous worldwide knowledge base.

Mplus (<http://www.statmodel.com/>), written by Linda and Bengt Muthén, is one of the newest entrants, but it inherits a legacy from Bengt Muthén's earlier program, LISCOMP. Besides maintaining LISCOMP's focus on nonnormal variables, Mplus has quickly built a reputation as one of the most statistically sophisticated SEM packages. Mplus includes tools for finite mixture modeling and latent class analysis that go well beyond conventional SEM, but which may point to the future of the discipline.

Mx (<http://griffin.vcu.edu/mx/>), written by Michael Neale, may be as technically sophisticated as any product on the market, and it has one distinguishing advantage: it's free. The software, the manual, and a graphical interface are all available free via the Internet. At its core, perhaps, Mx is really a matrix algebra program, but it includes everything that most users would expect in a SEM program, as well as leading edge capabilities in modeling incomplete data and in finite mixture modeling (*see* **Finite Mixture Distributions**).

SEPATH, written by James Steiger, is part of the Statistica statistical package (<http://www.statsoftinc.com/products/advanced.html#structural>). SEPATH incorporates many of Steiger's innovations relating to the analysis of correlation matrices. It is one of the few packages that automatically provides correct estimated standard errors for parameter estimates from SEM analysis of a Pearson correlation matrix (*see* **Correlation and Covariance Matrices**). Most packages still produce biased estimated standard errors in this case, unless the user takes some additional steps.

Other SEM packages may be less widely distributed but they are still worth serious consideration. LINCOS, written by Ronald Schoenberg, and MECOSA, written by Gerhard Arminger, use the GAUSS (<http://www.aptech.com/>) matrix algebra programming language and require a GAUSS installation. RAMONA, written by Michael Browne

and Gerhard Mels, is distributed as part of the Systat statistical package (<http://www.systat.com/products/Systat/productinfo/?sec=1006>). SEM, by John Fox, is written in the open-source R statistical programming language (<http://socserv.socsci.mcmaster.ca/jfox/Misc/sem/index.html>). Like Mx, Fox's SEM is free, under an open-source license.

Finally, STREAMS (<http://www.mwstreams.com/>) is not a SEM package itself, but it may make SEM packages more user-friendly and easier to use. STREAMS generates language for, and formats output from, a variety of SEM packages. Users who need to take advantage of specific SEM package capabilities might use STREAMS to minimize the training burden.

Criteria

Choosing a SEM software package involves potential tradeoffs among a variety of criteria. The capabilities of different SEM packages change regularly – Amos, EQS, LISREL, Mplus, and Mx have all announced or released major upgrades in the last year or two—so it is important to obtain updated information before making a choice.

Users may be inclined to choose a SEM package that is associated with a more general statistical package which they already license. Thus, licensees of SPSS, SAS, Statistica, Systat, or GAUSS might favor Amos, Proc Calis, SEPath, RAMONA or MECOSA, respectively. Keep in mind that leading SEM packages will generally have the ability to import data in a variety of file formats, so users do not need to use the SEM software associated with their general statistical package in order to be able to share data across applications.

Many SEM packages are associated with specific contributors to SEM, as noted previously. Researchers who are familiar with a particular contributor's approach to SEM may prefer to use the associated software. Beyond general orientation, a given contributor's package may be the only one that includes some of the contributor's innovations. Over time, successful innovations do tend to be duplicated across programs, but users cannot assume that any given package includes tools for dealing with all modeling situations. Currently, for example, Amos does not include any procedures for modeling ordinal data, while Proc Calis does not support multiple

group analysis, except when all groups have exactly the same sample size. Some features are largely exclusive, at least for the moment. Mplus and Mx are the only packages with tools for mixture modeling and latent class analysis. This allows these programs to model behavior that is intrinsically categorical, such as voting behavior, and also provides additional options for dealing with heterogeneity in a data set [3]. Similarly, only LISREL and Mplus have procedures for obtaining correct statistical results from modeling ordinal data without requiring exceptionally large sample sizes.

'Ease of use' is a major consideration for most users, but there are many different ways in which a package can be easy to use. Amos pioneered the graphical specification of models, using drawing tools to specify networks of observed and latent variables. Other packages followed suit, to one extent or another. Currently, LISREL allows users to modify a model with drawing tools but not to specify an initial model graphically. On the other hand, specifying a model with drawing tools can become tedious when the model involves many variables, and there is no universal agreement on how to graphically represent certain statistical features of a model. Menus, 'wizards' or other helps may actually better facilitate model specification in such cases. Several packages allow users to specify models through equation-like statements, which can provide a convenient basis for specifying relations among groups of variables.

'Ease of use' can mean ease of interpreting results from an analysis. Researchers working with multiple, competing models can easily become overwhelmed by the volume of output. In this regard, Amos and Mx have special features to facilitate comparisons among a set of competing models.

'Ease of use' can also mean that it is easy to find help when something goes wrong. Unfortunately, no SEM package does an excellent job of helping users in such a situation. To a great extent, users of any package find themselves relying on fellow users for support and advice. Thus, the popularity and history of a particular package can be important considerations. On that score, veteran packages like LISREL offer a broad population of users and a deep knowledge base, while less widely used packages like Proc Calis present special problems.

'Ease of use' could also relate to the ability to try out a product before committing to a major

purchase. Many SEM packages offer 'student' or 'demo' versions of the software, usually offering full functionality but only for a limited number of variables. Some packages do not offer a demo version, which makes the purchase process risky and inconvenient. Obviously, free packages like Mx do not require demo versions.

Finally, users may be concerned about price. Developing a full-featured SEM package is no small task. Add to that the costs of supporting and promoting the package, factor in the small user base, relative to more general statistical packages, and it is not surprising that SEM packages tend to be expensive. That said, users should give special consideration to the Mx package, which is available free via the Internet. Mx offers a graphical interface and a high degree of flexibility, although specifying some model forms will involve some programming work. Still, templates are available for conducting many types of analysis with Mx.

Room for Improvement

SEM software has come a long way from the opaque, balky, idiosyncratic, mainframe-oriented packages of the early 1980s, but today's SEM packages still frustrate and inconvenience users and fail to facilitate SEM analysis as well as they could, given only the tools available today. SEM packages have made it easier to specify basic models, but specifying advanced models may require substantial programming, often giving rise to tedious errors, even though there is very little user discretion involved, once the general form of the advanced model is chosen. Users sometimes turn to bootstrap and Monte Carlo methods, as when sample size is too small for stable estimation, and several packages offer this capability. Yet, users find themselves 'jumping through hoops' to incorporate these results into their analyses, even though, once a few choices are made, there really is nothing left but tedium. There is much more to be done in designing SEM packages to maximize the efficiency of the researcher.

Most SEM packages do a poor job of helping users when something goes wrong. For example, when a user's structural equation model is not identified – meaning that the model's parameters cannot be uniquely estimated – SEM packages

either simply fail or point the user to one particular parameter that is involved in the problem. Pointing to one parameter, however, may direct the user more to the symptoms and away from the fundamental problem in the model. Bekker, Merckens and Wansbeek [1] demonstrated a procedure, implemented via a Pascal program, that indicates all model parameters that are not identified, but this procedure has not been adopted by any major SEM package.

Several packages have taken steps in the area of visualization, but much more could be done. Confirmatory factor analysis measurement models imply networks of constraints on the patterns of covariances among a set of observed variables. When such a model performs poorly, there are sets of covariances that do not conform to the implied constraints. Currently, SEM packages will point to particular elements of the empirical covariance matrix where the model does a poor job of reproducing the data, and they will also point to particular parameter constraints that contribute to lack of fit. But again, as with identification, this is not the same as showing the user just how the data contradict the network of constraints implied by the model.

Packages could probably improve the advice that they provide about how a researcher might improve a poorly fitting model. Spirtes, Scheines, and Glymour [5] have demonstrated algorithms for quickly finding structures that are consistent with data, subject to constraints. Alongside the diagnostics currently provided, SEM packages could offer more insight to researchers by incorporating algorithms from the stream of research associated with the Tetrad program.

SEM users often find themselves struggling in isolation to interpret results that are somewhat vague, even though there is a large body of researcher experience with SEM generally and with any given program in particular. One day, perhaps, programs will not only generate results but will also help researchers evaluate those results, drawing on this body of experience to give the individual research greater perspective. This type of innovation – giving meaning to numbers – is emerging from the field of artificial intelligence, and it will surely come to structural equation modeling, one day.

References

- [1] Bekker, P.A., Merckens, A. & Wansbeek, T.J. (1994). *Identification, Equivalent Models, and Computer Algebra*, Academic Press, Boston.
- [2] Kenny, D.A. & Judd, C.M. (1984). Estimating the non-linear and interactive effects of latent variables, *Psychological Bulletin* **96**, 201–210.
- [3] Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations: presidential address to the psychometric society, *Psychometrika* **54**, 557–585.
- [4] Satorra, A. & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis, in 1988 Proceedings of the Business and Economic Statistics Section of the American Statistical Association, Washington, DC, 308–313.
- [5] Spirtes, P., Scheines, R. & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL Programs, *Sociological Methods & Research* **19**, 3–66.

(See also **Software for Statistical Analyses**)

EDWARD E. RIGDON

Structural Equation Modeling and Test Validation

BRUNO D. ZUMBO

Volume 4, pp. 1951–1958

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Equation Modeling and Test Validation

Ideally, test and measurement validation entails theoretical as well as empirical studies (*see* **Validity Theory and Applications**). Moreover, the term validation implies a process that takes place over time, often in a sequentially articulated fashion. The choice of statistical methods and research methodology for empirical data analyses is of course central to the viability of validation studies. The purpose of this entry is to describe developments in test and measurement validation as well as an important advancement in the statistical methods used in test validation research, structural equation modeling. In particular, a generalized linear **structural equation model** (GLISEM) that is a **latent variable** extension of a **generalized linear model** (GLIM) is introduced and shown to be particularly useful as a statistical methodology for test and measurement validation research.

A Brief Overview of Current Thinking in Test Validation

Measurement or test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals from a given sample and in a given context. The concept, method, and process of validation are central to constructing and evaluating measures used in the social, behavioral, health, and human sciences, for without validation, any inferences made from a measure are potentially meaningless.

The above definition highlights two central features in current thinking about validation. First, it is not the measure per se that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from scores obtained from a scale may appear subtle at first blush but, in fact, it has significant implications for measurement and testing because it highlights that the validity of the inferences one makes from test scores is somewhat bounded by place, time, and use of the scores resulting from a measurement operation.

The second central feature in the above definition is the clear statement that inferences made from all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an 'objective' educational, economic, or health indicator such as number of students finishing grade 12, or a more psychological measure such as a self-report depression measure, one must be concerned with the validity of the inferences.

It is instructive to contrast contemporary thinking in **validity theory** with what is commonly seen in many introductory texts in research methodology in the social, behavioral, health, and human sciences.

The Traditional View of Validity

The traditional view of validity focuses on (a) validity as a property of the measurement tool, (b) a measure is either valid or invalid, various types of validity – usually four – with the test user, evaluator, or researcher typically assuming only one of the four types is needed to have demonstrated validity, (c) validity as defined by a set of statistical methodologies, such as correlation with a gold-standard, and (d) reliability is a necessary, but not sufficient, condition for validity.

The traditional view of validity can be summarized in Table 1.

The process of validation then simply portrayed as picking the most suitable strategy from Table 1 and conducting the statistical analyses. The basis for much validation research is often described as a correlation with the 'gold standard'; this correlation is commonly referred to as a validity coefficient.

The Contemporary View of Validity

Several papers are available that describe important current developments in validity theory [4, 5, 9, 12, 13, 20]. The purpose of the contemporary view of validity, as it has evolved over the last two decades, is to expand upon the conceptual framework and power of the traditional view of validity seen in most introductory methodology texts. In brief, the recent history of validity theory is perhaps best captured by the following observations.

2 Structural Equation Modeling and Test Validation

Table 1 The traditional categories of validity

Type of validity	What does one do to show this type of validity?
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure with the criterion measure obtained in the present
B. Predictive	Select a criterion and correlate the measure with the criterion measure obtained in the future
Construct (A. Convergent and B. Discriminant):	Can be done several different ways. Some common ones are (a) correlate to a 'gold standard', (b) factor analysis, (c) multitrait multimethod approaches

1. Validity is no longer a property of the measurement tool but rather of the inferences made from the scores.
2. Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum.
3. Construct validity is the central most important feature of validity.
4. There are no longer various types of validity but rather different sources of evidence that can be gathered to aid in demonstrating the validity of inferences.
5. Validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods.
6. As one can see in Zumbo's [20] volume, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall validity framework for evaluating test interpretation and use came the consideration of the role of ethical and social consequences as validity evidence contributing to score meaning. This movement has been met with some resistance. In the end, Messick [14] made the point most succinctly when he stated that one should not be simply concerned with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of 'normal' test use. The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results affecting educational financing and staffing, as well as health care outcomes and financing in the United States and Canada.
7. Although it was initially set aside in the move to elevate construct validity, content-based evidence is gaining momentum again in part due to the work of Sireci [19].
8. Of all the threats to valid inferences from test scores, test translation is growing in awareness due to the number of international efforts in testing and measurement (see, for example, [3]).
9. And finally, there is debate as to whether reliability is a necessary but not sufficient condition for validity; it seems that this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in their inferences. Specifically, reliability is a question of *data quality*, whereas validity is a question of *inferential quality*. Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter.

In a broad sense, then, validity is about evaluating the inferences made from a measure. All of the methods discussed in this encyclopedia (e.g., factor analysis, **reliability**, **item analysis**, **item response modeling**, **regression**, etc.) are directed at building the evidential basis for establishing valid inferences. There is, however, one class of methods that are particularly central to the validation process, **structural equation models**. These models are particularly important to test validation research because they are a marriage of regression, path analysis, and latent variable modeling (often called **factor analysis**). Given that the use of latent variable structural equation models presents one of the most exciting new developments with implications for validity theory, the next section discusses these models in detail.

Generalized Linear Structural Equation Modeling

In the framework of modern statistical theory, test validation research involves the analysis of covariance matrices among the observed empirical data that arise from a validation study using **covariance structure models**. There are two classes of models that are key to validation research: **confirmatory factor analysis** (CFA) (*see Factor Analysis: Confirmatory*) and multiple indicators multiple causes (MIMIC) models. The former have a long and rich history in validation research, whereas the latter are more novel and are representative of the merger of the structural equation modeling and item response theory traditions to what will be referred to as generalized linear structural equation models. Many very good examples and excellent texts describing CFA are widely available (e.g., [1, 2, 10]). MIMIC models are a relatively novel methodology with only heavily statistical descriptions available.

An Example to Motivate the Statistical Problem

Test validation with SEM will be described using the Center for Epidemiologic Studies Depression scale (CES-D) as an example. The CES-D is useful as a demonstration because it is commonly used in the life and social sciences. The CES-D is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. The CES-D prompts the respondent to reflect upon his/her last week and respond to questions such as ‘My sleep was restless’ using an ordered or Likert response format of ‘not even one day’, ‘1 to 2 days’, ‘3 to 4 days’, ‘5 to 7 days’ during the last week. The items typically are scored from zero (not even one day) to three (5–7 days). Composite scores, therefore, range from 0 to 60, with higher scores indicating higher levels of depressive symptoms. The data presented herein is a subsample of a larger data set collected in northern British Columbia, Canada. As part of a larger survey, responses were obtained from 600 adults in the general population -290 females with an average age of 42 years with a range of 18 to 87 years, and 310 males with an average age of 46 years and a range of 17 to 82 years.

Of course, the composite scale score is not the phenomenon of depression, per se, but rather is

related to depression such that a higher composite scale score reflects higher levels of the latent variable depression. Cast in this way, two central questions of test validation are of interest: (a) Given that the items are combined to create one scale score, do they measure just one latent variable? and (b) Are the age and gender of the respondents predictive of the latent variable score on the CES-D? The former question is motivated by psychometric necessities whereas the latter question is motivated by theoretical predictions.

CFA Models in Test Validation

The first validation question described above is addressed by using CFA. In the typical CFA model, the score obtained on each item is considered to be a linear function of a latent variable and a stochastic error term. Assuming p items and one latent variable, the linear relationship may be represented in matrix notation as

$$y = \Lambda\eta + \varepsilon, \quad (1)$$

where y is a $(p \times 1)$ column vector of continuous scores for person i on the p items, Λ is a $(p \times 1)$ column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, η is the latent variable score for person i , and ε is $(p \times 1)$ column vector of measurement residuals. It is then straightforward to show that for items that measure one latent variable, (1) implies the following equation:

$$\Sigma = \Lambda\Lambda' + \Psi, \quad (2)$$

where Σ is the $(p \times p)$ population covariance matrix among the items and Ψ is a $(p \times p)$ matrix of covariances among the measurement residuals or unique factors, Λ' is the transpose of Λ , and Λ is as defined above. In words, (2) tells us that the goal of CFA, like all factor analyses, is to account for the covariation among the items by some latent variables. In fact, it is this accounting for the observed covariation that is fundamental definition of a latent variable – that is, a latent variable is defined by local or conditional independence.

More generally, CFA models are members of a larger class of general linear structural models for a p -variate vector of variables in which the empirical data to be modeled consist of the $p \times p$ unstructured estimator, the sample covariance

matrix, S , of the population covariance matrix, Σ . A confirmatory factor model is specified by a vector of q unknown parameters, θ , which in turn may generate a covariance matrix, $\Sigma(\theta)$, for the model. Accordingly, there are various estimation methods such as generalized **least-squares** or **maximum likelihood** with their own criterion to yield an estimator $\hat{\theta}$ for the parameters, and a legion of test statistics that indicate the similarity between the estimated model and the population covariance matrix from which a sample has been drawn (i.e., $\Sigma = \Sigma(\theta)$). That is, formally, one is trying to ascertain whether the covariance matrix implied by the measurement model is the same as the observed covariance matrix,

$$S \cong \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi} = \Sigma(\hat{\theta}) = \hat{\Sigma}, \quad (3)$$

where the symbols above the Greek letters are meant to imply sample estimates of these population quantities.

As in regression, the goal of CFA is to minimize the error (in this case, the off-diagonal elements of the residual covariance matrix) and maximize the fit between the model and the data. Most current indices of model fit assess how well the model reproduces the observed covariance matrix.

In the example with the CES-D, a CFA model with one latent variable was specified and tested using a recent version of the software LISREL (*see Structural Equation Modeling: Software*). Because the CES-D items are ordinal (and hence not continuous) in nature (in our case a four-point response scale) a polychoric covariance matrix was used as input for the analyses. Using a polychoric matrix is an underlying variable approach to modeling ordinal data (as opposed to an item response theory approach). For a polychoric correlation matrix (*see Polychoric Correlation*), an underlying continuum for the polytomous scores is assumed and the observed responses are considered manifestations of respondents exceeding a certain number of latent thresholds on that underlying continuum. Conceptually, the idea is to estimate the latent thresholds and model the observed cross-classification of response categories via the underlying latent continuous variables. Formally, for item j with response categories $c = 0, 1, 2, \dots, C - 1$, define the latent variable y^* such that

$$y_j = c \text{ if } \tau_c < y_j^* < \tau_{c+1}, \quad (4)$$

where τ_c , τ_{c+1} are the latent thresholds on the underlying latent continuum, which are typically spaced at nonequal intervals and satisfy the constraint $-\infty = \tau_0 < \tau_1 < \dots < \tau_{C-1} < \tau_C = \infty$. It is worth mentioning at this point that the latent distribution does not necessarily have to be normally distributed, although it commonly is due to its well understood nature and beneficial mathematical properties, and that one should be willing to believe that this model with an underlying latent dimension is actually realistic for the data at hand.

Suffice it to say that an examination of the fit indices for our example data with the CES-D, such as the root mean-squared error of approximation (RMSEA), a measure of model fit, showed that the one latent variable model was considered adequate, RMSEA = 0.069, with a 90% confidence interval for RMSEA of 0.063 to 0.074.

The single population CFA model, as described above, has been generalized to allow one to test the same model simultaneously across several populations. This is a particularly useful statistical strategy if one wants to ascertain whether their measurement instrument is functioning the same away in subpopulations of participants (e.g., if a measure functioning the same for males and females). This multigroup CFA operates with the same statistical engine described above with the exception of taking advantage of the statistical capacity of partitioning a likelihood ratio Chi-square and hence testing a series of nested models for a variety of tests of scale level measurement invariance (see [1], for details).

MIMIC Models in Test Validation

The second validation question described above (i.e., are age and gender predictive of CES-D scale scores?) is often addressed by using ordinary least-squares **regression** by regressing the observed composite score of the CES-D onto age and the dummy coded gender variables. The problem with this approach is that the regression results are biased by the measurement error in the observed composite score. Although widely known among psychometricians and statisticians, this bias is ignored in a lot of day-to-day validation research.

The more optimal statistical analysis than using OLS regression is to use SEM and MIMIC models. MIMIC models were first described by Jöreskog

and Goldberger [7]. MIMIC models, in their essence, posit a model stating that a set of possible observed explanatory variables (sometimes called *predictors or covariates*) affects latent variables, which are themselves indicated by other observed variables. In our example of the CES-D, the age and gender variables are predictors of the CES-D latent variable, which itself is indicated by the 20 CES-D items. Our example highlights an important distinction between the original MIMIC models discussed over the last three decades and the most recent developments in MIMIC methodology – in the original MIMIC work the indicators of the latent variable(s) were all continuous variables. In our case, the indicators for the CES-D latent variables (i.e., the CES-D items) are ordinal or Likert variables. This complicates the MIMIC modeling substantially and, until relatively recently, was a major impediment to using MIMIC models in validation research.

The recent MIMIC model for ordinal indicator variables is, in short, an example of the merging of statistical ideas in generalized linear models (e.g., logit and probit models) and structural equation modeling into a generalized linear structural modeling framework [6, 8, 16, 17, 18]. This new framework builds on the correspondence between factor analytic models and **item response theory** (IRT) models (see, e.g., [11]) and is a very general class of models that allow one to estimate group differences, investigate predictors, easily compute IRT with multiple latent variables (i.e., multidimensional IRT), investigate differential item functioning, and easily model complex data structures involving complex item and test formats such as testlets, item bundles, test method effects, or correlated errors all with relatively short scales, such as the CES-D.

A recent paper by Moustaki, Jöreskog, and Mavridis [15] provides much of the technical detail for the generalized linear structural equation modeling framework discussed in this entry; therefore, I will provide only a sketch of the statistical approach to motivate the example with the CES-D. In this light, it should be noted that these models can be fit with either Mplus or PRELIS-LISREL. I chose to use the PRELIS-LISREL software, and hence my description of the generalized linear structural equation model will use Jöreskog's notation.

To write a general model allowing for predictors of the observed (manifest) and latent variables, one extends (1) with a new matrix that contains the

predictors x

$$\begin{aligned} y^* &= \Lambda z + Bx + u, \text{ where} \\ z &= Dw + \delta, \end{aligned} \quad (5)$$

and u is an error term representing a specific factor and measurement error and y^* is an unobserved continuous variable underlying the observed ordinal variable denoted y , z is a vector of latent variables, w is a vector of fixed predictors (also called *covariates*), D is a matrix of regression coefficients and δ is a vector of error terms which follows a $N(0, I)$. Recall that in (1) the variable being modeled is directly observed (and assumed to be continuous), but in (5) it is not.

Note that because the PRELIS-LISREL approach does not specify a model for the complete p -dimensional response pattern observed in the data, one needs to estimate the model in (5) with PRELIS-LISREL one follows two steps. In the first step (the PRELIS step), one models the univariate and bivariate marginal distributions to estimate the thresholds and the joint covariance matrix of y^* , x , and w and their asymptotic covariance matrix. In the PRELIS step there is no latent variable imposed on the estimated joint covariance matrix hence making that matrix an unconstrained covariance matrix that is just like a sample covariance matrix, S , in (3) above for continuous variables. It can therefore be used in LISREL for modeling just as if y^* was directly observed using (robust) maximum likelihood or weighted least-squares estimation methods.

Turning to the CES-D example, the validity researcher is interested in the question of whether age and gender are predictive of CES-D scale scores. Figure 1 is the resulting generalized MIMIC model. One can see in Figure 1 that the correlation of age and gender is, as expected from descriptive statistics of age for each gender, negative. Likewise, if one were to examine the t values in the LISREL output, both the age and gender predictors are statistically significant. Given the female respondents are coded 1 in the binary gender variable, as a group the female respondents scored higher on the latent variable of depression. Likewise, the older respondents tended to have a lower level of depression compared to the younger respondents in this sample, as reflected in the negative regression coefficient in Figure 1. When the predictive relationship of age was investigated separately for males and females via

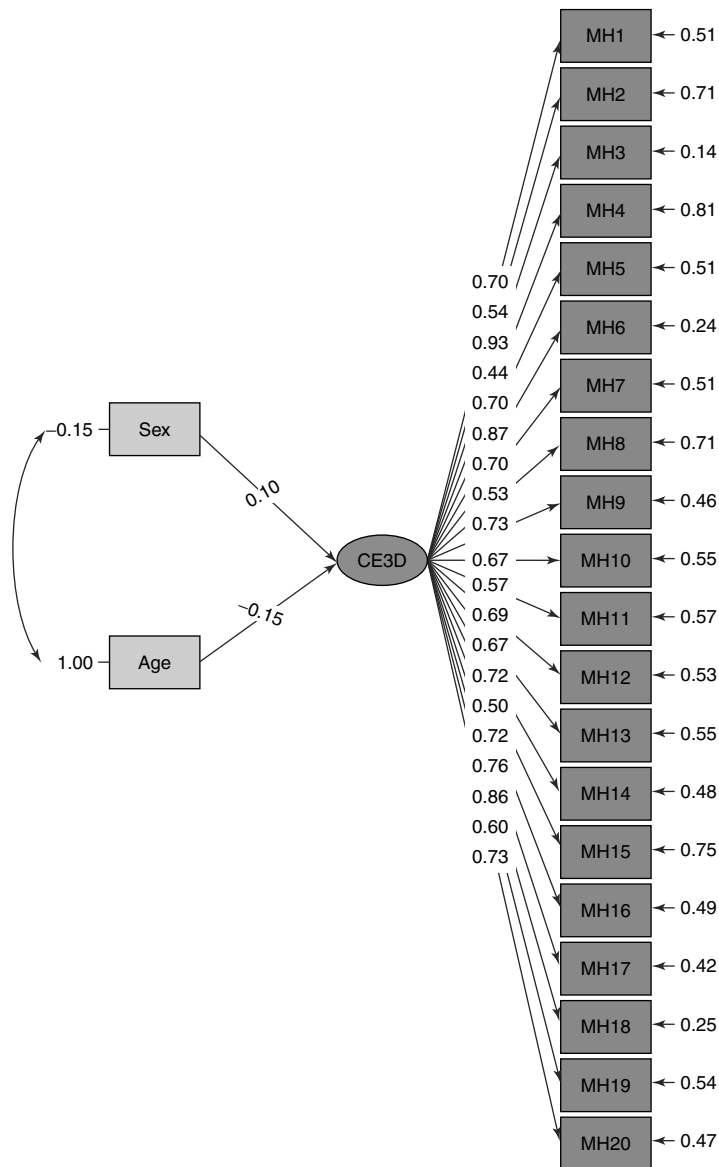


Figure 1 MIMIC model of age and gender for the CES-D (Standardized solution)

this generalized MIMIC model, age was a statistically significant (negative) predictor for the female respondents and age was not a statistically significant for male respondents. Age is unrelated to depression level for men, whereas older women in this sample are less depressed than younger women. This sort of predictive validity information is useful to researchers using the CES-D and hence supports, as described at the beginning of this entry, the inferences made from CES-D test scores.

References

- [1] Byrne, B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*, Lawrence Erlbaum, Hillsdale, N.J.
- [2] Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, Lawrence Erlbaum, Mahwah.
- [3] Hambleton, R.K. & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, Netherlands, pp. 153–171.
- [4] Hubley, A.M. & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going, *The Journal of General Psychology* **123**, 207–215.
- [5] Johnson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices, *Educational and Psychological Measurement* **58**, 736–753.
- [6] Jöreskog, K.G. (2002). Structural equation modeling with ordinal variables using LISREL. Retrieved December 2002 from <http://www.ssicentral.com/lisrel/ordinal.htm>
- [7] Jöreskog, K.G. & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable, *Journal of the American Statistical Association* **10**, 631–639.
- [8] Jöreskog, K.G. & Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches, *Multivariate Behavioral Research* **36**, 341–387.
- [9] Kane, M.T. (2001). Current concerns in validity theory, *Journal of Educational Measurement* **38**, 319–342.
- [10] Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, Newbury Park.
- [11] Lu, I.R.R., Thomas, D.R. & Zumbo, B.D. (in press). Embedding IRT in structural equation models: A comparison with regression based on IRT scores, *Structural Equation Modeling*.
- [12] Messick, S. (1989). Validity, in *Educational Measurement*, R.L. Linn, ed., 3rd Edition, American Council on Education/Macmillan, New York, pp. 13–103.
- [13] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *American Psychologist* **50**, 741–749.
- [14] Messick, S. (1998). Test validity: A matter of consequence, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 35–44.
- [15] Moustaki, I., Jöreskog, K.G. & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: a comparison of LISREL and IRT approaches, *Structural Equation Modeling* **11**, 487–513.
- [16] Muthen, B.O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables, *Journal of Educational Statistics* **10**, 121–132.
- [17] Muthen, B.O. (1988). Some uses of structural equation modeling in validity studies: extending IRT to external variables, in *Test validity*, H. Wainer & H. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 213–238.
- [18] Muthen, B.O. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika* **54**, 551–585.
- [19] Sireci, S.G. (1998). The construct of content validity, in *Validity Theory and the Methods used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 83–117.
- [20] Zumbo, B.D., ed. (1998). Validity theory and the methods used in validation: perspectives from the social and behavioral sciences, Special issue of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* **45**(1–3), 1–359.

BRUNO D. ZUMBO

Structural Zeros

VANCE W. BERGER AND JIALU ZHANG

Volume 4, pp. 1958–1959

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Structural Zeros

Empty or zero cells in a **contingency table** can be classified as either structural zeros or random (sampling) zeros. A sampling zero occurs when the observed cell count in the table is zero, while its expected value is not. This is especially likely when both the sample size and the cell probability are small. For any positive cell probability, however, increasing the sample size sufficiently will ensure that with high probability, the cell count will not be zero; that is, there will not be a random zero. In contrast, increasing the sample size does not have this effect on structural zeros. This is because a cell with a structural zero has an expected value of zero. Clearly, as a nonnegative random variable, this means that its variance is also zero, and that not only did no observations in the data set at hand fall into that cell, but in fact that no observation *could* fall into that cell. The cell count is zero with probability one.

Sampling zeros are part of the data and contribute to the likelihood function (see **Maximum Likelihood Estimation**) and model fitting, while structural zeros are not part of the data [1]. Therefore, they do not contribute to the likelihood function or model fitting. A contingency table containing structural zeros is, in some sense, an incomplete table and special analysis methods are needed to deal with structural zeros. Agresti [1] gave an example of a contingency table (Table 1) with a structural zero [1]. The study investigated the effect of primary pneumonia infections in calves on secondary pneumonia infections. The 2×2 table has primary infection and secondary infection as the row variable and the column variable, respectively. Since a secondary infection is not possible without a primary infection, the lower left cell has a structural zero. If any of the other three cells had turned out to have a zero count, then they would have been sampling zeros.

Table 1 An Example of a Structural Zero in a 2×2 Contingency Table

	Secondary infection		
	Yes	No	
Primary			
Yes	a (π_{11})	b (π_{12})	π_{1+}
No	—	c (π_{22})	π_{2+}
	π_{+1}	π_{+2}	

Other examples in which structural zeros may arise include cross-classifications by number of children in a household and number of smoking children in a household, number of felonies committed on a given day in a given area and number of these felonies for which at least one suspect was arrested and charged with the felony in question, and number of infections experienced and number of serious infections experienced for patients in a study. Yu and Zelterman [4] presented a triangular table with families classified by both number of siblings (1–6) and number of siblings with interstitial pulmonary fibrosis (0–6). The common theme is that the initial classification bears some resemblance to a Poisson variable, or a count variable, and the secondary variable bears some resemblance to a binary classification of each Poisson observation.

The structural zero occurs because of the restriction on the number of binary ‘successes’ imposed by the total Poisson count; that is, there cannot be a family with two children and three children smokers, or a day with six felonies and eight felonies with suspects arrested, or a patient with no infection but one serious infection. Table 1 of [4] is triangular for this reason too; that is, every cell above the main diagonal, in which the number of affected siblings would exceed the number of siblings, is a structural zero.

For two-way frequency tables, the typical analyses are based on Pearson’s χ^2 test or Fisher’s exact test (see **Exact Methods for Categorical Data**). These are tests of association (see **Measures of Association**) of the row and column variables. The null hypothesis is that the row and column variables are independent, while the alternative hypothesis is that the variables are associated. The usual formulation of ‘no association’ is that the cell probabilities in any given column are common across the rows, or $p(1,1) = p(2,1)$ and $p(1,2) = p(2,2)$, with more such equalities if the table has more than two rows and/or more than two columns. But if one cell is a structural zero and the corresponding cell in the same column is not, then the usual null hypothesis makes no sense. Even under the null hypothesis, the cell probabilities cannot be the same owing to the zero count cells, so the row and column variables are not independent even under the null hypothesis. This is not as big a problem as it may first appear, however, because further thought reveals that interest would not lie in the usual null hypothesis anyway. In fact, there is no reason to even insist on the usual two-way structure at all.

The three possible outcomes in the problem of primary and secondary pneumonia infections of calves can be displayed alternatively as a one-way layout, or a 1×3 table with categories for no infection, only a primary infection, or a combination of a primary and a secondary infection. This new variable is the information-preserving composite endpoint [2]. Agresti considered testing if the probability of the primary infection is the same as the conditional probability of a secondary infection given that the calf got the primary infection [1]; that is, the null hypothesis can be written as $H_0 : \pi_{1+} = \pi_{11}/\pi_{1+}$. Tang and Tang [3] developed several exact unconditional methods on the basis of the above null hypothesis. The one-way layout can be used for the other examples as well, but two-way structures may be used with different parameterizations.

Instead of number of children in a household and number of smoking children in a household, for example, one could cross-classify by number of non-smoking children and number of smoking children. This would avoid the structural zero. Likewise, structural zeros could be avoided by cross-classifying by

the number of felonies committed on a given day in a given area without an arrest and the number of these felonies for which at least one suspect was arrested and charged with the felony in question. Finally, the number of nonserious infections experienced and the number of serious infections experienced for patients in a study could be tabulated with two-way structure and no structural zeros.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition, Wiley – Interscience.
- [2] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.
- [3] Tang, N.S. & Tang, M.L. (2002). Exact unconditional inference for risk ratio in a correlated 2×2 table with structural zero, *Biometrics* **58**, 972–980.
- [4] Yu, C. & Zelterman, D. (2002). Statistical inference for familial disease clusters, *Biometrics* **58**, 481–491.

VANCE W. BERGER AND JIALU ZHANG

Subjective Probability and Human Judgement

PETER AYTON

Volume 4, pp. 1960–1967

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Subjective Probability and Human Judgement

How good are people at judging probabilities? One early benchmark used for comparison was Bayes' theorem (*see Bayesian Belief Networks*). Bayes' theorem defines mathematically how probabilities should be combined and can be used as a normative theory of the way in which subjective probabilities representing degrees of belief attached to the truth of hypotheses should be revised in the light of new information. Bayes' theorem states that the posterior odds of the hypothesis being correct in the light of new information is a product of two elements: the *prior odds* of the hypothesis being correct before the information is observed and the *likelihood ratio* of the information, given that the hypothesis is correct or incorrect (*see Bayesian Statistics*).

In the 1960s, Ward Edwards and his colleagues conducted a number of studies using the book-bag and poker-chip paradigm. A typical experiment would involve two opaque bags. Each bag contained 100 colored poker-chips in different, but stated, proportions of red to blue. One – bag A contains 70 red chips and 30 blue, while the second – bag B contains 30 red chips and 70 blue. The experimenter first chooses one bag at random and then draws a series of chips from it. After each draw, the poker-chip is replaced and the bag well shaken before the next chip is drawn. The subject's task is to say how confident he/she is – in probability terms – that the chosen bag is bag A, containing predominantly red chips, or bag B, containing predominantly blue chips. As the bag was drawn randomly from two bags, our prior odds that we have bag A (or bag B) are 0.5/0.5. If we draw (say) a red chip, we know the likelihood of this is 0.7 if we have bag A and 0.3 if we have bag B. We thus multiply $0.5/0.5 \times 0.07/0.03$ to discover the posterior odds ($0.35/0.15 = 0.7/0.3$). The posterior odds computed after the first draw then become the prior odds for computing the impact of the second draw and the process repeated subsequently.

A crucial aspect of the logic of these studies is that the experimenter is able to say what the correct subjective probabilities should be for the participants by the simple expedient of calculating them using Bayes' theorem. All of the information required as inputs to Bayes' theorem is explicit and

unambiguous. Ironically, though this meant that the *subjectivity* of probability was not a part of the studies, in the sense that the experimenters assumed that they could objectively compute that the correct answer – which they would be able to assume – should be the same for all the participants faced with the same evidence.

The fact that the experimenter assumes he is able to calculate what the subjective probabilities *should* be for all of the participants was absolutely necessary if one was to be able to judge judgment by this method. However, it is also an indication of the artificiality of the task – and is at the root of the difficulties that were to emerge with interpreting the participants' behavior. The experiments conducted with this procedure produced a good deal of evidence that human judgment under these conditions is not well described by Bayes' theorem. Although participants' opinion revisions were proportional to the values calculated from Bayes' rule, they did not revise their opinions sufficiently in the light of the evidence, a phenomenon that was labeled *conservatism*. The clear suggestion was that human judgment was to this extent poor, although there was some debate as to the precise reason for this. It might be due to a failure to understand the impact of the evidence or to an inability to aggregate the assessments according to Bayes' theorem. Aside from any theoretical interest in these possibilities, there were practical implications of this debate. If people are good at assessing probabilities, but poor at combining them (as Edwards [5] suggested), then perhaps they could be helped; a relatively simple remedy would be to design a support system that took the human assessments and combined them using Bayes' theorem. However, if they were poor at assessing the component probabilities, then there would not be much point in devising systems to help them aggregate these.

Before any firm conclusions were reached as to the cause of conservatism, however, the research exploring the phenomenon fizzled out. The reasons for this seem to be twofold. One cause was the emergence of the heuristics and biases research and, in particular, the discovery of what Kahneman and **Tversky** [19] called *base-rate neglect*. Base-rate neglect is the exact opposite of conservatism – according to this account of judgment, people, far from being conservative about opinion revision, disregard prior odds and are *only* influenced by the likelihood ratio. Before this development occurred, however, there

was growing disquiet as to the validity of the book-bag experimental method as a basis for judging real-world judgment.

A number of studies had shown considerable variability in the amount of conservatism manifested according to various, quite subtle differences in the task set to participants. For example, the *diagnosticity* of the data seemed an important variable. Imagine, instead of our two bags with a 70/30 split in the proportions of blue and red poker-chips, the bags contained 49 red and 51 blue or 49 blue and 51 red chips. Clearly, two consecutive draws of a blue chip would not be very diagnostic as to which of the two bags we were sampling from. Experiments have shown that the more diagnostic the information, the more conservative is the subject. When information is very weakly diagnostic, as in our example, human probability revision, rather than being conservative, can be too extreme [29].

DuCharme and Peterson [4] argued that the fact that the information was restricted to one of two different possibilities (red chip or blue chip) meant that there were very few possible revisions that could be made. In the real world, information leading to revision of opinion does not have discrete values, but may more fairly be described as varying along a continuum. In an experimental study, DuCharme and Peterson used a hypothesis test consisting of the population of male heights and the population of female heights. The participants' task was to decide which population was being sampled from, based on the information given by randomly sampling heights from one of the populations. Using this task, DuCharme and Peterson found conservatism greatly reduced to half the level found in the more artificial tasks. They concluded that this was due to their participants' greater familiarity with the data-generating process underlying their task.

The argument concerning the validity of the conclusions from the book-bag and poker-chip paradigm was taken further by Winkler and Murphy [35]. Their article entitled 'Experiments in the laboratory and the real world' argued that the standard task differed in several crucial aspects from the real world. For example, the bits of evidence that are presented to experimental participants are conditionally independent; knowing one piece of information does not change the impact of the other. Producing one red chip from the urn and then replacing it does not affect the likelihood of drawing another red chip. However,

in real-world probability revision, this assumption often does not make sense.

For example, consider a problem posed by medical diagnosis. Loss of appetite is a symptom which, used in conjunction with other symptoms, can be useful for identifying the cause of certain illnesses. However, if I know that a patient is nauseous, I know that they are more likely (than in the absence of nausea) to experience loss of appetite. These two pieces of information, therefore, are not conditionally independent and so, when making my diagnosis, I should not revise my opinion on seeing the loss of appetite symptom as much as I might, before knowing about the nausea symptom, to diagnose diseases indicated by loss of appetite.

Winkler and Murphy argued that in many real-world situations lack of conditional independence of the information would render much of it redundant. In the book-bag task, participants may have been behaving much as they do in more familiar situations involving redundant information sources. Winkler and Murphy considered a range of other artificialities with this task and concluded that 'conservatism may be an artifact caused by dissimilarities between the laboratory and the real world'.

Heuristics and Biases

From the early 1970s, Kahneman and Tversky provided a formidable series of demonstrations of human judgmental error and linked these to the operation of a set of mental heuristics – mental rules of thumb – that they proposed the brain uses to simplify the process of judgment. For example, Tversky and Kahneman [30] claimed that human judgment is overconfident, ignores base rates, is insufficiently regressive, is influenced by arbitrary anchors, induces illusory correlations, and misconceives randomness. These foibles, they argued, indicated that the underlying judgment process was not normative (i.e., it did not compute probabilities using any kind of mental approximation to Bayes' theorem), but instead used simpler rules that were easier for the brain to implement quickly.

The idea, spelled out in [18], is that, due to limited mental processing capacity, strategies of simplification are required to reduce the complexity of judgment tasks and make them tractable for the kind of mind that people happen to have. Accordingly,

the principal reason for interest in judgmental biases was not merely that participants made errors, but that it supported the notion that people made use of relatively simple, but error-prone, heuristics for making judgments.

One such heuristic is *representativeness*. This heuristic determines how likely it is that an event is a member of a category according to how similar or typical the event is to the category. For example, people may judge the likelihood that a given individual is employed as a librarian by the extent to which the individual resembles a typical librarian. This may seem a reasonable strategy, but it neglects consideration of the relative prevalence of librarians. Tversky and Kahneman found that when base rates of different categories vary, judgments of the occupations of described people were correspondingly biased due to base-rate neglect. People using the representativeness heuristic for forecasting were employing a form of stereotyping in which similarity dominates other cues as a basis for judgment and decision-making.

In Kahneman and Tversky's [19] experiments demonstrating neglect of base rates, participants were found to ignore information concerning the prior probabilities of the hypotheses. For example, in one study, participants were presented with the following brief personal description of an individual called Jack:

Jack is a 45-year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

Half the participants were told that the description had been randomly drawn from a sample of 70 engineers and 30 lawyers, while the other half were told that the description was drawn from a sample of 30 engineers and 70 lawyers. Both groups were asked to estimate the probability that Jack was an engineer (or a lawyer). The mean estimates of the two groups of participants were only very slightly different (50 vs 55%). On the basis of this result and others, Kahneman and Tversky concluded that prior probabilities are largely ignored when individuating information was made available.

Although participants used the base rates when told to suppose that they had no information whatsoever about the individual (a 'null description'), when a description designed to be totally uninformative

with regard to the profession of an individual called Dick was presented, complete neglect of the base rates resulted.

Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

When confronted with this description, participants in both base rate groups gave median estimates of 50%. Kahneman and Tversky concluded that when no specific evidence is given, the base rates were properly utilized; but when worthless information is given, base rates were neglected.

Judgment by representativeness was also invoked by Tversky and Kahneman [32] to explain the *conjunction fallacy* whereby a conjunction of events is judged more likely than one of its constituents. This is a violation of a perfectly simple principle of probability logic: If A includes B, then the probability of B cannot exceed A. Nevertheless, participants who read a description of a woman called Linda who had a history of interest in liberal causes gave a higher likelihood to the possibility that she was a feminist bank clerk than to the possibility that she was a bank clerk – thereby violating the conjunction rule. Although it may seem unlikely that someone who had interests in liberal causes would be a bank clerk, but a bit more likely that she were a feminist bank clerk, all feminist bank clerks are of course bank clerks.

Another heuristic used for probabilistic judgment is *availability*. This heuristic is invoked when people estimate likelihood or relative frequency by the ease with which instances can be brought to mind. Instances of frequent events are typically easier to recall than instances of less frequent events, so availability will often be a valid cue for estimates of likelihood. However, availability is affected by factors other than likelihood. For example, recent events and emotionally salient events are easier to recollect. It is a common experience that the perceived riskiness of air travel rises in the immediate wake of an air disaster. Applications for earthquake insurance in California are apparently higher in the immediate wake of a major quake. Judgments made on the basis of availability then are vulnerable to bias whenever availability and likelihood are uncorrelated.

The *anchor and adjust* heuristic is used when people make estimates by starting from an initial value that is adjusted to yield a final value. The

claim is that adjustment is typically insufficient. For instance, one experimental task required participants to estimate various quantities stated in percentages (e.g., the percentage of African countries in the UN). Participants communicated their answers by using a spinner wheel showing numbers between 0 and 100. For each question, the wheel was spun and then participants were first asked whether the true answer was above or below this arbitrary value. They then gave their estimate of the actual value. Estimates were found to be considerably influenced by the initial (entirely random) starting point (cf. [34]).

The research into heuristics and biases provided a methodology, a very vivid explanatory framework and a strong suggestion that judgment is not as good as it might be. However, the idea that all of this should be taken for granted was denied by the proponents of the research some time ago. For example, Kahneman and Tversky [20] made clear that the main goal of the research was to understand the processes that produce both valid and invalid judgments. However, it soon became apparent that: 'although errors of judgment are but a method by which some cognitive processes are studied, the method has become a significant part of the message' [20, p. 494]. So, how should we regard human judgment?

There has been an enormous amount of discussion of Tversky and Kahneman's findings and claims. Researchers in the heuristics and biases tradition have sometimes generated shock and astonishment that people seem so bad at reasoning with probability despite the fact that we all live in an uncertain world. Not surprisingly, and as a consequence, the claims have been challenged. The basis of the challenges has varied. Some have questioned whether these demonstrations of biases in judgment apply merely to student samples or also to experts operating in their domain of expertise. Another argument is that the nature of the tasks set to participants gives a misleading perspective of their competence. A third argument is that the standards for the assessment of judgment are inappropriate.

Criticisms of Heuristics and Biases Research

Research following Tversky and Kahneman's original demonstration of base-rate neglect established that base rates might be attended to more (though

usually not sufficiently) if they were perceived as relevant [1], had a causal role [31], or were 'vivid' rather than 'pallid' in their impact on the decision-maker [27]. However, Gigerenzer, Hell, and Blank [10] have argued that the real reason for variations in base-rate neglect has nothing to do with any of these factors *per se*, but because the different tasks may, to varying degrees, encourage the subject to represent the problem as a Bayesian revision problem. They claimed that there are few inferences in real life that correspond directly to Bayesian revision where a known base-rate is revised on the basis of new information. Just because the experimenter assumes that he has defined a Bayesian revision problem does not imply that the subject will see it the same way. In particular, the participants may not take the base rate asserted by the experimenter as their subjective prior probability. In Kahneman and Tversky's original experiments, the descriptions were not actually randomly sampled (as the participants were told), but especially selected to be 'representative' of the professions. To the extent that the participants suspected that this was the case then they would be entitled to ignore the offered base rate and replace it with one of their own perception.

In an experiment, Gigerenzer et al. [10] found that when they let the participants experience the sampling themselves, base-rate neglect 'disappeared'. In the experiment, their participants could examine 10 pieces of paper, each marked lawyer or engineer in accord to the base rates. Participants then drew one of the pieces of paper from an urn and it was unfolded so they could read a description of an individual without being able to see the mark defining it as being of a lawyer or engineer. In these circumstances, participants clearly used the base rates in a proper fashion. However, in a replication of the verbal presentation where base rates were asserted, rather than sampled, Kahneman and Tversky's base-rate neglect was replicated.

In response to this, Kahneman and Tversky [21] argued that a fair summary of the research would be that explicitly presented base rates are generally underweighted, but not ignored. They have also pointed out that in Gigerenzer et al.'s experiment [10], participants who sampled the information themselves still produced judgments that deviated from the Bayesian solution in the direction predicted by representativeness. Plainly, then

representativeness is useful for predicting judgments. However, to the extent that base rates are not entirely ignored (as argued in an extensive review of the literature by Koehler [23]), the heuristic rationale for representativeness is limited. Recall that the original explanation for base-rate neglect was the operation of a simple heuristic that reduced the need for integration of multiple bits of information. If judgments in these experiments reflect base rates – even to a limited extent – it is hard to account for by the operation of the representativeness heuristic.

Tversky and Kahneman [32] reported evidence that violations of the conjunction rule largely disappeared when participants were requested to assess the relative frequency of events rather than the probability of a single event. Thus, instead of being asked about likelihood for a particular individual, participants were requested to assess how many people in a survey of 100 adult males had had heart attacks and then were asked to assess the number of those who were both over 55 years old *and* had had heart attacks. Only 25% of participants violated the conjunction rule by giving higher values to the latter than to the former. When asked about likelihoods for single events, it is typically the vast majority of participants who violate the rule. This difference in performance between frequency and single-event versions of the conjunction problem has been replicated several times since (cf. [8]).

Gigerenzer (e.g., [8], [9]) has suggested that people are naturally adapted to reasoning with information in the form of frequencies and that the conjunction fallacy ‘disappears’ if reasoning is in the form of frequencies for this reason. This suggests that the difficulties that people experience in solving probability problems can be reduced if the problems require participants to assess relative frequency for a class of events rather than the probability of a single event. Thus, it follows that if judgments were elicited with frequency formats there would be no biases. Kahneman and Tversky [21] disagree and argue that the frequency format serves to provide participants with a powerful cue to the relation of inclusion between sets that are explicitly compared, or evaluated in immediate succession. When the structure of the conjunction is made more apparent, then participants who appreciate the constraint supplied by the rule will be less likely to violate it. According to their account,

salient cues to set inclusion, not the frequency information *per se*, prompted participants to adjust their judgment.

To test this explanation, Kahneman and Tversky [21] reported a new variation of the conjunction problem experiment where participants made judgments of frequencies, but the cues to set inclusion were removed. They presented participants with the description of Linda and then asked their participants to suppose that there were 1000 women who fit the description. They then asked one group of participants to estimate how many of them would be bank tellers; a second independent group of participants were asked how many were bank tellers and active feminists; a third group made evaluations for both categories. As predicted, those participants who evaluated both categories mostly conformed to the conjunction rule. However, in a between-groups comparison of the other two groups, the estimates for ‘bank tellers and active feminists’ were found to be significantly higher than the estimates for bank tellers. Kahneman and Tversky argue that these results show that participants use the representativeness heuristic to generate their judgments and then edit their responses to respect class inclusion where they detect cues to that relation. Thus, they concluded that the key variable controlling adherence to the conjunction rule is not the relative frequency format *per se*, but the opportunity to detect the relation of class inclusion.

Other authors have investigated the impact of frequency information [7, 12, 13, 24] and concluded that it is not the frequency information *per se*, but the perceived relations between the entities that is affected by different versions of the problem, though this is rejected by Hoffrage, Gigerenzer, Krauss, and Martignon [15].

We need to understand more of the reasons underlying the limiting conditions of cognitive biases – how it is that seemingly inconsequential changes in the format of information can so radically alter the quality of judgment. Biases that can be cured so simply cannot be held to reveal fundamental characteristics of the processes of judgment. Gigerenzer’s group has recently developed an alternative program of research studying the *efficacy* of simple heuristics – rather than their association with biases (see **Heuristics: Fast and Frugal**). We consider the changing and disputed interpretations given to another claimed judgmental bias next.

Overconfidence

In the 1970s and 1980s, a considerable amount of evidence was marshaled for the view that people suffer from an overconfidence bias. Typical laboratory studies of calibration ask participants to answer question such as

‘Which is the world’s longest canal?’ (a) Panama
(b) Suez

Participants are informed that one of the answers is correct and are then required to indicate the answer that they think is correct and state how confident they are on a probability scale ranging from 50 to 100% (as one of the answers is always correct, 50% is the probability of guessing correctly). To be well calibrated, assessed probability should equal percentage correct over a number of assessments of equal probability. For example, if you assign a probability of 70% to each of 10 predictions, then you should get 7 of those predictions correct. Typically, however, people tend to give overconfident responses – their average confidence is higher than their proportion of correct answers. For a full review of this aspect of probabilistic judgment, see [25] and [14].

Overconfidence of judgments made under uncertainty is commonly found in calibration studies and has been recorded in the judgments of experts. For example, Christensen-Szalanski and Bushyhead [3] explored the validity of the probabilities given by physicians to diagnoses of pneumonia. They found that the probabilities were poorly calibrated and very overconfident; the proportion of patients who turned out to have pneumonia was far less than the probability statements implied. These authors had previously established that the physicians’ estimates of the probability of a patient having pneumonia were significantly correlated with their decision to give a patient a chest X ray and to assign a pneumonia diagnosis.

Wagenaar and Keren [33] found overconfidence in lawyers’ attempts to anticipate the outcome of court trials in which they represented one side. As they point out, it is inconceivable that the lawyers do not pay attention to the outcomes of trials in which they have participated, so why do they not learn to make well-calibrated judgments? Nonetheless, it is possible that the circumstances in which the lawyers, and other experts, make their judgments, and the circumstances in which they receive feedback, combine to

impede the proper monitoring of feedback necessary for the development of well-calibrated judgments. A consideration of the reports of well-calibrated experts supports this notion; they all appear to be cases where some explicit unambiguous quantification of uncertainty is initially made and the outcome feedback is prompt and unambiguous.

The most commonly cited example of well-calibrated judgments is weather forecasters’ estimates of the likelihood of precipitation [26], but there are a few other cases. Keren [22] found highly experienced tournament bridge players (but not experienced nontournament players) made well-calibrated forecasts of the likelihood that a contract, reached during the bidding phase, would be made, and Phillips [28] reports well-calibrated forecasts of horse races by bookmakers. In each of these three cases, the judgments made by the experts are precise numerical statements and the outcome feedback is unambiguous and received promptly and so can be easily compared with the initial forecast. Under these circumstances, the experts are unlikely to be insensitive to the experience of being surprised; there is very little scope for neglecting, or denying, any mismatch between forecast and outcome.

However, ‘ecological’ theorists (cf. [25]) claim that overconfidence is an artifact of the artificial experimental tasks and the nonrepresentative sampling of stimulus materials. Gigerenzer et al. [11] and Juslin [16] claim that individuals are well adapted to their environments and do not make biased judgments. Overconfidence is observed because the typical general knowledge quiz used in most experiments contains a disproportionate number of misleading items. These authors have found that when knowledge items are randomly sampled, the overconfidence phenomenon disappears. For example, Gigerenzer et al. [11] presented their participants with items generated with random pairs of the German cities with more than 100 000 inhabitants and asked them to select the biggest and indicate their confidence they had done so correctly. With this randomly sampled set of items, there was no overconfidence.

Moreover, with conventional general knowledge quizzes, participants are aware of how well they are likely to perform overall. Gigerenzer et al. [11] found that participants are really quite accurate at indicating *the proportion of items* that they have correctly answered. Such quizzes are representative of general knowledge quizzes experienced in the

past. Thus, even when they appear overconfident with their answers to the individual items, participants are not overconfident about their performance on the same items as a set. Note that this observation is consistent with Gigerenzer's claim that though people may be poor at representing and so reasoning with probabilities about single events they can effectively infer probabilities when represented as frequencies.

Juslin et al. [17] report a **meta-analysis** comparing 35 studies, where items were randomly selected from a defined domain, with 95 studies where items were selected by experimenters. While overconfidence was evident for selected items, it was close to zero for randomly sampled items, which suggests that overconfidence is not simply a ubiquitous cognitive bias. This analysis suggests that the appearance of overconfidence may be an illusion created by research and not a cognitive failure by respondents.

Moreover, in cases of judgments of repeated events (weather forecasters, horse race bookmakers, tournament bridge players), experts make well-calibrated forecasts. In these cases, respondents might be identifying relative frequencies for sets of similar events rather than judging likelihood for individual events. And, if we compare studies of the calibration of probability assessments concerning individual events (e.g., [36]) with those where subjective assessments have been made for repetitive predictions of events [26], we observe that relatively poor calibration has been observed in the former, whereas relatively good calibration has been observed in the latter.

Another idea relevant to the interpretation of the evidence of overconfidence comes from Erev, Wallsten, and Budescu [6], who have suggested that overconfidence may, to some degree, reflect an underlying stochastic component of judgment. Any degree of error variance in judgment would create a regression that appears as overconfidence in the typical calibration analysis of judgment. When any two variables are not perfectly correlated – and confidence and accuracy are not perfectly correlated – there will be a regression effect. So, it is that a sample of the (adult) sons of very tall fathers will, on average, be shorter than their fathers, and, *at the same time*, a sample of the fathers of very tall sons will, on average, be shorter than their sons.

Exploring this idea, Budescu, Erev, and Wallsten [2] presented a generalization of the results from

the Erev et al. [6] article, which shows that overconfidence and its apparent opposite, underconfidence, can be observed *simultaneously* in one study, depending upon whether probabilistic judgments are analyzed by conditionalizing accuracy as a function of confidence (the usual method showing overconfidence) or vice versa.

Conclusions

Although there has been a substantial amassing of evidence for the view that humans are inept at dealing with uncertainty using judged – subjective – probability, we also find evidence for a counterargument. It seems that disparities with basic requirements of probability theory can be observed when people are asked to make judgments of probability as a measure of propensity or strength of belief. The counterargument proposes that people may be very much better at reasoning under uncertainty than this research suggests when they are presented with tasks in a manner that permits them to conceive of probability in frequentist terms. This debate is currently unresolved and highly contentious. Nevertheless, for those with the hope of using subjective probabilities as inputs into decision support systems, we hope we have gone some way toward demonstrating that human judgments of uncertainty are worth considering as a valuable resource, rather than as objects to be regarded with suspicion or disdain.

References

- [1] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgements, *Acta Psychologica* **44**, 211–233.
- [2] Budescu, D.V., Erev, I. & Wallsten, T.S. (1997). On the importance of random error in the study of probability judgment: Part I: new theoretical developments, *Journal of Behavioral Decision Making* **10**, 157–171.
- [3] Christensen-Szalanski, J.J.J. & Bushyhead, J.B. (1981). Physicians use of probabilistic information in a real clinical setting, *Journal of Experimental Psychology, Human Perception and Performance* **7**, 928–935.
- [4] DuCharme, W.M. & Peterson, C.R. (1968). Intuitive inference about normally distributed populations, *Journal of Experimental Psychology* **78**, 269–275.
- [5] Edwards, W. (1968). Conservatism in human information processing, in *Formal Representation of Human Judgment*, B. Kleinmuntz, ed., Wiley, New York.
- [6] Erev, I., Wallsten, T.S. & Budescu, D.V. (1994). Simultaneous over-and underconfidence: the role of error in judgment processes, *Psychological Review* **101**, 519–528.

- [7] Evans, J.St.B.T., Handley, S.J., Perham, N., Over, D.E. & Thompson, V.A. (2000). Frequency versus probability formats in statistical word problems, *Cognition* **77**, 197–213.
- [8] Gigerenzer, G. (1994). Why the distinction between single event probabilities and frequencies is important for Psychology and vice-versa, in *Subjective Probability*, G. Wright & P. Ayton, eds, Wiley, Chichester.
- [9] Gigerenzer, G. (1996). On narrow norms and vague heuristics: a rebuttal to Kahneman and Tversky, *Psychological Review* **103**, 592–596.
- [10] Gigerenzer, G., Hell, W. & Blank, H. (1988). Presentation and content: the use of base rates as a continuous variable, *Journal of Experimental Psychology: Human Perception and Performance* **14**, 513–525.
- [11] Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence, *Psychological Review* **98**, 506–528.
- [12] Girotto, V. & Gonzales, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form, *Cognition* **78**, 247–276.
- [13] Girotto, V. & Gonzales, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon, *Cognition* **84**, 353–359.
- [14] Harvey, N. (1998). Confidence in judgment, *Trends in Cognitive Sciences* **1**, 78–82.
- [15] Hoffrage, U., Gigerenzer, G., Krauss, S. & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not, *Cognition* **84**, 343–352.
- [16] Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items, *Organizational Behavior and Human Decision Processes* **57**, 226–246.
- [17] Juslin, P., Winman, A. & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect, *Psychological Review* **107**, 384–396.
- [18] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press.
- [19] Kahneman, D. & Tversky, A. (1973). On the psychology of prediction, *Psychological Review* **80**, 237–251.
- [20] Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions, in *Judgement Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press.
- [21] Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions: a reply to Gigerenzer's critique, *Psychological Review* **103**, 582–591.
- [22] Keren, G.B. (1987). Facing uncertainty in the game of bridge: a calibration study, *Organizational Behavior and Human Decision Processes* **39**, 98–114.
- [23] Koehler, J.J.J. (1995). The base-rate fallacy reconsidered - descriptive, normative, and methodological challenges, *Behavioral and Brain Sciences* **19**, 1–55.
- [24] Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations, *Organizational Behavior and Human Decision Processes* **82**, 217–236.
- [25] McClelland, A.G.R. & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–1994, in *Subjective Probability*, G. Wright & P. Ayton, eds, Wiley, New York.
- [26] Murphy, A.H. & Winkler, R.L. (1984). Probability forecasting in meteorology, *Journal of the American Statistical Association* **79**, 489–500.
- [27] Nisbett, R. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings*, Prentice Hall, Englewood Cliffs.
- [28] Phillips, L.D. (1987). On the adequacy of judgmental probability forecasts, in *Judgmental Forecasting*, G. Wright & P. Ayton, eds, Wiley, Chichester.
- [29] Phillips, L.D. & Edwards, W. (1966). Conservatism in simple probability inference tasks, *Journal of Experimental Psychology* **72**, 346–357.
- [30] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases, *Science* **185**, 1124–1131.
- [31] Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates, in *Judgement Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press.
- [32] Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment, *Psychological Review* **90**, 293–315.
- [33] Wagenaar, W.A. & Keren, G.B. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts, in *Intelligent Decision Support In Process Environments*, E. Hollnagel, G. Mancini & D.D. Woods, eds, Springer-Verlag, Berlin.
- [34] Wilson, T.D., Houston, C.E., Etling, K.M. & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents, *Journal of Experimental Psychology: General* **125**, 387–402.
- [35] Winkler, R.L. & Murphy, A.M. (1973). Experiments in the laboratory and the real world, *Organizational Behavior and Human Performance* **20**, 252–270.
- [36] Wright, G. & Ayton, P. (1992). Judgmental probability forecasting in the immediate and medium term, *Organizational Behavior and Human Decision Processes* **51**, 344–363.

Further Reading

Savage, L. (1954). *The Foundations of Statistics*, Wiley, New York.

(See also **Heuristics: Fast and Frugal**)

PETER AYTON

Summary Measure Analysis of Longitudinal Data

BRIAN S. EVERITT

Volume 4, pp. 1968–1969

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Summary Measure Analysis of Longitudinal Data

There are a variety of approaches to the analysis of **longitudinal data**, including **linear mixed effects models** and **generalized estimating equations**. But many investigators may prefer (initially at least) to use a less complex procedure. One that may fit the bill is summary measure analysis, the essential feature of which is the reduction of the repeated response measurements available on each individual in the study, to a single number that is considered to capture an essential feature of the individual's response over time. In this way, the multivariate nature of the repeated observations is transformed to univariate. The approach has been in use for many years – see [3].

Choosing a Summary Measure

The most important consideration when applying a summary measure analysis is the choice of a suitable summary measure, a choice that needs to be made before any data are collected. The measure chosen needs to be relevant to the particular questions of interest in the study and in the broader scientific context in which the study takes place. A wide range of summary measures have been proposed as we see in Table 1. In [1], it is suggested that the average response over time is often likely to be the most

relevant, particularly in *intervention studies* such as **clinical trials**.

Having chosen a suitable summary measure, the analysis of the longitudinal data becomes relatively straightforward. If two groups are being compared and normality of the summary measure is thought to be a valid assumption, then an independent samples *t* Test can be used to test for a group difference, or (preferably) a **confidence interval** for this difference can be constructed in the usual way. A one-way **analysis of variance** can be applied when there are more than two groups if again the necessary assumptions of normality and homogeneity hold. If the distributional properties of the selected summary measure are such that normality seems difficult to justify, then nonparametric analogues of these procedures might be used (*see Catalogue of Parametric Tests; Distribution-free Inference, an Overview*).

An Example of Summary Measure Analysis

The summary measure approach can be illustrated using the data shown in Table 2 that arise from a study of alcohol dependence. Two groups of subjects, one with severe dependence and one with moderate dependence on alcohol, had their salsolinol excretion levels (in millimoles) recorded on four consecutive days. (Salsolinol is an alkaloid with a structure similar to heroin.)

Using the mean of the four measurements available for each subject as the summary measure followed by the application of a *t* Test and the construction of a **confidence interval** leads to the results

Table 1 Possible summary measures (taken from [2])

Type of data	Question of interest	Summary measure
Peaked	Is overall value of outcome variable the same in different groups?	Overall mean (equal time intervals) or area under curve (unequal intervals)
Peaked	Is maximum (minimum) response different between groups?	Maximum (minimum) value
Peaked	Is time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is rate of change of outcome different between groups?	Regression coefficient
Growth	Is eventual value of outcome different between groups?	Final value of outcome or difference between last and first values or percentage change between first and last values
Growth	Is response in one group delayed relative to the other?	Time to reach a particular value (e.g., a fixed percentage of baseline)

2 Summary Measure Analysis of Longitudinal Data

Table 2 Salsolinol excretion data

Subject	Day			
	1	2	3	4
Group 1 (Moderate dependence)				
1	0.33	0.70	2.33	3.20
2	5.30	0.90	1.80	0.70
3	2.50	2.10	1.12	1.01
4	0.98	0.32	3.91	0.66
5	0.39	0.69	0.73	3.86
6	0.31	6.34	0.63	3.86
Group 2 (Severe dependence)				
7	0.64	0.70	1.00	1.40
8	0.73	1.85	3.60	2.60
9	0.70	4.20	7.30	5.40
10	0.40	1.60	1.40	7.10
11	2.50	1.30	0.70	0.70
12	7.80	1.20	2.60	1.80
13	1.90	1.30	4.40	2.80
14	0.50	0.40	1.10	8.10

Table 3 Results from using the mean as a summary measure for the data in Table 2

	Moderate	Severe
Mean	1.80	2.49
sd	0.60	1.09
<i>n</i>	6	8
$t = -1.40$, $df = 12$, $p = 0.19$, 95% CI: $[-1.77, 0.39]$		

shown in Table 3. There is no evidence of a group difference in salsolinol excretion levels.

A possible alternative to the use of the mean as summary measure is the maximum excretion level recorded over the four days. Testing the null hypothesis that the measure has the same median in both populations using a Mann-Whitney (see **Wilcoxon–Mann–Whitney Test**) test results in a test statistic of 36 and associated P value of 0.28. Again,

there is no evidence of a difference in salsolinol excretion levels in the two groups.

Problems with the Summary Measure Approach

In some situations, it may be impossible to identify a suitable summary measure and where interest centers on assessing details of how a response changes over time and how this change differs between groups, then summary measure analysis has little to offer. The summary measure approach to the analysis of longitudinal data can accommodate **missing data** by simply using only the available observations in its calculation, but the implicit assumption is that values are missing completely at random. Consequently, in, say, clinical trials in which a substantial number of participants drop out before the scheduled end of the trial, the summary measure approach is probably not to be recommended.

References

- [1] Frison, L. & Pocock, S.J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **11**, 1685–1704.
- [2] Matthews, J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1989). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.
- [3] Oldham, P.D. (1962). A note on the analysis of repeated measurements of the same subjects, *Journal of Chronic Disorders* **15**, 969–977.

(See also **Repeated Measures Analysis of Variance**)

BRIAN S. EVERITT

Survey Questionnaire Design

ROGER TOURANGEAU

Volume 4, pp. 1969–1977

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Survey Questionnaire Design

Introduction

Survey statisticians have long distinguished two major sources of error in survey estimates – sampling error and measurement error. Sampling error arises because the survey does not collect data from the entire population and the characteristics of the sample may not perfectly match those of the population from which it was drawn (*see Survey Sampling Procedures*). Measurement error arises because the information collected in the survey differs from the true values for the variables of interest. The discrepancies between survey reports and true values can arise because the survey questions measure the wrong thing or because they measure the right thing but do it imperfectly. For example, the survey designers may want to measure unemployment, and, in fact, most developed countries conduct regular surveys to monitor employment and unemployment rates. Measuring unemployment can be tricky. How does one classify workers with a job but on extended sick leave? A major problem is asking the right questions, the questions that are needed to classify each respondent correctly. Another major problem is inaccurate reporting. Even if the questions represent the concept of interest, respondents may still not report the right information because they misunderstand the questions or they do not know all the relevant facts. For example, they may not know about the job search activities of other family members. The goal of survey questionnaire design is simple – it is to reduce such measurement errors to a minimum, subject to whatever cost constraints apply to the survey.

There are two basic methods for attaining this goal. First, questionnaire designers attempt to write questions that follow well-established principles for survey questions. Texts with guidelines for writing survey questions have been around for at least fifty years, appearing at about the same time as the first texts on survey sampling (for an early example, *see* [9]). Initially, these texts offered guidelines that were based on the experiences of the authors, but over the years a large base of methodological research has accumulated and this work has provided

an empirical foundation for the questionnaire design guidelines (*see*, e.g. [14]). In addition, over the last twenty years or so, survey researchers have drawn more systematically on research in cognitive psychology to understand how respondents answer questions in surveys. This work, summarized in [15] and [17], has provided a theoretical grounding for the principles of questionnaire design. Traditionally, survey researchers have thought of writing survey questions as more on an art than a science, but, increasingly, because of the empirical and theoretical advances of the last twenty years, survey researchers have begun referring to the science of asking questions [13].

Aside from writing good questions in the first place, the other strategy for minimizing measurement error is to test survey questions and cull out or improve questions that do not seem to yield accurate information. There are several tools questionnaire designers use in developing and testing survey questions. These include both pretesting methods such as cognitive interviews or pilot tests used before the survey questionnaire is fielded, and methods that can be applied as the survey is carried out such as recontacting some of the respondents and asking some questions a second time.

The Questionnaire Design Process

Questionnaire design encompasses four major activities. The first step often consists of library research, in which the questionnaire designers look for existing items with desirable measurement properties. There are few general compilations of existing survey items. As a result, survey researchers generally rely on subject matter experts and earlier surveys on the same topic as sources for existing questions.

The next step is to assemble the questions into a draft questionnaire; the draft is usually a blend of both existing and newly written items. The initial draft often undergoes some sort of expert review. Questionnaire design experts may review the questionnaire to make sure that the questions adhere to questionnaire design principles and that they are easy for the interviewers to administer and for the respondents to understand and answer. Subject matter experts may review the draft to make sure that the survey will yield all the information needed in the analysis and that the questions correspond to the concepts of interest. For instance, the concept of unemployment

2 Survey Questionnaire Design

involves both not having a job and wanting one; the survey questions must adequately cover both aspects of the concept. Subject matter experts are typically in the best position to decide whether the survey questions as drafted will meet the analytical requirements of the survey. When the questionnaire includes questions about a new topic, the questionnaire designers may also conduct one or more focus groups, a procedure described in more detail below, to discover how members of the survey population think about the topic of interest and the words and phrases they use in talking about it. Respondents are more likely to answer the questions accurately when the questions match their experiences and circumstances, and when they use familiar terminology.

Rarely does a survey questionnaire go into the field without having been pretested in some way. Thus, the third step in the questionnaire design process typically involves testing, evaluating, and revising the questionnaire prior to conducting the survey. Two types of pretesting are commonly used. The first is called *cognitive interviewing*. Cognitive interviews are generally conducted in a centralized laboratory setting rather than in the field. The purpose of these interviews is to discover how respondents answer the questions and whether they encounter any cognitive difficulties in formulating their answers. The cognitive interviewers administer a draft of the survey questions. They may encourage the respondents to think out loud as they answer the questions or they may administer follow-up probes designed to explore potential problems with the draft survey items. The second type of pretest is a pilot test, or a small-scale version of the main study. Pretest interviewers may interview 50 to 100 respondents. The size of the pretest sample often reflects the size and budget of the main survey. The survey designers may evaluate this trial run of the questionnaire by drawing on several sources of information. Often, they examine the pilot test responses, looking for items with low variances or high rates of missing data. In addition, the questionnaire designers may conduct a ‘debriefing’ with the pilot test interviewers, eliciting their input on such matters as the items that seemed to cause problems for the interviewers or the respondents. Pilot tests sometimes incorporate experiments comparing two or more methods for asking the questions and, in such cases, the evaluation of the pretest results will include an assessment of the experimental results. Or the pilot test may include recording or monitoring the

interviews. The point of such monitoring is to detect items that interviewers often do not read as written or items that elicit frequent requests for clarification from the respondents. On the basis of the results of cognitive or pilot test interviews, the draft questionnaire may be revised, often substantially. The pilot test may reveal other information about the questionnaire, such as the average time needed to administer the questions, which may also feed into the evaluation and revision of the questions.

The final step is to administer questions in the real survey. The evaluation of the questions does not necessarily stop when the questionnaire is fielded, because the survey itself may collect data that are useful for evaluating the questions. For example, some education surveys collect data from both the students and their parents, and the analysis can assess the degree that the information from the two sources agrees upon. Low levels of agreement between sources would suggest high levels of measurement error in one or both sources. Similarly, surveys on health care may collect information both from the patient and from medical records, allowing an assessment of the accuracy of the survey responses. Many surveys also recontact some of the respondents to make sure the interviewers have not fabricated the data; these ‘validation’ interviews may readminister some of the original questions, allowing an assessment of the reliability of the questions. Thus, the final step in the development of the questionnaire is sometimes an after-the-fact assessment of the questions, based on the results of the survey.

Tools for Testing and Evaluating Survey Questions

Our overview of the questionnaire design process mentioned a number of tools survey researchers use in developing and testing survey questions. This section describes these tools – expert reviews, focus groups, cognitive testing, pilot tests, and split-ballot experiments – in more detail.

Expert Reviews. Expert reviews refer to two distinct activities. One type of review is carried out by substantive experts or even the eventual analysts of the survey data. The purpose of these substantive reviews is to ensure that the questionnaire collects all the information needed to meet the analytic objectives

of the survey. The other type of review features questionnaire design experts, who review the wording of the questions, the response format and the particular response options offered, the order of the questions, the instructions to interviewers for administering the questionnaire, and the navigational instructions (e.g., 'If yes, please go to Section B'). Empirical evaluations [10] suggest that questionnaire design experts often point out a large number of problems with draft questionnaires.

Sometimes, the experts employ formal checklists of potential problems with questions. Several checklists are available. Lessler and Forsyth [8], for example, present a list of 25 types of potential problems with questions. Such checklists are generally derived from a cognitive analysis of the survey response process and the problems that can arise during that process (see, e.g., [15] and [17]). The checklists often distinguish various problems in comprehension of the questions, the recall of information needed to answer the questions, the use of judgment and estimation strategies, and the reporting of the answer. One set of researchers [6] has even developed a computer program that diagnoses 12 major problems with survey questions, most of them involving comprehension issues, thus providing an automated, if preliminary, expert appraisal. To illustrate the types of problems included in the checklists, here are the 12 detected by Graesser and his colleagues' program:

1. complex syntax,
2. working memory overload,
3. vague or ambiguous noun phrase,
4. unfamiliar technical term,
5. vague or imprecise predicate or relative term,
6. misleading or incorrect presupposition,
7. unclear question category,
8. amalgamation of more than one question category,
9. mismatch between the question category and the answer option,
10. difficulty in accessing (that is, recalling) information,
11. respondent unlikely to know answer, and
12. unclear question purpose.

Focus Groups. Before they start writing the survey questions, questionnaire designers often listen to volunteers discussing the topic of the survey. These focus group discussions typically include 6 to 10

members of the survey population and a moderator who leads the discussion. Questionnaire designers often use focus groups in the early stages of the questionnaire design process to learn more about the survey topic and to discover how the members of the survey population think and talk about it (*see Focus Group Techniques*).

Suppose, for example, one is developing a questionnaire on medical care. It is useful to know what kinds of doctors and medical plans respondents actually use and it is also useful to know whether they are aware of the differences between HMOs, other types of managed care plans, and fee-for-service plans. In addition, it is helpful to know what terminology they use in describing each sort of plan and how they describe different types of medical visits. The more that the questions fit the situations of the respondents, the easier it will be for the respondents to answer them. Similarly, the more closely the questions mirror the terminology that the respondents use in everyday life, the more likely it is that the respondents will understand the questions as intended.

Focus groups can be an efficient method for getting information from several people in a short period of time, but the method does have several limitations. Those who take part in focus groups are typically volunteers and they may or may not accurately represent the survey population. In addition, the number of participants in a focus group may be a poor guide to the amount of information produced by the discussion. No matter how good the moderator is, the discussion often reflects the views of the most articulate participants. In addition, the discussion may veer off onto some tangent, reflecting the group dynamic rather than the considered views of the participants. Finally, the conclusions from a focus group discussion are often simply the impressions of the observers; they may be unreliable and subject to the biases of those conducting the discussions.

Cognitive Interviewing. Cognitive interviewing is a family of methods designed to reveal the strategies that respondents use in answering survey questions. It is descended from a technique called 'protocol analysis', invented by Herbert Simon and his colleagues (*see, e.g., [5]*). Its purpose is to explore how people deal with higher-level cognitive problems, like solving chess problems or proving algebraic theorems. Simon asked his subjects to think aloud as they worked on such problems and recorded what

they said. These verbalizations were the ‘protocols’ that Simon and his colleagues used in testing their hypotheses about problem solving. The term ‘cognitive interviewing’ is used somewhat more broadly to cover a range of procedures, including:

1. concurrent protocols, in which respondents verbalize their thoughts while they answer a question;
2. retrospective protocols, in which they describe how they arrived at their answers after they provide them;
3. confidence ratings, in which they rate their confidence in their answers;
4. definitions of key terms, in which respondents are asked to define terms in the questions;
5. paraphrasing, in which respondents restate the question in their own words, and
6. follow-up probes, in which respondents answer questions designed to reveal their response strategies.

This list is adopted from a longer one found in [7]. Like focus groups, cognitive interviews are typically conducted with paid volunteers. The questionnaires for cognitive interviews include both draft survey questions and prescribed probes designed to reveal how the respondents understood the questions and arrived at their answers. The interviewers may also ask respondents to think aloud as they answer some or all of the questions. The interviewers generally are not field interviewers but have received special training in cognitive interviewing.

Although cognitive interviewing has become a very popular technique, it is not very well standardized. Different organizations emphasize different methods and no two interviewers conduct cognitive interviews in exactly the same way. Cognitive interviews share two of the main drawbacks with focus groups. First, the samples of respondents are typically volunteers so the results may not be representative of the survey population. Second, the conclusions from the interviews are often based on the impressions of the interviewers rather than objective data such as the frequency with which specific problems with an item are encountered.

Pilot Tests. Pilot tests or field tests of a questionnaire are mini versions of the actual survey conducted by field interviewers under realistic survey conditions. The pilot tests use the same mode of data

collection as the main survey. For instance, if the main survey is done over the telephone, then the pilot test is done by the telephone as well. Pilot tests have some important advantages over focus groups and cognitive interviews; they often use probability samples of the survey population and they are done using the same procedures as the main survey. As a result, they can provide information about the data collection and sampling procedures – Are they practical? Do the interviews take longer than planned to complete? – as well as information about the draft questionnaire.

Pilot tests typically yield two main types of information about the survey questionnaire. One type consists of the feedback from the pilot test interviewers. The reactions of the interviewers are often obtained in an interviewer debriefing, where some or all of the field test interviewers meet for a discussion of their experiences with the questionnaire during the pilot study. The questionnaire designers attend these debriefing sessions to hear the interviewers present their views about questions that do not seem to be working and other problems they experienced during the field test. The second type of information from field test is the data – the survey responses themselves. Analysis of the pretest data may produce various signs diagnostic of questionnaire problems, such as items with high rates of missing data, out-of-range values, or inconsistencies with other questions. Such items may be dropped or rewritten.

Sometimes pilot tests gather additional quantitative data that is useful for evaluating the questions. These data are derived from monitoring or recording the pilot test interviews and then coding the interchanges between the interviewers and respondents. Several schemes have been developed for systematically coding these interactions. Fowler and Cannel [4] have proposed and applied a simple scheme for assessing survey questions. They argue that coders should record for each question whether the interviewer read the question exactly as worded, with minor changes, or with major changes that altered the meaning of the questions. In addition, the coders should record whether the respondent interrupted the interviewer before he or she had finished reading the question, asked for clarification of the question, and gave an adequate answer, a don’t know or refusal response, or an answer that required further probing from the interviewer. Once the interviews have been coded, the questionnaire designers can examine

a variety of statistics for each item, including the percentage of times the interviewers departed from the verbatim text in administering the item, the percentage of respondents who asked for clarification of the question, and the percentage of times the interviewer had to ask additional questions to obtain an acceptable answer. All three behaviors are considered signs of poorly written questions.

Relative to expert reviews, focus groups, and cognitive testing, field tests, especially when supplemented by behavior coding, yields data that are objective, quantitative, and replicable.

Split-ballot Experiments. A final method used in developing and testing questionnaires is the split-ballot experiment. In a split-ballot experiment, random subsamples of the respondents receive different versions of the questions or different methods of data collection, for example, self-administered questions versus questions administered by interviewers, or both. Such experiments are generally conducted as part of the development of questionnaires and procedures for large-scale surveys, and they may be embedded in a pilot study for such surveys. A discussion of the design issues raised by such experiments can be found in [16]; examples can be found in [3] and [16].

Split-ballot experiments have the great virtue that they can clearly show what features of the questions or data collection procedures affect the answers. For example, the experiment can compare different wordings of the questions or different question orders. But the fact that two versions of the questionnaire produce different results does not necessarily resolve the question of which version is the right one to use. Experiments can produce more definitive results when they also collect some external validation data that can be used to measure the accuracy of the responses. For example, in a study of medical care, the researchers can compare the survey responses to medical records, thereby providing some basis for deciding which version of the questions yielded the more accurate information. In other cases, the results of a methodological experiment may be unambiguous even in the absence of external information. Sometimes there is a strong *a priori* reason for thinking that reporting errors follow a particular pattern or direction. For example, respondents are likely to underreport embarrassing or illegal behaviors, like illicit drug use. Thus, a strong reason for thinking

that a shift in a specific direction (such as higher levels of reported drug use) represents an increase in accuracy.

The major drawback to methodological experiments is their complexity and expense. Detecting even moderate differences between experimental groups may require substantial numbers of respondents. In addition, such experiments add to the burden on the survey designers, requiring them to develop multiple questionnaires or data collection protocols rather than just one. The additional time and expense of this effort may exceed the budget for questionnaire development or delay the schedule for the main survey too long. For these reasons, split-ballot experiments are not a routine part of the questionnaire design process.

Combining the Tools. For any particular survey, the questionnaire design effort is likely to employ several of these tools. The early stages of the process are likely to rely on relatively fast and inexpensive methods such as expert reviews. Depending on how much of the questionnaire asks about a new topic, the questionnaire designers may also conduct one or more focus groups. If the survey is fielding a substantial number of new items, the researchers are likely to conduct one or more rounds of cognitive testing and a pilot test prior to the main survey. Typically, the researchers analyze the pilot study data and carry out a debriefing of the pilot test interviewers. They may supplement this with the coding and analysis of data on the exchanges between respondents and interviewers. If there are major unresolved questions about how the draft questions should be organized or worded, the survey designers may conduct an experiment to settle them.

The particular combination of methods used in developing and testing the survey questionnaire for a given survey will reflect several considerations, including the relative strengths and weaknesses of the different methods, the amount of time and money available for the questionnaire design effort, and the issues that concern the survey designers the most. The different questionnaire design tools yield different kinds of information. Cognitive interviews, for example, provide information about respondents' cognitive difficulties with the questions but are less useful for deciding how easy it will be for interviewers to administer the questions in the field. Expert reviews are a good, all-purpose tool but they yield

educated guesses rather than objective data on how the questions are likely to work. Pilot tests are essential if there are concerns about the length of the questionnaire or other practical issues that only can be addressed by a dry run of the questionnaire and data collection procedures under realistic field conditions. If the questionnaire requires the use of complicated response aids or new software for administering the questions, a field test and interviewer debriefing are likely to be deemed essential. And, of course, decisions about what tools to use and how many rounds of the testing to carry out are likely to reflect the overall survey budget, the amount of prior experience with this or similar questionnaires, and other factors related to cost and schedule.

Standards for Evaluating Questions

Although the goal for questionnaire design is straightforward in principle – to minimize survey error and cost – as a practical matter, the researchers may have to examine various indirect measures of cost or quality. In theory, the most relevant standards for judging the questions are their reliability and validity; but such direct measures of error are often not available and the researchers are forced to fall back on such indirect indicators of measurement error as the results of cognitive interviews. This section takes a more systematic look at the criteria questionnaire designers use in judging survey questions.

Content Standards. One important nonstatistical criterion that the questionnaire must meet is whether it covers all the topics of interest and yields all the variables needed in the analysis. It does not matter much whether the questions elicit accurate information if it is not the right information. Although it might seem a simple matter to make sure the survey includes the questions needed to meet the analytical objectives, there is often disagreement about the best strategy for measuring a given concept and there are always limits on the time or space available in a questionnaire. Thus, there may be deliberate compromises between full coverage of a particular topic of interest and cost. To keep the questionnaire to a manageable length, the designers may include a subset of the items from a standard battery in place of the full battery or they may explore certain topics superficially rather than in the depth they would prefer.

Statistical Standards: Validity and reliability. Of course, the most fundamental standards for a question are whether it yields consistent and accurate information. Reliability and validity (*see* **Validity Theory and Applications**) are the chief statistical measures of these properties.

The simplest mathematical model for a survey response treats it as consisting of components – a true score and an error:

$$Y_{it} = \mu_i + \varepsilon_{it}, \quad (1)$$

in which Y_{it} refers to the reported value for respondent i on occasion t , μ_i refers to the true score for that respondent, and ε_{it} to the error for the respondent on occasion t (*see* **Measurement: Overview**). The true score is the actual value for the respondent on the variable of interest and the error is just the discrepancy between the true score and the reported value. The idea of a true score makes more intuitive sense when the variable involves some readily verifiable fact or behavior – say, the number of times the respondent visited a doctor in the past month. Still, many survey researchers find the concept useful even for subjective variables, for example, how much the respondent favors or opposes some policy. In such cases, the true score is defined as the mean across the hypothetical population of measures for the concept that are on the same scale (*see*, e.g., [1]).

Several assumptions are often made about the errors. The simplest model assumes first that, for any given respondent, the expected value of the errors is zero and, second, that the correlation between the errors for any two respondents or between those for the same respondent on any two occasions is zero. The validity of the item is usually defined as the correlation between Y_{it} and μ_i . The reliability is defined as the correlation between Y_{it} and $Y_{it'}$, where t and t' represent two different occasions. Under the simplest model, it is easy to show that validity (V) is just:

$$\begin{aligned} V &= \frac{\text{Cov}(Y, \mu)}{[\text{Var}(Y)\text{Var}(\mu)]^{1/2}} \\ &= \frac{\text{Var}(\mu)}{[\text{Var}(Y)\text{Var}(\mu)]^{1/2}} \\ &= \frac{\text{Var}(\mu)^{1/2}}{\text{Var}(Y)^{1/2}} \end{aligned} \quad (2)$$

in which $\text{Cov}(Y, \mu)$ is the covariance between the observed values and the true scores and $\text{Var}(Y)$ and

$\text{Var}(\mu)$ are their variances. Under this model, the validity is just the square of the reliability.

As a practical matter, the validity of a survey item is often estimated by measuring the correlation between the survey reports and some external ‘gold standard,’ such as administrative records or some other measure of the variable of interest that is assumed to be error-free or nearly error-free. Reliability is estimated in one of two ways. For simple variables derived from a single item, the item may be administered to the respondent a second time in a reinterview. Rather than assessing the reliability by calculating the correlation between the two responses, survey researchers often calculate the gross discrepancy rate – the proportion of respondents classified differently in the original interview and the reinterview. This approach is particularly common when the survey report yields a simple categorical variable such as whether the respondent is employed or not. The gross discrepancy rate is a measure of *unreliability* rather than reliability. For variables derived from multi-item batteries, the average correlation among the items in the battery or some related index, such as Cronbach’s alpha [2], is typically used to assess reliability.

The model summarized in (1) is often unrealistically simple and more sophisticated models relax one or more of its assumptions. For example, it is sometimes reasonable to assume that errors for two different respondents are correlated when the same interviewer collects the data from both. Other models allow the true scores and observed values to be on different scales of measurement or allow the observed scores to reflect the impact of other underlying variables besides the true score on the variable of interest. These other variables affecting the observed score might be other substantive constructs or measurement factors, such as the format of the questions (see [12] for an example).

Cognitive Standards. Most questionnaire design efforts do not yield direct estimates of the reliability or validity of the key survey items. As noted, the typical procedure for estimating validity is to compare the survey responses to some external measure of the same variable, and this requires additional data collection for each survey item to be validated. Even obtaining an estimate of reliability typically requires recontacting some or all of the original respondents and administering the items

a second time. The additional data collection may exceed time or budget constraints. As a result, many survey design efforts rely on cognitive testing or other indirect methods to assess the measurement properties of a given survey item. The assumptions of this approach are that if respondents consistently have trouble understanding a question or remembering the information needed to answer it, then the question is unlikely to yield accurate answers. The evidence that respondents have difficulty comprehending the question, retrieving the necessary information, making the requisite judgments, and so on often comes from cognitive interviews. Alternatively, questionnaire design experts may flag the question as likely to produce problems for the respondents or evidence of such problems may arise from the behavior observed in the pilot interviews. For example, a high percentage of respondents may ask for clarification of the question or give inadequate answers. Whatever the basis for their judgments, the developers of survey questionnaires are likely to assess whether the draft questions seem to pose a reasonable cognitive challenge to the respondents or are too hard for respondents to understand or answer accurately.

Practical Standards. Surveys are often large-scale efforts and may involve thousands of respondents and hundreds of interviewers. Thus, a final test for a survey item or survey questionnaire is whether it can actually be administered in the field in a standardized way and at a reasonable cost. Interviewers may have difficulty reading long questions without stumbling; or they may misread questions involving unfamiliar vocabulary. Any instructions the interviewers are supposed to follow should be clear. Both individual items and the questionnaire as a whole should be easy for the interviewers to administer. The better the questionnaire design, the less training the interviewers will need. In addition, it is important to determine whether the time actually needed to complete the interview is consistent with the budget for the survey. These practical considerations are often the main reasons for conducting a pilot test of the questionnaire and debriefing the pilot interviewers.

Increasingly, survey questionnaires take the form of computer programs. The reliance on electronic questionnaires raises additional practical issues – Is the software user-friendly? Does it administer the

items as the authors intended? Do interviewers or respondents make systematic errors in interacting with the program? The questionnaire design process may include an additional effort – usability testing – to address these practical questions.

Conclusion

Designing survey questionnaires is a complex activity, blending data collection and statistical analysis with subjective impressions and expert opinions. Well-validated principles for writing survey questions are gradually emerging and questionnaire designers can now consult a substantial body of evidence about how to ask questions. For a summary of recent developments, see [11]. Still, for any particular survey, the questionnaire designers often have to rely on low-cost indirect indicators of measurement error in writing specific questions. In addition, the questionnaire that is ultimately fielded is likely to represent multiple compromises that balance statistical considerations, such as reliability and validity, against practical considerations, such as length, usability, and cost. Survey questionnaire design remains a mix of art and science, a blend of practicality and principle.

References

- [1] Biemer, P. & Stokes, L. (1991). Approaches to the modeling of measurement error, in *Measurement Errors in Surveys*, P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman, eds, John Wiley & Sons, New York.
- [2] Cronbach, L. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.
- [3] Fowler, F.J. (2004). Getting beyond pretesting and cognitive interviews: the case for more experimental pilot studies, in *Questionnaire Development Evaluation and Testing Methods*, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer, eds, John Wiley & Sons, New York.
- [4] Fowler, F.J. & Cannell, C.F. (1996). Using behavioral coding to identify cognitive problems with survey questions, in *Answering Questions*, N.A. Schwarz & S. Sudman, eds, Jossey-Bass, San Francisco.
- [5] Ericsson, K.A. & Simon, H.A. (1980). Verbal reports as data, *Psychological Review* **87**, 215–257.
- [6] Graesser, A.C., Kennedy, T., Wiemer-Hastings, P. & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires, in *Cognition in Survey Research*, M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J. Tanur & R. Tourangeau, eds, John Wiley & Sons, New York.
- [7] Jobe, J.B. & Mingay, D.J. (1989). Cognitive research improves questionnaires, *American Journal of Public Health* **79**, 1053–1055.
- [8] Lessler, J.T. & Forsyth, B.H. (1996). A coding system for appraising questionnaires, in *Answering Questions*, N.A. Schwarz & S. Sudman, eds, Jossey-Bass, San Francisco.
- [9] Payne, S. (1951). *The Art of Asking Questions*, Princeton University Press, Princeton.
- [10] Presser, S. & Blair, J. (1994). Survey pretesting: do different methods produce different results?, in *Sociological Methodology*, P.V. Marsden, ed., American Sociological Association, Washington.
- [11] Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. & Singer, E. (2004). *Questionnaire Development Evaluation and Testing Methods*, John Wiley & Sons, New York.
- [12] Saris, W.E. & Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modeling approach, in *Measurement Errors in Surveys*, P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman, eds, John Wiley & Sons, New York.
- [13] Schaeffer, N.C. & Presser, S. (2003). The science of asking questions, *Annual Review of Sociology* **29**, 65–88.
- [14] Sudman, S. & Bradburn, N. (1982). *Asking Questions*, Jossey-Bass, San Francisco.
- [15] Sudman, S., Bradburn, N. & Schwarz, N. (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*, Jossey-Bass.
- [16] Tourangeau, R. (2004). Design considerations for questionnaire testing and evaluation, in S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer, eds, *Questionnaire Development Evaluation and Testing Methods*, John Wiley, New York.
- [17] Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The Psychology of Survey Responses*, Cambridge University Press, New York.

(See also **Telephone Surveys**)

ROGER TOURANGEAU

Survey Sampling Procedures

KRIS K. MOORE

Volume 4, pp. 1977–1980

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Survey Sampling Procedures

Survey sampling provides a method of obtaining information about a population based on a sample selected from the population. A population is the totality of elements or individuals about which one wants to obtain information. A sample is a portion of the population that has been selected randomly. A sample characteristic is used to estimate the population characteristic. The cost of surveying all the elements in a population can be high, and the time for surveying each member of the population can be enormous. Sampling can provide a savings in cost and time.

Parameters or characteristics of the population are estimated from statistics that are characteristics of the sample. The estimates may differ by only $\pm 2.5\%$ from the actual population values. For example, in 2004, public opinion polls in the United States obtained information about approximately 300 million people from a sample of 1500 people. If in the sample of 1500, say, 600 favor the president's performance, the sample statistic is $600/1500 = 0.4$. The statistic 0.4 or 40% differs by only 2.5% from the population parameter 19 out of 20 times a sample of this size is taken. Thus, this sample would indicate that the interval 37.5 to 42.5% estimates the lower and upper bound of an interval that contains the percentage who favor the president's performance. About 95 out of 100 times a sample of this size (1500) would produce an interval that contains the population value.

Instrument of Survey Sampling

A questionnaire is usually designed to obtain information about a population from sampled values (*see Survey Questionnaire Design*). The questionnaire should be as brief as possible, preferably not more than two or three pages. The questionnaire should have a sponsor that is well known and respected by the sampled individuals. For example, if the population is a group of teachers, the teachers will be more likely to respond if the survey is sponsored by a recognizable and respected teacher organization. Also, money incentives such as 50¢ or a \$1.00 included

with the mailing improve the response rate on mailed questionnaires.

The questionnaire should begin with easy-to-answer questions; more difficult questions should be placed toward the end of the questionnaire. Demographic questions should be placed toward the beginning of the questionnaire because these questions are easy to answer. Care should be taken in constructing questions so that the questions are not offensive. For example rather than asking, 'How old are you' a choice of appropriate categories is less offensive. For example, 'Are you 20 or younger?' 'Are you 20 to 30?' and so on is less offensive. A few open-ended questions should be included to allow the respondent to clearly describe his or her position on issues not included in the body of the questionnaire.

Questionnaires should be tested in a small pilot survey before the survey is implemented to be sure the desired information is selected and the questions are easily understood. The pilot survey should include about 30 respondents.

Type of Question

Again, some open-ended questions should be used to allow the respondent freedom in defining his or her concerns. The Likert scale of measurement should be used on the majority of the questions. For example,

Strongly disagree	Mildly disagree	Disagree	Neutral	Mildly agree	Strongly agree
1	2	3	4	5	6

The Likert scale is easily converted to a form that computers can process.

The Frame

The sample in survey sampling is randomly selected from a frame or list of elements in the population. A random selection means each element of the population has an equal chance of being selected. Numbers can be assigned to each member of the population 1, 2, ..., N , where N is the total number in the population. Then random numbers can be used to select the sample elements.

Often a list of the population elements does not exist. Other methods will work for this type of

2 Survey Sampling Procedures

problem and are discussed in the Selection Methods section.

Collection of the Sample Information

The basic methods of obtaining responses in survey sampling are personal interview, mail questionnaire, telephone interview, or electronic responses via computers. Personal interviews are accurate but very expensive and difficult if the population includes a large geographic area. The person conducting the interview must be careful to be neutral and must not solicit preferred responses.

Mail questionnaires (*see Mail Surveys*) are inexpensive but the response rate is usually low, sometimes less than 10%. Incentives such as enclosing a dollar or offering the chance to win a prize increase response rates to 20 to 30%. Additional responses can be obtained by second and third mailings to those who failed to respond. Responses from first, second, and third mailings should be compared to see if trends are evident from one mailing to the other. For example, people who feel strongly about an issue may be more likely to respond to the first mailing than people who are neutral about an issue.

Telephone interviews are becoming more difficult to obtain because the general public has grown tired of tele-marketing and the aggravation of telephone calls at inconvenient times. In the United States, approximately 95% of the general population has telephones and of these, 10 to 15% of telephone numbers are unlisted. Random digit dialing allows unlisted numbers to be included in the sample by using the prefix for an area to be sampled, for example 756-XXXX. The XXXX is a four digit random number that is selected from a list of random numbers. This procedure randomly selects telephone numbers in 756 exchange.

Electronic methods of sampling are the most recently developed procedures (*see Internet Research Methods*). Internet users are sampled and incentives are used to produce a high response rate. Samples are easily selected at low cost via the computer.

Selection Methods

The simple random sample in which each element has an equal chance of selection is the most frequently used selection method when a frame or list of

elements exists (see [2]). A systematic sample in which every k th element is selected is often easier to obtain than a random sample. If N is the number in the population and n the sample size, then $k = N/n$, where k is rounded to the nearest whole number. The starting point 1 to k is randomly selected. If the sampled elements are increasing in magnitude, for example, inventory ordered by value from low-cost items to high-cost items, a systematic sample is better than a random sample. If the elements to be sampled are periodic, for example sales Monday through Saturday, then a random sample is better than a systematic sample because a systematic sample could select the same day each time. If the population is completely random, then systematic and random sample produce equivalent results.

If a population can be divided into groups of similar elements called strata, then a stratified random sample is appropriate (*see Stratification*). Random samples are selected from each stratum, which insures that the diversity of the population is represented in the sample and an estimate is also obtained for each stratum.

If no frame exists, a cluster sample is possible. For example, to estimate the number of deer on 100 acres, the 100 acres can be divided into one-acre plots on a map. Then a random sample of the one-acre plots can be selected and the number of deer counted for each selected plot. Suppose five acres are selected and a total of 10 deer are found. Then the estimate for the 100 acres would be 2 deer per acre and 200 for the 100 acres.

Sample Size

An approximate estimate of the sample size can be determined from the following two equations (see [1]). If you are estimating an average value for a quantitative variable, (1) can be used.

$$n = \left(\frac{2\sigma}{B} \right)^2, \quad (1)$$

where n is the sample size, σ is the population standard deviation, and B is the bound on the error. The bound on the error is the maximum differences between the true value and the stated value with probability .9544. An estimate of σ may be obtained in three ways: (a) estimate σ from historical studies, (b) estimate σ from (range of values / 6) $\cong \sigma$, and

(c) obtain a pilot sample of 30 or more elements and estimate σ by calculating $\hat{\sigma}$ (the sample standard deviation), (see 2).

For example, if you wanted to estimate miles per gallon, mpg, for a large population of automobiles within 2 mpg, then B would equal 2. A pilot sample can be taken and $\hat{\sigma}$ is used to estimate σ where $\hat{\sigma}$ is found from equation

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}, \quad (2)$$

where X_i is the sample value and $\bar{X}$ is the sample mean. For the above example, suppose $\sigma \cong 4$. Then using (1)

$$n = \left(\frac{2(4)}{2} \right)^2 = 16. \quad (3)$$

A sample of size 16 would produce a sample mean mpg that differs from the population mean mpg by 2 mpg or less.

If a population proportion (percentage) is to be estimated, the sample size is found from (4).

$$n = \left(\frac{2}{B} \right)^2 p(1-p), \quad (4)$$

where B is the bound on the error of the estimate, n is the sample size, and, p is the population proportion. The population proportion can be estimated three ways: (a) use historical values of p , (b) obtain a pilot sample of 30 or more elements and estimate p , and (c) use $p = .5$ because this produces the widest interval.

An example of determining the sample size necessary to estimate the population proportion is given

below when no information exists about the value of p . Suppose an estimate of the proportion of voters that favor a candidate is to be estimated within $\pm 3\%$ points. If there is no information from previous research, we select $p = .5$. Then $B = .03$ and n is determined from (4)

$$n = \left(\frac{2}{.03} \right)^2 .5(.5) = 1111.11 \text{ or } 1112. \quad (5)$$

Consider another example. If a low percentage is to be estimated like the proportion of defective light bulbs that is known to be 5% or less, then use $p = .05$ to estimate the proportion of defectives. If we want to estimate the percentage within $\pm 1\%$, we use $B = .01$. Then from (4)

$$n = \left(\frac{2}{.01} \right)^2 .05(.95) = 1900. \quad (6)$$

In summary, survey sampling procedures allow estimates of characteristics of populations (parameters) from characteristics of a sample (statistics). The procedure of survey sampling saves time and cost, and the accuracy of estimates is relatively high.

References

- [1] Keller, G. & Warrack, B. (2003). *Statistics for Management and Economics*, 6th edition, Thomson Learning Academic Resource Center.
- [2] Scheaffer, R.L., Mendenhall, I.I.I.W. & Ott, L.O. (1996). *Elementary Survey Sampling*, 5th edition, Duxbury Press.

KRIS K. MOORE

Survival Analysis

SABINE LANDAU

Volume 4, pp. 1980–1987

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Survival Analysis

In many studies the main outcome is the time from a well-defined *time origin* to the occurrence of a particular event or *end-point*. If the end-point is the death of a patient the resulting data are literally *survival times*. However, other end-points are possible, for example, the time to relief of symptoms or to recurrence of a particular condition, or simply to the completion of an experimental task. Such observations are often referred to as *time to event data* although the generic term *survival data* is commonly used to indicate any time to event data.

Standard statistical methodology is not usually appropriate for survival data, for two main reasons:

1. The distribution of survival time in general is likely to be *positively skewed* and so assuming normality for an analysis (as done for example, by a *t* Test or a regression) is probably not reasonable.
2. More critical than doubts about normality, however, is the presence of *censored* observations, where the survival time of an individual is referred to as censored when the end-point of interest has not yet been reached (more precisely *right-censored*). For true survival times this might be because the data from a study are analyzed at a time point when some participants are still alive. Another reason for censored event times is that an individual might have been *lost to follow-up* for reasons unrelated to the event of interest, for example, due to moving to a location which cannot be traced. When censoring occurs all that is known is that the actual, but unknown, survival time is larger than the censored survival time.

Specialized statistical techniques that have been developed to analyze such censored and possibly skewed outcomes are known as *survival analysis*. An important assumption made in standard survival analysis is that the censoring is *noninformative*, that is that the actual survival time of an individual is independent of any mechanism that causes that individual's survival time to be censored. For simplicity this description also concentrates on techniques for continuous survival times - for the analysis of discrete survival times see [3, 6].

A Survival Data Example

As an example, consider the data in Table 1 that contain the times heroin addicts remained in a clinic for methadone maintenance treatment [2]. The study ($n = 238$) recorded the duration spent in the clinic, and whether the recorded time corresponds to the time the patient leaves the programme or the end of the observation period (for reasons other than the patient deciding that the treatment is 'complete'). In this study the time of origin is the date on which the addicts first attended the methadone maintenance clinic and the end-point is methadone treatment cessation (whether by patient's or doctor's choice). The durations of patients who were lost during the follow-up process are regarded as right-censored. The 'status' variable in Table 1 takes the value unity if methadone treatment was stopped, and zero if the patient was lost to follow-up. In addition a number of prognostic variables were recorded;

1. one of two clinics,
2. presence of a prison record
3. and maximum methadone dose prescribed.

The main aim of the study was to identify predictors of the length of the methadone maintenance period.

Survival Analysis Concepts

To describe survival two functions of time are of central interest. The *survival function* $S(t)$ is defined as the probability that an individual's survival time, T , is greater than or equal to time t , that is,

$$S(t) = \text{Prob}(T \geq t) \quad (1)$$

The graph of $S(t)$ against t is known as the *survival curve*. The survival curve can be thought of as a particular way of displaying the frequency distribution of the event times, rather than by say a histogram.

In the analysis of survival data it is often of some interest to assess which periods have the highest and which the lowest chance of death (or whatever the event of interest happens to be), amongst those people at risk at the time. The appropriate quantity for such risks is the *hazard function*, $h(t)$, defined as the (scaled) probability that an individual experiences the event in a small time interval δt , given that

2 Survival Analysis

Table 1 Durations of heroin addicts remaining in a methadone treatment programme (only five patients from each clinic are shown)

Patient ID	Clinic	Status	Time (days)	Prison record (1 = 'present', 0 = 'absent')	Maximum methadone (mg/day)
1	1	1	428	0	50
2	1	1	275	1	55
3	1	1	262	0	55
4	1	1	183	0	30
5	1	1	259	1	65
...					
103	2	1	708	1	60
104	2	0	713	0	50
105	2	0	146	0	50
106	2	1	450	0	55
109	2	0	555	0	80
...					

the individual has not experienced the event up to the beginning of the interval. The hazard function therefore represents the *instantaneous event rate* for an individual at risk at time t . It is a measure of how likely an individual is to experience an event as a function of the age of the individual. The hazard function may remain constant, increase or decrease with time, or take some more complex form. The hazard function of death in human beings, for example, has a 'bath tub' shape. It is relatively high immediately after birth, declines rapidly in the early years and then remains pretty much constant until beginning to rise during late middle age.

In formal, mathematical terms, the hazard function is defined as the following limiting value

$$h(t) = \lim_{\delta t \rightarrow 0} \left[\frac{\text{Prob}(t \leq T < t + \delta t | T \geq t)}{\delta t} \right] \quad (2)$$

The conditional probability is expressed as a probability per unit time and therefore converted into a rate by dividing by the size of the time interval, δt .

A further function that features widely in survival analysis is the *integrated* or *cumulative hazard function*, $H(t)$, defined as

$$H(t) = \int_0^t h(u) du \quad (3)$$

The hazard function is mathematically related to the survivor functions. Hence, once a hazard function is specified so is the survivor function and *vice versa*.

Nonparametric Procedures

An initial step in the analysis of a set of survival data is the numerical or graphical description of the survival times. However, owing to the censoring this is not readily achieved using conventional descriptive methods such as boxplots and summary statistics. Instead survival data are conveniently summarized through estimates of the survivor or hazard function obtained from the sample.

When there are no censored observations in the sample of survival times, the survival function can be estimated by the *empirical survivor function*

$$\hat{S}(t) = \frac{\text{Number of individuals with event times} \geq t}{\text{Number of individuals in the data set}} \quad (4)$$

Since every subject is 'alive' at the beginning of the study and no-one is observed to survive longer than the largest of the observed survival times then

$$\hat{S}(0) = 1 \text{ and } \hat{S}(t) = 0 \text{ for } t > t_{\max} \quad (5)$$

Furthermore the estimated survivor function is assumed constant between two adjacent death times so that a plot of $\hat{S}(t)$ against t is a step function that decreases immediately after each 'death'. However, this simple method cannot be used when there are censored observations since the method does not allow for information provided by an individual whose survival time is censored before time t to be used in the computing of the estimate at t .

The most commonly used method for estimating the survival function for survival data containing censored observations is the *Kaplan–Meier* or *product-limit-estimator* [8]. The essence of this approach is the use of a product of a series of conditional probabilities. This involves ordering the r sample event times from the smallest to the largest such that

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)} \quad (6)$$

Then the survival curve is estimated from the formula

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \quad (7)$$

where r_j is the number of individuals at risk at $t_{(j)}$ and d_j is the number experiencing the event of interest at $t_{(j)}$. (Individuals censored at $t_{(j)}$ are included in r_j .) For example, the estimated survivor function at the second event time $t_{(2)}$ is equal to the estimated probability of not experiencing the event at time $t_{(1)}$ times the estimated probability, given that the individual is still at risk at time $t_{(2)}$, of not experiencing it at time $t_{(2)}$.

The Kaplan-Meier estimators of the survivor curves for the two methadone clinics are displayed in Figure 1. The survivor curves are step functions with decrease at the time points when patients ceased methadone treatment. The censored observations in the data are indicated by the ‘cross’ marks on the curves.

The variance of the Kaplan-Meier estimator of the survival curve can itself be estimated from *Greenwood’s formula* and once the standard error has been determined point-wise symmetric confidence intervals can be found by assuming a normal distribution on the original scale or asymmetric intervals can be constructed after transforming $\hat{S}(t)$ to a value on the continuous scale, for details see [3, 6].

A *Kaplan-Meier type estimator* of the hazard function is given by the proportion of individuals experiencing an event in an interval per unit time, given that they are at risk at the beginning of the interval, that is

$$\tilde{h}(t) = \frac{d_j}{r_j (t_{(j+1)} - t_{(j)})} \quad (8)$$

Integration leads to the *Nelson-Aalen* or *Altshuler’s* estimator of the cumulative hazard function, $\tilde{H}(t)$, and employing the theoretical relationship between the survivor function and the cumulative hazard function to the Nelson-Aalen estimator of the survivor function. Finally, it needs to be noted that relevant functions can be estimated using the so-called *life-table* or *Actuarial estimator*. This approach is, however, sensitive to the choice of intervals used in its construction and therefore not generally recommended for continuous survival data (readers are referred to [3, 6]).

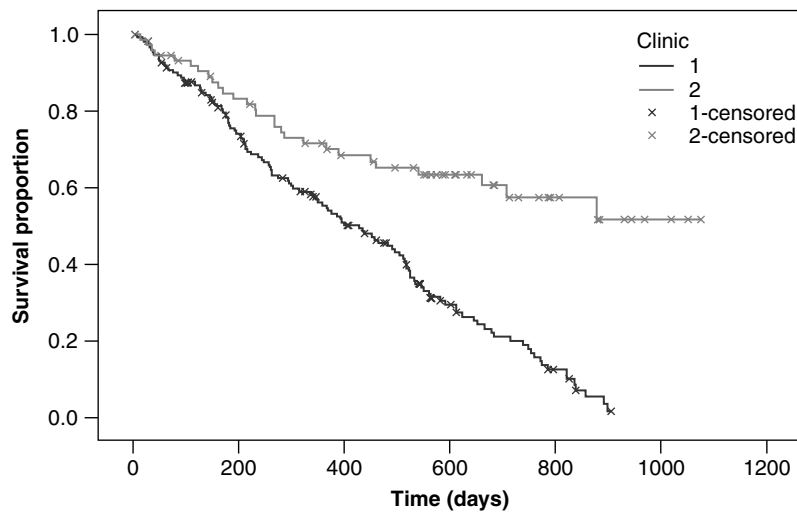


Figure 1 Kaplan-Meier survivor curves for the heroin addicts data

Standard errors and confidence intervals can be constructed for all three functions although the estimated hazard function is generally considered ‘too noisy’ for practical use. The Nelson-Aalen estimator is typically used to describe the cumulative hazard function while the Kaplan-Meier estimator is used for the survival function.

Since the distribution of survival times tends to be positively skewed the median is the preferred summary measure of location. The *median event time* is the time beyond which 50% of the individuals in the population under study are expected to ‘survive’, and, once the survivor function has been estimated by $\hat{S}(t)$, can be estimated by the smallest observed survival time, t_{50} , for which the value of the estimated survivor function is less than 0.5. The estimated median survival time can be read from the survival curve by finding the smallest value on the x -axis for which the survival proportions reaches less than 0.5. Figure 1 shows that the median methadone treatment duration in clinic 1 group can be estimated as 428 days while an estimate is not available for clinic 2 since more than 50% of the patients continued treatment throughout the study period. A similar procedure can be used to estimate other percentiles of the distribution of the survival times and approximate confidence intervals can be found once the variance of the estimated percentile has been derived from the variance of the estimator of the survivor function.

In addition to comparing survivor functions graphically a more formal statistical test for a group difference is often required. In the absence of censoring a nonparametric test, like the Mann-Whitney test could be used (*see Distribution-free Inference, an Overview*). In the presence of censoring the *log-rank* or *Mantel-Haenszel test* [9] is the most commonly used nonparametric test. It tests the null hypothesis that the population survival functions $S_1(t), S_2(t), \dots, S_k(t)$ are the same in k groups.

Briefly, the test is based on computing the expected number of events for each observed event time in the data set, assuming that the chances of the event, given that subjects are at risk, are the same in the groups. The total number of expected events is then computed for each group by adding the expected number of events for each event time. The test finally compares the observed number of events in each group with the expected number of

events using a chi-squared test with $k - 1$ degrees of freedom, *see* [3, 6].

The log-rank test statistic, X^2 , weights contributions from all failure times equally. Several alternative test statistics have been proposed that give differential weights to the failure times. For example, the *generalized Wilcoxon test* (or *Breslow test*) uses weights equal to the number at risk. For the heroin addicts data in Table 1 the log-rank test ($X^2 = 27.9$ on 1 degree of freedom, $p < 0.0001$) detects a significant clinic difference in favor of longer treatment durations in clinic 2. The Wilcoxon test puts relatively more weight on differences between the survival curves at earlier times but also reaches significance ($X^2 = 11.6$ on 1 degree of freedom, $p = 0.0007$).

Modeling Survival Times

Modeling survival times is useful especially when there are several explanatory variables of interest. For example the methadone treatment durations of the heroin addicts might be affected by the prognostic variables maximum methadone dose and prison record as well as the clinic attended. The main approaches used for Modeling the effects of covariates on survival can be divided roughly into two classes – *models for the hazard function* and *models for the survival times themselves*. In essence these models act as analogies of **multiple linear regression** for survival times containing censored observations, for which regression itself is clearly not suitable.

Proportional Hazards Models

The main technique is due to Cox [4] and known as the *proportional hazards model* or, more simply, *Cox’s regression*. The approach stipulates a model for the hazard function. Central to the procedure is the assumption that the hazard functions for two individuals at any point in time are proportional, the so-called *proportional hazards assumption*. In other words, if an individual has a risk of the event at some initial time point that is twice as high as another individual, then at all later times the risk of the event remains twice as high. Cox’s model is made up of an unspecified *baseline hazard function*, $h_0(t)$, which is then multiplied by a suitable function of an individual’s explanatory variable values, to give the

individual's hazard function. Formally, for a set of p explanatory variables, $x_1, x_2, \dots, x_p$, the model is

$$h(t) = h_0(t) \exp \left(\sum_{i=1}^p \beta_i x_i \right) \quad (9)$$

where the terms $\beta_1, \dots, \beta_p$ are the parameters of the model which have to be estimated from sample data. Under this model the *hazard* or *incidence rate ratio*, h_{12} , for two individuals, with covariate values $x_{11}, x_{12}, \dots, x_{1p}$ and $x_{21}, x_{22}, \dots, x_{2p}$

$$h_{12} = \frac{h_1(t)}{h_2(t)} = \exp \left[\sum_{i=1}^p \beta_i (x_{1i} - x_{2i}) \right] \quad (10)$$

does not depend on t . The interpretation of the parameter β_i is that $\exp(\beta_i)$ gives the incidence rate change associated with an increase of one unit in x_i , all other explanatory variables remaining constant. Specifically, in the simple case of comparing hazards between two groups, $\exp(\beta)$, measures the hazard ratio between the two groups. The effect of the covariates is assumed multiplicative.

Cox's regression is considered a *semiparametric* procedure because the baseline hazard function, $h_0(t)$, and by implication the probability distribution of the survival times does not have to be specified. The baseline hazard is left unspecified; a different parameter is essentially included for each unique survival time. These parameters can be thought of as nuisance parameters whose purpose is merely to control the parameters of interest for any changes in the hazard over time. Cox's regression model can also be extended to allow the baseline hazard function to vary with the levels of a stratification variable. Such a *stratified proportional hazards model* is useful in situations where the stratifier is thought to affect the hazard function but the effect itself is not of primary interest.

A Cox regression can be used to model the methadone treatment times from Table 1. The model uses prison record and methadone dose as explanatory variables whereas the variable clinic, whose effect was not of interest, merely needed to be taken account of and did not fulfill the proportional hazards assumption, was used as a stratifier. The estimated regression coefficients are shown in Table 2. The coefficient of the prison record indicator variable is 0.389 with a standard error of 0.17. This translates into a hazard ratio of $\exp(0.389) = 1.475$ with a 95% confidence interval ranging from 1.059 to 2.054. In other words a prison record is estimated to increase the hazard of immediate treatment cessation by 47.5%. Similarly the hazard of treatment cessation was estimated to be reduced by 3.5% for every extra mg/day of methadone prescribed.

Statistical software packages typically report three different tests for testing regression coefficients, the *likelihood ratio (LR) test* (see **Maximum Likelihood Estimation**), the *score test* (which for Cox's proportional hazards model is equivalent to the log-rank test) and the *Wald test*. The test statistic of each of the tests can be compared with a chi-squared distribution to derive a P value. The three tests are asymptotically equivalent but differ in finite samples. The likelihood ratio test is generally considered the most reliable and the Wald test the least. Here presence of a prison record tests statistically significant after adjusting for clinic and methadone dose (LR test: $X^2 = 5.2$ on 1 degree of freedom, $p = 0.022$) and so does methadone dose after adjusting for clinic and prison record (LR test: $X^2 = 30.0$ on 1 degree of freedom, $p < 0.0001$).

Cox's model does not require specification of the probability distribution of the survival times. The hazard function is not restricted to a specific form and as a result the semiparametric model has flexibility and is widely used. However, if the assumption of

Table 2 Parameter estimates from Cox regression of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

Predictor variable	Effect estimate			95% CI for $\exp(\beta)$	
	Regression coefficient ($\hat{\beta}$)	Standard error ($\sqrt{\text{var}(\hat{\beta})}$)	Hazard ratio ($\exp(\hat{\beta})$)	Lower limit	Upper limit
Prison record	0.389	0.17	1.475	1.059	2.054
Maximum methadone dose	-0.035	0.006	0.965	0.953	0.978

a particular probability distribution for the data is valid, inferences based on such an assumption are more precise. For example estimates of hazard ratios or median survival times will have smaller standard errors. A *fully parametric proportional hazards model* makes the same assumptions as Cox's regression but in addition also assumes that the baseline hazard function, $h_0(t)$, can be parameterized according to a specific model for the distribution of the survival times. Survival time distributions that can be used for this purpose, that is that have the *proportional hazards property*, are principally the *Exponential*, *Weibull*, and *Gompertz* distributions (see **Catalogue of Probability Density Functions**). Different distributions imply different shapes of the hazard function, and in practice the distribution that best describes the functional form of the observed hazard function is chosen – for details see [3, 6].

Models for Direct Effects on Survival Times

A family of fully parametric models that assume direct multiplicative effects of covariates on survival times and hence do not rely on proportional hazards are *accelerated failure time models*. A wider range of survival time distributions possesses the *accelerated failure time property*, principally the *Exponential*, *Weibull*, *log-logistic*, *generalized gamma*, or *lognormal* distributions. In addition this family of parametric models includes distributions (e.g., the log-logistic distribution) that model unimodal hazard functions while all distributions suitable for the proportional hazards model imply hazard functions that increase or decrease monotonically. The latter property might be limiting, for example, for Modeling the hazard of dying after a complicated operation that peaks in the postoperative period.

The general accelerated failure time model for the effects of p explanatory variables, $x_1, x_2, \dots, x_p$, can

be represented as a **log-linear model** for survival time, T , namely,

$$\ln(T) = \alpha_0 + \sum_{i=1}^p \alpha_i x_i + \text{error} \quad (11)$$

where $\alpha_1, \dots, \alpha_p$ are the unknown coefficients of the explanatory variables and α_0 an intercept parameter. The parameter α_i reflects the effect that the i th covariate has on log-survival time with positive values indicating that the survival time increases with increasing values of the covariate and *vice versa*. In terms of the original time scale the model implies that the explanatory variables measured on an individual act multiplicatively, and so affect the speed of progression to the event of interest.

The interpretation of the parameter α_i then is that $\exp(\alpha_i)$ gives the factor by which any survival time percentile (e.g., the median survival time) changes per unit increase in x_i , all other explanatory variables remaining constant. Expressed differently, the probability, that an individual with covariate value $x_i + 1$ survives beyond t , is equal to the probability, that an individual with value x_i survives beyond $\exp(-\alpha_i)t$. Hence $\exp(-\alpha_i)$ determines the change in the speed with which individuals proceed along the time scale and the coefficient is known as the *acceleration factor* of the i th covariate.

Software packages typically use the log-linear formulation. The regression coefficients from fitting a log-logistic accelerated failure time model to the methadone treatment durations using prison record and methadone dose as explanatory variables and clinic as a stratifier are shown in Table 3. The negative regression coefficient for prison suggests that the treatment durations tend to be shorter for those with a prison record. The positive regression coefficient for dose suggests that treatment durations tend to be prolonged for those on larger methadone doses.

Table 3 Parameter estimates from log-logistic accelerated failure time model of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

Predictor variable	Effect estimate			95% CI for $\exp(\alpha)$	
	Regression coefficient ($\hat{\alpha}$)	Standard error ($\sqrt{\text{var}(\hat{\alpha})}$)	Acceleration factor ($\exp(-\hat{\alpha})$)	Lower limit	Upper limit
Prison record	-0.328	0.140	1.388	1.054	1.827
Maximum methadone dose	0.0315	0.0055	0.969	0.959	0.979

The estimated acceleration factor for an individual with a prison record compared with one without such a record is $\exp(0.328) = 1.388$, that is, a prison record is estimated to accelerate the progression to treatment cessation by a factor of about 1.4. Both explanatory variables, prison record (LR test: $X^2 = 5.4$ on 1 degree of freedom, $p = 0.021$) and maximum methadone dose prescribed (LR test: $X^2 = 31.9$ on 1 degree of freedom, $p < 0.0001$) are found to have statistically significant effects on treatment duration according to the log-logistic accelerated failure time model.

Summary

Survival analysis is a powerful tool for analyzing time to event data. The classical techniques Kaplan-Meier estimation, proportional hazards, and accelerated failure time Modeling are implemented in most general purpose statistical packages with the S-PLUS and R packages having particularly extensive facilities for fitting and checking nonstandard Cox models, *see* [10]. The area is complex and one of active current research. For additional topics such as other forms of censoring and *truncation* (delayed entry), *recurrent events*, models for *competing risks*, *multistate models* for different transition rates, and *frailty models* to include random effects the reader is referred to [1, 5, 7, 10].

References

- [1] Andersen, P.K., ed. (2002). *Multistate Models, Statistical Methods in Medical Research 11*, Arnold Publishing, London.
- [2] Caplethorn, J. (1991). Methadone dose and retention of patients in maintenance treatment, *Medical Journal of Australia* **154**, 195–199.
- [3] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Edition, Chapman & Hall/CRC, London.
- [4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society: Series B* **74**, 187–220.
- [5] Crowder, K.J. (2001). *Classical Competing Risks*, Chapman & Hall/CRC, Boca Raton.
- [6] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis*, Wiley, New York.
- [7] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.
- [8] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
- [9] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion), *Journal of the Royal Statistical Society: Series A* **135**, 185–206.
- [10] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

SABINE LANDAU

Symmetry: Distribution Free Tests for

CLIFFORD E. LUNNEBORG

Volume 4, pp. 1987–1988

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Symmetry: Distribution Free Tests for

Nonparametric tests for the median of a distribution and the related estimation of confidence intervals for that parameter assume that the distribution sampled is symmetric about the median. To avoid obtaining misleading results, a preliminary test for distribution symmetry is advisable.

Tests for Symmetry of Distribution

The four tests of symmetry selected for mention here have been shown to have reasonable power to detect asymmetric distributions and are either widely used or easy to apply.

The Gupta test proposed in 1967 [3] and [4] is based on the set of pair-wise comparisons of the sample data. For each pair of data points, the statistic δ_{ij} is assigned a value of 1 if $(x_i + x_j)$ is greater than twice the sample median or 0 otherwise. The sum of these δ_{ij} s will be large if the underlying distribution is skewed to the right, small if it is skewed to the left, and intermediate in value if the distribution is symmetric. After centering and standardization, the test statistic is asymptotically distributed as the standard normal random variable. The approach is explained in detail in [4], and has been built into the SC statistical package (www.mole-software.demon.co.uk) as the *Gupta* procedure.

The Randles et al. test, published in 1980 [5] and [6], is based on the set of triplets of data points. For each unique set of three data points, the statistic ζ_{ijk} is assigned the value 1/3 if the mean of the three data points, (x_i, x_j, x_k) , is greater than their median, 0 if the mean and median are equal, and $-1/3$ if the median is larger than the mean. The sum of these ζ_{ijk} s will be large and positive if the underlying distribution is skewed to the right, large and negative if the distribution is skewed to the left, and small if the distribution is symmetric. After standardization, this sum enjoys, at least asymptotically, a normal sampling distribution under the null, symmetric hypothesis. Details are given in [5].

The Boos test, described in 1982 [1], is based on the set of absolute differences between the $n(n+1)/2$ **Walsh averages** and their median, the one-sample **Hodges–Lehmann estimator**. When summed, large values are suggestive of an underlying asymmetric distribution. Critical asymptotic values for a scaling of the sum of absolute deviations are given in [1].

The Cabilio & Masaro test, presented in 1996 [2], features simplicity of computation. The test statistic,

$$S_K = \frac{[\sqrt{n}(Mn - Mdn)]}{Sd}, \quad (1)$$

requires for its computation only four sample quantities – size, mean, median, and standard deviation. The test’s authors recommend comparing the value of the test statistic, S_K , against the critical values in

Table 1 Empirical quantiles of the distribution of S_K (normal samples)

n	Q_{90}	Q_{95}	$Q_{97.5}$	n	Q_{90}	Q_{95}	$Q_{97.5}$
5	0.88	1.07	1.21	6	0.71	0.88	1.02
7	0.91	1.13	1.30	8	0.77	0.96	1.13
9	0.92	1.15	1.35	10	0.81	1.01	1.19
11	0.93	1.17	1.37	12	0.83	1.05	1.24
13	0.93	1.18	1.39	14	0.85	1.08	1.27
15	0.94	1.19	1.41	16	0.86	1.10	1.30
17	0.94	1.19	1.41	18	0.87	1.11	1.32
19	0.94	1.20	1.42	20	0.88	1.12	1.33
21	0.95	1.20	1.43	22	0.89	1.13	1.35
23	0.95	1.21	1.43	24	0.89	1.14	1.36
25	0.95	1.21	1.44	26	0.90	1.15	1.37
27	0.95	1.21	1.44	28	0.90	1.15	1.37
29	0.95	1.21	1.44	30	0.91	1.16	1.38
∞	0.97	1.24	1.48				

Table 1, reproduced here from [2] with the permission of the authors.

These critical values were obtained by calibrating the test against samples from a particular symmetric distribution, the normal. However, the nominal significance level of the test appears to hold for other symmetric distributions, with the exception of the Cauchy and uniform distributions (*see Catalogue of Probability Density Functions*) [2].

Comments

The power of the four tests of distribution symmetry have been evaluated and compared, [2] and [6]. Briefly, the Randles et al. test dominates the Gupta test [6], while the Boos and Cabilio–Masaro tests appear to be superior to Randles et al. [2]. Although the Boos test may have a slight power advantage over the Cabilio–Masaro procedure, the ease of application of the latter suggests that the assumption of symmetry might more frequently be checked than it has been in the past. It should be noted, however,

that tests of symmetry appear to have fairly low power to detect asymmetric distributions when the sample size is smaller than 20 [6].

References

- [1] Boos, D.D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator, *Journal of the American Statistical Association* **77**, 647–651.
- [2] Cabilio, P. & Masaro, J. (1996). A simple test of symmetry about an unknown median, *The Canadian Journal of Statistics* **24**, 349–361.
- [3] Gupta, M.K. (1967). An asymptotically nonparametric test of symmetry, *Annals of Mathematical Statistics* **38**, 849–866.
- [4] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*, Wiley, New York.
- [5] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.
- [6] Randles, H., Fligner, M.A., Pollicello, G. & Wolfe, D.A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry, *Journal of the American Statistical Association* **75**, 168–172.

CLIFFORD E. LUNNEBORG

Symmetry Plot

SANDY LOVIE

Volume 4, pp. 1989–1990

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Symmetry Plot

This plot does exactly what it says on the tin, that is, it provides a graphical test of whether a sample is symmetrically distributed about a measure of location; in this case, the median. Having such information about a sample is useful in that just about all tests of significance assume that the parent population from which the sample came is at least symmetrical about some location parameter and, in effect, that the sample should not markedly violate this condition either. A further linked role for the plot is its use in evaluating transformations to *achieve* symmetry, particularly following schemes like the *ladder of powers* advocated for **Exploratory Data Analysis** (see, for example [2]).

The plot itself is built up by first ordering the data and calculating a median, if necessary, by interpolation for even numbered samples. Secondly, each reading in the sample is subtracted from the median, thus reexpressing all the sample values as (signed and ordered) distances from the median. Then, these distances are expressed as unsigned values, whilst still keeping separate those ordered distances that lie above the median from those below it. Next, as with the **empirical quantile-quantile (EQQ) plot**, the ordered values above and below the median are paired in increasing order of size, and then plotted on a conventional **scatterplot**. Here the zero/zero origin represents the median itself, with the ascending dots representing the ordered pairs, where the lowest one represents the two smallest distances above and below the median, the next higher dot the next smallest pair, and so on. Also, as with the EQQ plot, a 45° comparison line is placed on the plot to represent perfect symmetry about the median (note that the x and y axes of the scatterplot are equal in all respects, hence the angle of the line). All judgements as to the symmetry of the sample are therefore made relative to this line. The statistics package Minitab adds a simple histogram of the data to its version of the plot, thus aiding in its interpretation (see **Software for Statistical Analyses**).

The three illustrative plots below are from Minitab and use the *Pulse* data set. Figure 1 is a symmetry plot of the raw data for 35 human pulse readings after exercise (running on the spot for one minute). Here, the histogram shows a marked skew to the left, which shows up on the full plot as the data, both lying below

the comparison line and increasingly divergent from it, as one moves to the right. However, if the data had been skewed to the right, then the plotted data would have appeared above the comparison line.

The next two plots draw on transforms from the ladder of powers to improve the symmetry of the data. The first applies a $\log_{10}$ transform to the data, while the second uses a reciprocal ($1/x$) transform (see **Transformation**). Notice that the log transform in Figure 2 improves the symmetry of the histogram a little, which shows up in the symmetry plot as data, which are now somewhat closer to the comparison line and less divergent from it than in the raw plot.

However, this is improved on even further in Figure 3, where the histogram is even more symmetric, and the data of the symmetry plot is much closer and much less divergent than in the log transformed plot.

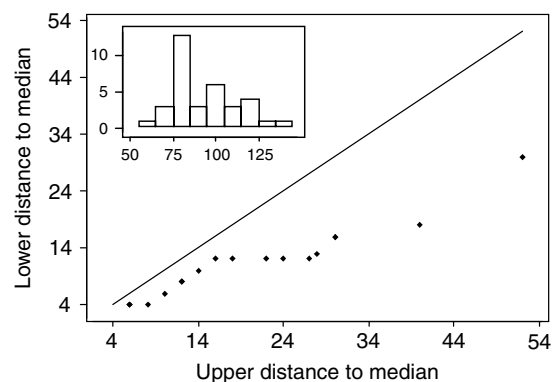


Figure 1 Symmetry plot of raw 'pulse after exercise' data

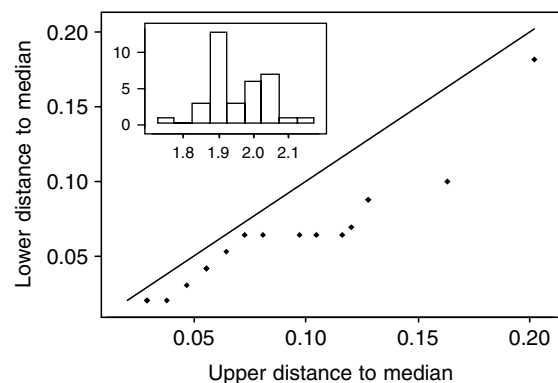


Figure 2 Symmetry plot of the log transformed 'pulse after exercise' data

2 Symmetry Plot

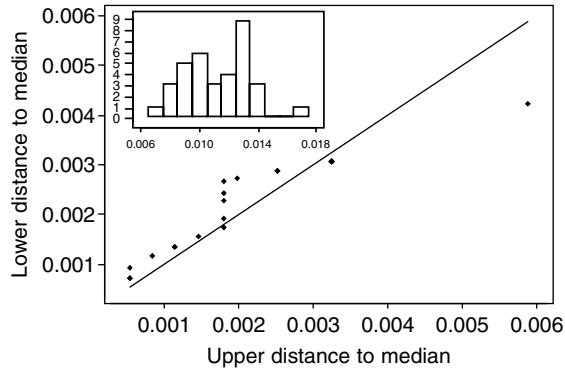


Figure 3 Symmetry plot of the reciprocal transformed 'pulse after exercise' data

Interestingly, a somewhat more complex transform lying between the two chosen here from the ladder of powers, the *reciprocal/square root* ($1/\sqrt{x}$), generates a symmetry plot (not included here), which reproduces many of the aspects of Figure 3 for the

bulk of the data on the left-hand side of the plot, and also draws the somewhat anomalous data point on the far right, much closer to the comparison line. Although analyses of the resulting data might be more difficult to interpret, this transform is probably the one to choose if you want your (transformed) results to conform to the symmetry assumption behind all those tests of significance!

More information on symmetry plots can be found in [1], pages 29 to 32.

References

- [1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.
- [2] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

SANDY LOVIE

Tau-Equivalent and Congeneric Measurements

JOS M.F. TEN BERGE

Volume 4, pp. 1991–1994

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tau-Equivalent and Congeneric Measurements

It is a well-known fact that the reliability of a test, defined as the ratio of true score to observed score variance, cannot generally be determined from a single test administration, but requires the use of a parallel test. More often than not, parallel tests are not available. In such cases, two approaches are popular to obtain indirect information on the reliability of the test: either lower bounds to reliability can be used, or one may resort to hypotheses about the nature of the test parts.

Evaluating lower bounds to the reliability, such as Guttman's λ_3 [6], better known as coefficient alpha [4] has gained wide popularity. A lower bound that is nearly always better than alpha is Guttman's λ_4 . It is the highest alpha that can be obtained by splitting up the items in two parts (not necessarily of equal numbers) and treating those two parts as novel 'items'. Jackson and Agunwamba [8] proposed the greatest lower bound (glb) to reliability. It exceeds all conceivable lower bounds by using the available information implied by the observed covariance matrix exhaustively. A computational method for the glb has been proposed by Bentler & Woodward [3], also see [19]. Computation of the glb has been implemented in EQS 6.

When lower bounds are high enough, the reliability has been shown adequate by implication. However, when lower bounds are low, they are of limited value. Also, some lower bounds to reliability involve a considerable degree of sampling bias. To avoid these problems, it is tempting to look to alternative approaches, by introducing hypotheses on the nature of the test parts, from which the reliability can be determined at once. Two of such hypotheses are well-known in **classical test theory**.

Tau Equivalence

The first hypothesis is that of (essentially) tau-equivalent tests. Test parts $X_1, \dots, X_k$ are essentially tau-equivalent when for $i, j = 1, \dots, k$,

$$T_j = T_i + a_{ij}. \quad (1)$$

This implies that the true scores of the test parts are equal up to an additive constant. When the additive constants are zero, the test parts are said to be tau-equivalent. Novick and Lewis [14] have shown that coefficient alpha is the reliability (instead of merely a lower bound to it) if and only if the test parts are essentially tau-equivalent.

Unfortunately, the condition for essential tau-equivalence to hold in practice is prohibitive: All covariances between test parts must be equal. This will only be observed when $k = 2$ or with contrived data. Moreover, the condition of equal covariances is necessary, but not sufficient for essential tau-equivalence. For instance, let Y_1, Y_2 and Y_3 be three uncorrelated variables with zero means and unit variances, and let the test parts be $X_1 = Y_2 + Y_3$, $X_2 = Y_1 + Y_3$, and $X_3 = Y_1 + Y_2$. Then, the covariance between any two test parts is 1, but the test parts are far from being essentially tau-equivalent, which shows that having equal covariances is necessary but not sufficient for essential tau-equivalence. Because the necessary condition will never be satisfied in practice, coefficient alpha is best thought of as a lower bound (underestimate) to the reliability of a test.

Congeneric Tests

A weaker, and far more popular hypothesis, is that of a congeneric test, consisting of k test parts satisfying

$$T_j = c_{ij}T_i + a_{ij}, \quad (2)$$

which means that the test parts have perfectly correlated true scores. Equivalently, the test parts are assumed to fit a one-factor model. Essential tau equivalence is more restricted because it requires that the weights c_{ij} in (2) are unity. For the case $k = 3$, Kristof has derived a closed-form expression for the reliability of the test, based on this hypothesis. It is always at least as high as alpha, and typically better [10]. Specifically, there are cases where it coincides with or even exceeds the greatest lower bound to reliability [20].

For $k > 3$, generalized versions of Kristof's coefficients have been proposed. For instance, Gilmer and Feldt [5] offered coefficients that could be evaluated without having access to powerful computers. They were fully aware that these coefficients would be supplanted by common factor analysis (see **History of Factor Analysis: A Psychological Perspective**)

based coefficients by the time large computers would be generally available. Nowadays, even the smallest of personal computers can evaluate the reliability of a test in the framework of common factor analysis, assuming that the one-factor hypothesis is true. For instance, McDonald [13], also see Jöreskog [9] for a similar method, proposed estimating the loadings on a single factor and evaluating the reliability as the ratio of the squared sum of loadings to the test variance. When $k = 3$, this yields Kristof's coefficient. Coefficients like Kristof's and McDonald's have been considered useful alternatives to lower bounds like glb, because they aim to estimate, rather than underestimate, reliability, and they lack any reputation of sampling bias. However, much like the hypothesis of essential tau-equivalence, the one-factor hypothesis is problematic.

The Hypothesis of Congeneric Tests is Untenable for $k > 3$, and Undecided Otherwise

The hypothesis of congeneric tests relies on the existence of communalities to be placed in the diagonal cells of the item covariance matrix, in order to reduce the rank of that matrix to one. The conditions under which this is possible have been known for a long time. **Spearman** [17] already noted that unidimensionality is impossible (except in contrived cases) when $k > 3$. Accordingly, when $k > 3$, factor analysis with only one common factor will never give perfect fit. More generally, Wilson and Worcester [23], Guttman [7], and Bekker and De Leeuw [1] have argued that rank reduction of a covariance matrix by communalities does not carry a long way. Shapiro [15] has proven that the minimal reduced rank that can possibly be achieved will be at or above the Ledermann bound [11] almost surely. It means that the minimal reduced rank is almost surely at or above 1 when $k = 3$, at or above 2 when $k = 4$, at or above 3 when $k = 5$ or 6, and so on. The notion of 'almost surely' reflects the fact that, although covariance matrices that do have lower reduced rank are easily constructed, they will never be observed in practice. It follows that the hypothesis of congeneric tests is nearly as unrealistic as that of essential tau-equivalence. It may be true only when there are three or fewer items.

Even when reduction to rank 1 is possible, this is not sufficient for the hypothesis to be true: We merely have a necessary condition that is satisfied. The example of X_1 , X_2 , and X_3 with three uncorrelated underlying factors Y_1 , Y_2 , and Y_3 , given above in the context of tau-equivalence, may again be used to demonstrate this: There are three underlying factors, yet communalities that do reduce the rank to 1 do exist (being 1, 1, and 1). The bottom line is that the hypothesis of congeneric tests cannot be rejected, but still may be false when $k = 3$ or less, and it has to be rejected when $k > 3$. 'Model-based coefficients are not useful if the models are not consistent with the empirical data' [2]. Reliability coefficients based on the single-factor hypothesis are indeed a case where this applies.

Sampling Bias

Lower bounds to reliability do not rest on any assumption other than that error scores of the test parts correlate only with themselves and with the observed scores they belong with. On the other hand, lower bounds do have a reputation for sampling bias. Whereas coefficient alpha tends to slightly underestimate the population alpha [21, 24], Guttman's λ_4 , and the greatest lower bound in particular, may grossly overestimate the population value when computed in small samples. For instance, when $k = 10$ and the population glb is 0.68, its average sample estimate may be as high as 0.77 in samples of size 100 [16].

It may seem that one-factor-based coefficients have a strong advantage here. But this is not true. When $k = 3$, Kristof's coefficient often coincides with glb, and for $k > 3$, numerical values of McDonald's coefficient are typically very close to the glb. In fact, McDonald's coefficient demonstrates the same sampling bias as glb in Monte Carlo studies [20]. Because McDonald's and other factor analysis-based coefficients behave very similarly to the glb and have the same bias problem, and, in addition, rely on a single-factor hypothesis, which is either undecided or false, the glb is to be preferred.

Bias Correction of the glb

Although the glb seems superior to single-factor-based coefficients of reliability, when the test is hypothesized to be unidimensional, this does not

mean that the glb must be evaluated routinely for the single administration of an arbitrary test. The glb has gained little popularity, mainly because of the sampling bias problem. Bias correction methods are under construction [12, 22], but still have not reached the level of accuracy required for practical applications. The problem is especially bad when the number of items is large relative to sample size. Until these bias problems are over, alpha will prevail as the lower bound to reliability.

Reliability versus Unidimensionality

Reliability is often confused with unidimensionality. A test can be congeneric, a property of the true score parts of the items, yet have large error variances, a property of the error parts of the items, and the reverse is also possible. Assessing the degree of unidimensionality is a matter of assessing how closely the single factor fits in common factor analysis. Ten Berge and Sočan [20] have proposed a method of expressing unidimensionality as the percentage of common variance explained by a single factor in factor analysis, using the so-called Minimum Rank Factor Method of Ten Berge and Kiers [18]. However, this is a matter of taste and others prefer goodness-of-fit measures derived from maximum-likelihood factor analysis.

References

- [1] Bekker, P.A. & De Leeuw, J. (1987). The rank of reduced dispersion matrices, *Psychometrika* **52**, 125–135.
- [2] Bentler, P.M. (2003). Should coefficient alpha be replaced by model-based reliability coefficients? *Paper Presented at the 2003 Annual Meeting of the Psychometric Society*, Sardinia.
- [3] Bentler, P.M. & Woodward, J.A. (1980). Inequalities among lower bounds to reliability: with applications to test construction and factor analysis, *Psychometrika* **45**, 249–267.
- [4] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.
- [5] Gilmer, J.S. & Feldt, L.S. (1983). Reliability estimation for a test with parts of unknown lengths, *Psychometrika* **48**, 99–111.
- [6] Guttman, L. (1945). A basis for analyzing test-retest reliability, *Psychometrika* **10**, 255–282.
- [7] Guttman, L. (1958). To what extent can communalities reduce rank, *Psychometrika* **23**, 297–308.
- [8] Jackson, P.H. & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds, *Psychometrika* **42**, 567–578.
- [9] Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests, *Psychometrika* **36**, 109–133.
- [10] Kristof, W. (1974). Estimation of reliability and true score variance from a split of the test into three arbitrary parts, *Psychometrika* **39**, 245–249.
- [11] Ledermann, W. (1937). On the rank of reduced correlation matrices in multiple factor analysis, *Psychometrika* **2**, 85–93.
- [12] Li, L. & Bentler, P.M. (2004). *The greatest lower bound to reliability: Corrected and resampling estimators*. Paper presented at the 82nd symposium of the Behaviormetric Society, Tokyo.
- [13] McDonald, R.P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis, *British Journal of Mathematical and Statistical Psychology* **23**, 1–21.
- [14] Novick, M.R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements, *Psychometrika* **32**, 1–13.
- [15] Shapiro, A. (1982). Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis, *Psychometrika* **47**, 187–199.
- [16] Shapiro, A. & Ten Berge, J.M.F. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability, *Psychometrika* **65**, 413–425.
- [17] Spearman, C.E. (1927). *The Abilities of Man*, McMillan, London.
- [18] Ten Berge, J.M.F. & Kiers, H.A.L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix, *Psychometrika* **56**, 309–315.
- [19] Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis, *Psychometrika* **46**, 357–366.
- [20] Ten Berge, J.M.F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality, *Psychometrika* **69**, 611–623.
- [21] Van Zijl, J.M., Neudecker, H. & Nel, D.G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha, *Psychometrika* **65**, 271–280.
- [22] Verhelst, N.D. (1998). Estimating the reliability of a test from a single test administration, (Measurement and Research Department Report No. 98-2), Arnhem.
- [23] Wilson, E.B. & Worcester, J. (1939). The resolution of six tests into three general factors, *Proceedings of the National Academy of Sciences* **25**, 73–79.
- [24] Yuan, K.-H. & Bentler, P.M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates, *Psychometrika* **67**, 251–259.

JOS M.F. TEN BERGE

Teaching Statistics to Psychologists

GEORGE SPILICH

Volume 4, pp. 1994–1997

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Teaching Statistics to Psychologists

In the introduction to his 1962 classic statistical text, Winer [7] describes the role of the statistician in a research project as similar to that of an architect; that is, determining whether the efficacy of a new drug is superior to that of competing products is similar to designing a building with a particular purpose in mind and in each case, there is more than one possible solution. However, some solutions are more elegant than others and the particulars of the situation, whether they are actual patient data or the size and placement of the building site, place boundaries on what can and cannot be accomplished.

What is the best way to teach the science and art of designing and conducting data analysis to today's graduate students in psychology? Perhaps history can be our guide, and so we begin by first asking what graduate education in statistics was like when today's senior faculty were students, then ask what is current common practice, and finally ask what the future might hold for tomorrow's graduate students. I propose the following:

- Graduate training in statistics has greatly changed over the last few decades in ways that are both helpful and harmful to students attempting to master statistical methodology.
- The major factor in this change is the development of computerized statistical packages (*see Software for Statistical Analyses*) which, when used in graduate education, cause students trained in **experimental design** to be more broadly but less thoroughly trained.
- 'Point and click' statistical programs allow individuals without professional training access to procedures they do not understand. The availability of statistical packages to individuals without professional statistical training might lead to a guild environment where psychologists could play an important role.

The Recent Past

Perusal of 'classic' texts such as Lindquist [5], Kirk [4], McNemar [6], Winer [7], Guilford and Fruchter [1], Hayes [2], and Keppel [3] suggests that

30 to 35 years ago (or just one generation ago in terms of the approximate length of an academic career), psychology graduate education emphasized descriptive measures, correlational techniques, and especially multiple regression (*see Multiple Linear Regression*) and the **analysis of variance** including *post hoc* tests and trend analysis (*see Multiple Comparison Procedures*). Statistical techniques were taught via hand calculation methods using small data sets. Computationally demanding methods such as **factor analysis**, time series analysis, **discriminant** and cluster analysis (*see Cluster Analysis: Overview*) as well as residual analysis (*see Residuals*) in multiple regression and multivariate analysis of variance (*see Multivariate Analysis: Overview*) were not typically presented to all students. These advanced techniques were generally presented to students whose professional success would depend upon mastery of those particular skills. For instance, a graduate student preparing for a career investigating personality, intelligence, or social psychology was much more likely to receive thorough training in factor analytic techniques (*see History of Factor Analysis: A Psychological Perspective; Factor Analysis: Confirmatory*) than a student studying classical or operant conditioning. There is little need for understanding the difference between varimax and orthomax rotations in factor analysis if one is pointing to a career observing rats in operant chambers. Statistical training of that era was subdiscipline specific.

For all students of this bygone era, training heavily emphasized experimental design. Graduate students of 30–35 years ago were taught that the best way to fix a problem in the data was not to get into trouble in the first place, and so the wise course of action was to employ one of several standard experimental designs. Texts such as Lindquist [5] and Kirk [4] instructed several decades' worth of students in the pros and cons of various experimental designs as well as the proper course of analysis for each design.

Current State

What has changed in graduate training since then and why? The answer is that computer-based statistical packages have become commonplace and altered the face of graduate statistical education in psychology. In times past, data analysis was so time consuming

2 Teaching Statistics to Psychologists

and labor intensive that it had to be planned carefully, prior to conducting the actual study. In that era, one could not easily recover from design errors through statistical control, especially when the size of the data set was large. Today, studies involving neuroimaging techniques such as fMRI and EEG result in tens of thousands and even millions of data points per subject, all potentially requiring a baseline correction, an appropriate transformation, and, possibly, covariates. The sheer hand labor of such an analysis without computer assistance is beyond imagination.

Currently, most graduate programs in psychology expect their students to gain competency in some statistical package, although there is a considerable amount of variability in how that goal is met. A survey of 60 masters and doctoral programs in psychology at colleges and universities in the United States provides a glimpse of the statistical packages used in graduate education. Individual faculty members who indicated that they were responsible for graduate training in statistics were asked if their department had a standard statistical package for graduate training. All 60 respondents replied that some statistical package was a part of graduate training and their responses are presented in Table 1.

These same faculty members were asked how graduate students gained mastery of a statistical package; their responses are presented in Table 2. While most graduate programs appear to have a formal course dedicated to teaching a particular statistical

package, some programs integrate the statistical package into laboratory courses or teach it as part of an advisor's research program. One rationale for teaching the use of a statistical package during research training instead of in a formal course is that with the former, students are likely to learn material appropriate to their immediate professional life. A contrasting view is that in a formal course that cuts across the domain boundaries of psychology, students are exposed to a wide variety of statistical techniques that may be useful in the future. Because many professionals start their career in one field but end up elsewhere in academia or in industry, breadth of skill may be preferable to depth, at least for the intermediate level of graduate training.

Statistical packages also provide graduate students with tools that were not typically a part of the common curriculum a generation ago. Respondents to the survey indicated that cluster analysis, discriminant analysis, factor analysis, binary logistic regression, and categorical regression are processes generally presented to graduate students because they have access to a statistical package. Furthermore, a wide assortment of two- and three-dimensional graphs and other visual techniques for exploring relationships are now presented to graduate students.

What is the Impact of Statistical Packages on Today's Graduate Education?

Providing graduate students with access to statistical packages has dramatically changed the breadth of their education. To assay the magnitude of the change, faculty in the survey were asked what topics they would keep or drop if their current statistical package was no longer available for graduate education. Faculty replied that anova, extensive discussion of *post hoc* analysis, orthogonal contrasts, power analysis and eta-squared (see **Effect Size Measures**) would still be covered but with very simple examples. Currently, they expect their students to fully master such techniques for a wide variety of settings. They also indicated that a variety of regression techniques would still be taught as would the fundamentals of factor analysis.

However, the survey respondents agreed almost unanimously that coverage of statistical techniques that are computationally intensive such as manova, cluster analysis, discriminant analysis, and complex

Table 1 Standard statistical packages in graduate programs ($N = 60$)

Statistical program of choice	Percent respondents
SPSS	45
SAS	15
Minitab	5
Systat	5
No standard	25
Other	5

Table 2 Methods that graduate programs use to ensure student mastery of statistical packages ($N = 60$)

How is a statistical package taught?	Percent respondents
Through a dedicated course	55
No course; through lab and research	25
Both course and lab/research	15
No standard method	5

factor analytic techniques (*see* **Factorial Designs; Repeated Measures Analysis of Variance**) would be reduced or eliminated if students did not have access to a statistical package. The comments of these professional educators suggest that today's graduate students are exposed to a wider range of statistical techniques and are considerably more adept at data manipulations and massaging than their counterparts of 30–35 years ago.

What is the Future?

The evolution of statistical programs suggests that they will become 'smarter' and will interact with the user by suggesting analyses for particular data sets. As the programs become easier to use and available to a wider audience, will biostatisticians still be needed? Perhaps the answer lies in Winer's [7] description of statisticians as architects. Today, one can easily purchase a computer program that helps design a kitchen, office, or home. Have these programs eliminated the need for architects? The answer is 'not at all', except for the simplest applications. When homeowners can use a software product to play with the design of an addition or a house, they are more likely to conceive of a project that requires the skill of a professional. The same may happen in biostatistics. Just as certification as an architect is necessary for a practitioner's license, we may see the emergence of a 'guild' that certifies professional statisticians after sufficient coursework and some demonstration of technical ability. Statistical packages can impart skill, but they cannot impart wisdom. The challenge for faculty who are training tomorrow's graduate students in biostatistics is to ensure that we impart more than knowledge of statistics; we need to teach the wisdom that comes from experience.

Conclusion

It would be easy to assume that the impact of statistical packages upon graduate education is merely to increase the range of techniques that are taught. The real sea change is in how the analytic process itself is approached conceptually. Before the advent of statistical packages, a student was taught to carefully plan analyses prior to actual computation, and this plan guided as well as constrained the design. In contrast, today's students are taught to break the data into various subsets and then to examine it from all angles. Such sifting and winnowing is so time consuming if performed by hand that extensive exploratory data analyses that are common today were all but impossible in the earlier era. Students now have the possibility of seeing more in the data than was possible a generation ago but at the cost of a deeper understanding of how the view was derived.

References

- [1] Guilford, J.P. & Fruchter, B. (1973). *Fundamental Statistics in Psychology and Education*, 5th Edition. McGraw-Hill, New York.
- [2] Hayes, W.L. (1973). *Statistics for the Social Sciences*, 2nd Edition, Harcourt Brace, New York.
- [3] Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook*, Prentice Hall, Englewood Cliffs.
- [4] Kirk, R.E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks-Cole, Belmont.
- [5] Lindquist, E.F. (1953). *Design and Analysis of Experiments in Psychology and Education*, Houghton-Mifflin, Boston.
- [6] McNemar, Q. (1969). *Psychological Statistics*, 4th Edition, Wiley, New York.
- [7] Winer, B.J. (1962). *Statistical Principles in Experimental Design*, McGraw-Hill, New York.

GEORGE SPILICH

Teaching Statistics: Sources

BERNARD C. BEINS

Volume 4, pp. 1997–1999

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Teaching Statistics: Sources

Teachers of statistics are aware of the dismal reputation of their discipline among students, but they work to improve it by displaying delight in the topic and by using clever demonstrations and activities. Fortunately, there is a population of useful, interesting material available. The books and periodicals reviewed here include sources that stand alone as fascinating reading for teachers and students, as well as books that appeal to teachers who want to enliven their classes. The comments here proceed from books to periodicals, with more general material preceding the more technical.

The 1954 classic by Huff [4], *How to lie with statistics*, is an entertaining depiction of statistics in everyday life. Its main limitation is that its content is, at times, quite outdated. Few of us would now be impressed with college graduates who earn a beginning salary of \$10 000. Although the examples do not always lend themselves to updating, they illustrate universal and timeless pitfalls. Instructors can generate current examples of these pitfalls.

A recent philosophical descendent of Huff's book is Best's [1] *Damned Lies and Statistics*, which provides many examples of dubious statistics and how they originated. The book opens with what Best considers the worst possible social statistic: the 'fact' that the number of children gunned down in the United States has doubled every year since 1950. The book documents the history of this number. It also includes sources of bad statistics, 'mutant' statistics that arise when original values are misinterpreted or distorted, and inappropriate comparisons using superficially plausible statistics. Best illustrates how statistics that are important to social issues can take on a life of their own and confuse rather than illuminate those issues. Students will enjoy this book and instructors can use it to enhance classroom presentations and discussions.

Paulos's [5] entertaining book, *Once upon a number: The hidden mathematical logic of stories*, takes the same approach as Best's. Both cite examples of events in everyday life that relate to social controversies and public policies. The anecdotes selected by Paulos illustrate the general principles of statistics

and probabilities (see **Probability: An Introduction**) found in many facets of life. He offers compelling examples of how to think using statistics. One of his important messages is that generalizations from a single instance can be faulty because they may be highly idiosyncratic. He also identifies questions that statistics can answer and how to appropriately apply those statistics. For instance, he discusses the statistics associated with why we invariably have to wait longer in line than we think we should, a phenomenon of interest to those of us who always think we pick the wrong line. Paulos invokes probability to show that minority populations are statistically (and realistically) more prone to encountering racism than are people in majority populations. Because many of the illustrations involve people and their behavior, this book is particularly useful for statistics classes in the behavioral and social sciences. This volume uses the same clear approach that Paulos took in his previous books on numerical literacy.

Readers get a more technical, but still highly readable, presentation of statistics and probability in Holland's [3] volume, *What are the chances?* Examples range widely. They include how the random distribution of rare events such as cancer creates clusters in a few locations (neighborhoods or buildings), the probabilities of Prussian soldiers being kicked to death by their horses between 1875 and 1894, and probabilities linked to waiting times in queues and traffic jams. The examples are supported by formulas, tables, and graphs. Holland emphasizes that the meaning of the data comes from their interpretation, but that interpretations are fraught with their own difficulties. Sampling is emphasized as is measurement error, which is explained quite nicely. Readers generally familiar with statistical notation and the rudiments of normal distributions (see **Catalogue of Probability Density Functions**) and factorials (i.e., instructors who have completed a graduate level statistics course) will be able to fathom the points Holland makes. This book shows both the use and limitations of statistics.

Both the Paulos and the Holland books can serve to augment classroom discussions. In addition, both are likely to be accessible to competent and motivated students who have mastered the rudimentary statistical concepts.

A quite different book by Gelman and Nolan [2], *Teaching statistics: A bag of tricks*, focuses on activities and demonstrations. The authors' audience is

2 Teaching Statistics: Sources

statistics teachers in secondary schools and colleges. An introductory topics section contains demonstrations and activities that match the contents of virtually any beginning statistics course in the behavioral and social sciences. The numerous demonstrations and activities illustrate important statistical concepts and provide excellent exercises in critical thinking. Consequently, the instructor can use varied exercises in different classes. The in-class activities include a simple memory demonstration that can be used to illustrate confidence intervals and an evaluation of newspaper articles that report data. The authors present the important message that researchers are bound by ethical principles in their use of data. The second segment of Gelman and Nolan's book gives the logistics for the successful implementation of the demonstrations and activities. This part of the book will certainly be of interest to beginning teachers, but it contains helpful hints for veterans as well.

Statistics teachers can benefit from two of the handbooks of articles reprinted from the journal *Teaching of Psychology* [6] and [7]. The two volumes devoted to teaching statistics contain nearly five dozen articles. These books resemble that of Gelman and Nolan in that they all feature activities and demonstrations that instructors have used successfully in their classes. The entries in the handbooks were written by a diverse set of statistics teachers and cover a wide variety of topics. The examples are broad enough to be useful to instructors in all of the behavioral sciences. The topics of many articles overlap with those of Gelman and Nolan, but the entries in the handbooks provide pedagogical advice as well as activities and demonstrations. For instance, there are sections on topics such as developing student skills, evaluating successes in statistics, and presenting research results. Many of the activities and demonstrations in these handbooks are accompanied by at least a basic empirical evaluation of their effects on student learning.

In addition to books, several periodicals publish activities, demonstrations, and lecture enhancements. The quarterly magazine *Chance* (not be confused with the gambling magazine with the same title) publishes articles on topics such as the misuse of statistics in the study of intelligence, **Bayesian statistics** in cancer research, teacher course evaluations, and the question of who wrote the 15th book of Oz. Feature articles illustrate the value of relying

on statistical knowledge to help address the important issues in our lives. The articles are typically very engaging, although the authors do not dumb down the content. Sometimes the reading requires diligence because of the complexity of the statistical issues, but the articles are worth the work. Most of the time, readers with a modicum of knowledge of statistical ideas and notation will find the writing accessible. (One vintage article [from 1997] discussed the statistics associated with waiting in lines for various services. Statisticians' fascination with time spent waiting suggests that they spend an inordinate amount of time doing nothing, but that they are quite productive during those times.) *Chance* has regular columns on sport, visual presentation of data, book reviews, and other topics.

A periodical with a more narrow orientation is the *Journal of Statistics Education*, published by the American Statistical Association three times a year. This online journal is available free on the Internet. Its featured articles involve some aspect of pedagogy. A recent volume included articles on teaching power and sample size, using the Internet in teaching statistics, the use of analogies and heuristics in teaching statistics, and the use of student-specific datasets in the classroom. In addition, a column 'Teaching Bits: A Resource for Teachers of Statistics' offers brief excerpts of current events that are of relevance to teaching statistics. The journal also offers data sets that can be downloaded from the Internet.

The journal *Teaching of Psychology* publishes articles on teaching statistics. As the journal title suggests, the material is oriented toward the discipline of psychology, but there are regular articles on teaching statistics that are suitable for other behavioral sciences. This journal is the organ of the Society for the Teaching of Psychology. A parallel journal, *Teaching Sociology*, also publishes occasional articles on teaching statistics and research methods.

The periodical, *American Statistician*, includes the 'Teacher's Corner'. The entries associated with teaching are often quite technical and involve advanced topics in statistics. They are less likely to be relevant to students in the behavioral sciences.

Psychological Methods appears quarterly. The articles are usually fairly technical and are addressed to sophisticated researchers. Recent topics included new approaches to regression analyses (see **Regression Models**) **meta-analysis**, and **item response**

theory (IRT). The material in this journal is best suited for advanced students. Although the intent of most articles is not pedagogical, the articles can help instructors bring emerging statistical ideas to their classrooms.

Finally, the Open Directory Project (<http://www.dmoz.org/Science/Math/Statistics/>) is an Internet site that provides a wealth of information on teaching statistics. This site provides a compendium of useful web addresses on a vast array of topics, including statistics education. The web pages to which the Open Directory Project sends you range from light-hearted and humorous (the three most common misspellings of statistics) to the very serious (e.g., an interactive Internet environment for teaching undergraduate statistics). There is also an interesting link to a web page that evaluates the use of statistics in the media. The Open Directory Project site has links to virtually any topic being covered in an introductory level statistics class, as well as links to more advanced topics for higher level students.

References

- [1] Best, J. (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*, University of California Press, Berkeley.
- [2] Gelman, A. & Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks*, Oxford University Press, Oxford.
- [3] Holland, B.K. (2002). *What are the Chances? Voodoo Deaths, Office Gossip, and other Adventures in Probability*, Johns Hopkins University Press, Baltimore.
- [4] Huff, D. (1954). *How to lie with Statistics*, Norton, New York.
- [5] Paulos, J.A. (1998). *Once upon a Number: The Hidden Mathematical Logic of Stories*, Basic Books, New York.
- [6] Ware, M.E. & Brewer, C.L. (1999). *Handbook for Teaching Statistics and Research Methods*, 2nd Edition, Erlbaum, Mahwah.
- [7] Ware, M.E. & Johnson, D.E. (2000). *Handbook of Demonstrations and Activities in the Teaching of Psychology, Vol. 1: Introductory, Statistics, Research Methods, and History*, 2nd Edition, Erlbaum, Mahwah.

BERNARD C. BEINS

Telephone Surveys

ROBYN BATEMAN DRISKELL

Volume 4, pp. 1999–2001

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Telephone Surveys

In recent years, telephone surveys have become increasingly popular and almost commonplace. When the methodology of telephone surveys was developed in the 1970s, many assumed that telephone surveys would replace face-to-face surveys for the most part. While this did not occur, telephone surveying is the preferred approach in many cases [2]. Telephone survey methodologies have undergone dramatic changes and examination in the last 20 years. Telephone survey methodology is widely used. As a result of recent advances in telephone technology, the methodology is viewed as both valid and reliable. Telephone surveys are efficient and effective means of collecting data that can aid decision-making processes in both the public and private sectors [5].

When designing the questionnaire, it is often easier and safer to borrow questions that have been used and tested previously (*see Survey Questionnaire Design*). This allows for comparability of the questions across time and place. The *introductory spiel* is the standardized introduction read by an interviewer when contact is made with a possible eligible household or respondent. A carefully worded introduction is of utmost importance. A weak introductory spiel can lead to refusals and nonresponse error. During the introduction, the potential respondent decides whether to cooperate [3]. The credibility of the interviewer and the survey must be established in the introduction. The introduction must reduce the respondent's fears and skepticism by providing assurances of legitimacy. Lyon suggests [5] that the introduction should *always* include (a) the interviewer's full name; (b) the organization and/or its sponsor that is conducting the research; (c) the survey's general topics; (d) the procedure for selection; (e) a screening technique; and (f) an assurance of confidentiality. Following the introduction, the screening technique selects a particular individual within the household to be interviewed. The screening technique is designed to systematically select respondents by age and sex, so that every individual in each sampled household has an equal probability of being selected, thereby ensuring a representative sample.

Advantages of Telephone Surveys

Telephone surveys have numerous advantages. The most important advantage is the ability to maintain quality control throughout the data collection process [4]. A second advantage is cost efficiency. Telephone surveys are much less expensive than face-to-face interviews but more expensive than mail surveys.

The third major advantage of telephone surveys is their short turn around time. The speed with which information is gathered and processed is much faster than that of any other survey method. A telephone survey takes 10 to 20% less time than the same questions asked in a face-to-face interview [4]. When a central interviewing facility with several phone banks is used, it is possible for 10 callers to complete approximately 100 interviews in an evening. Hence, large nationwide studies can be completed in a short time. By polling only a few hundred or a few thousand persons, researcher can obtain accurate and statistically reliable information about tens of thousands or millions of persons in the population. This assumes, of course, that proper techniques are implemented to avoid survey errors in sampling, coverage, measurement, and non-response [2].

A fourth advantage is the ability to reach most homes as a result of methodological advances in random-digit dialing (RDD) and the proliferation of phones. It is estimated by the US census that approximately 97% of US households have a telephone. RDD has improved telephone survey methodology. RDD uses a computer to select a telephone sample by random generation of telephone numbers. There are several different techniques for generating an RDD sample. The most common technique begins with a list of working exchanges in the geographical area from which the sample is to be drawn. The last four digits are computer generated by a random procedure. RDD procedures have the advantage of including unlisted numbers that would be missed if numbers were drawn from a telephone directory. Telephone numbers also can be purchased from companies that create the sampling pool within a geographic area, including numbers for RDD surveying (*see Survey Sampling Procedures*).

A fifth advantage is that telephone interviewers do not have to enter high-crime neighborhoods or

enter people's homes, as is required for face-to-face interviews. In some cases, a respondent will be more honest in giving socially disapproved answers if they do not have to face the interviewer. Likewise, it is possible to probe into more sensitive areas over the phone than it is in face-to-face interviews [1]. The major differences between telephone interviewing and face-to-face interviewing is that the interviewer's voice is the principal source of interviewing bias. By not seeing the interviewer, respondents are free from the biases that would be triggered by appearance, mannerisms, gestures, and expressions. On the other hand, the interviewer's voice must project an image of a warm, pleasant person who is stimulating to talk to and who is interested in the respondent's views.

Other advantages found in [5] include the ability to probe, the ability to ask complex questions with complex skip patterns (with computer-assisted telephone-interviewing, CATI), the ability to use long questionnaires, the assurance that the desired respondent completes the questionnaire, and the ability to monitor the interviewing process.

CATI is a survey method in which a printed questionnaire is not used. Instead, the questions appear on the screen of a computer terminal, and the answers are entered directly into a computer via the keyboard or mouse. The major advantages of this procedure are that it allows for the use of a complex questionnaire design with intricate skip patterns. It also provides instant feedback to the interviewer if an impossible answer is entered, and it speeds up data processing by eliminating intermediate steps. The computer is programmed not only to present the next question after a response is entered but also to determine from the response exactly which question should be asked next. The computer branches automatically to the next question according to the filter instructions. CATI can randomly order the sequence of possible response categories and incorporate previous answers into the wording of subsequent items. The CATI software also can control the distribution of the sampling pool and dial the appropriate phone number for the interviewer. During the polling process, the supervisory staff are able to access the interviewer's completed calls, the duration of each call, response rates, and listen in to assure the accuracy of the script. When CATI is properly implemented, the quality of

the data is improved and survey errors are often reduced [2, 4, 5].

Disadvantages of Telephone Surveys

Despite the numerous advantages, telephone surveys have limitations. During a telephone interview, it is not possible to use visual aids. The length of the survey is another concern. Owing to respondent fatigue, most telephone surveys should not be longer than 15 min. Face-to-face interviews can be longer, up to 20 to 30 min. In face-to-face interviews, the interviewer can assess body language and notice respondent fatigue. While complex skip patterns are easy to use with the CATI system, long, involved questions with many response categories are hard to follow in a telephone interview. The telephone methodology is hampered by the proliferation of marketing and sales calls veiled as 'research'. Many respondents distrust telephone surveys and want to know what you are selling. Also, a shortcoming of the telephone survey is the ease with which a potential respondent can hang up. It is quite easy to terminate the interview or make up some excuse for not participating at that time [1]. Another problem is the potential coverage error – every demographic category (i.e., sex, age, and gender) is not equally willing to answer the phone and complete a survey – although this can be avoided with an appropriate screening technique. Technological advances such as answering machines and caller ID have contributed to nonresponse rates as potential respondents screen incoming calls. Despite the several limitations of telephone survey methods, the advantages usually outweigh the disadvantages, resulting in an efficient and effective method for collecting data.

References

- [1] Babbie, E. (1992). *The Practice of Social Research*, 6th Edition, Wadsworth Publishing Company, Belmont.
- [2] Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*, 2nd Edition, John Wiley, New York.
- [3] Dijkstra, W. & Smit, J.H. (2002). Persuading reluctant recipients in telephone surveys, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge, J.L. Little, & J.A. Roderick, eds, John Wiley, New York.
- [4] Lavrakas, P.J. (1998). Methods for sampling and interviewing in telephone surveys, in *Handbook of Applied*

Social Research Methods, L. Brickman & D.J. Rog, eds, Sage Publications, Thousand Oaks, pp. 429–472.

(See also **Randomized Response Technique**)

- [5] Lyon, L. (1999). *The Community in Urban Society*, Waveland Press, Prospect Heights.

ROBYN BATEMAN DRISKELL

Test Bias Detection

MICHAL BELLER

Volume 4, pp. 2001–2007

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Test Bias Detection

During the 1970s considerable attention was given to developing fair selection models in the context of college admissions and job entry. These models put heavy emphasis on predictive validity, and in one way or another they all address the possibility of differences in the predictor–criterion relationship for different groups of interest. With the exception of Cleary's regression model, most other models proposed were shown to be mutually contradictory in their goals and assumptions.

There is a growing consensus in the measurement community that bias refers to any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees. In regard to test use, the core meaning of fairness is *comparable validity*: A fair test is one that yields comparable and valid scores from person to person, group to group, and setting to setting. However, fairness, like validity, is not just a psychometric issue. It is also a social value, and therefore alternative views about its essential features will persist.

As the use of educational and psychological tests continues to grow, an increasing number of decisions that have profound effects on individuals' lives are being made based on test scores, despite the fact that most test publishers caution against the use of a single score for decision-making purposes. Tests are instruments that are designed to provide evidence from which inferences are drawn and on the basis from which such decisions are made. The degree to which this evidence is credible constitutes the validity of the test, and it should hold for all groups among the intended test-taking population (*see Validity Theory and Applications*).

Concerns of possible gender and/or ethnic biases in the use of various tests have drawn the attention of test users and test developers as well as of the public in general. A view of fairness, frequently used by the public, involves equality of testing outcomes. In the public debate on this issue the terms *test unfairness*, *cultural unfairness*, and *test bias* are often used interchangeably to refer to differences in test performance among subgroups of social interest. However, the idea that fairness requires overall performance or passing rates to be equal across groups is not the one generally

accepted in the professional literature. In fact, Cole and Zieky [12] claim that: 'If the members of the measurement community currently agree on any aspect of fairness, it is that score differences are not proof of bias' (p. 375). This is because test outcomes may validly document group differences that are real and that may be reflective, in part, of unequal opportunity to learn and other possible reasons.

Bias refers to any construct¹ under-representation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers. Construct-irrelevant variance exists when the 'test contains excess reliable variance that is irrelevant to the interpreted construct' [30]. The effect of such irrelevant sources of variance on scores is referred to as *measurement bias*. Sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups are of potential concern for both predictors and criteria. Determining whether measurement bias is present is often difficult, as this requires evaluating an observed score in relation to the unobservable construct of interest.

Fairness in Terms of Equitable Treatment of All Examinees

One way to reduce measurement bias is to assure equitable treatment of all examinees through test design and development practices intended from an early stage to prevent bias. Equity in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and providing reasonable accommodations for test-takers with disabilities when appropriate, are important aspects of fairness. Equitable treatment of all examinees is directly related to establishing construct validity and the fairness of the testing process [42]. It is useful to distinguish two kinds of comparability relevant to construct validity and fairness – *comparability of score interpretation* and *task comparability*. *Comparability of score interpretation* means that the properties of the test itself and its external relationships with other variables are comparable across groups and settings. Comparability of score interpretation is important in justifying uniform score use for different groups and in different circumstances. *Task comparability* means that the test task elicits the same cognitive processes across different groups and different circumstances.

Fairness of Accommodations. Within task comparability, two types of processes may be distinguished: those that are relevant to the construct measured and those that are irrelevant to the construct but nonetheless involved in task performance (possibly like reading aloud in mathematics). Comparability of construct-relevant processes across groups is necessary for validity. Ancillary or construct-irrelevant processes may, however, be modified without jeopardizing score interpretation. This provides a fair and legitimate basis for accommodating tests to the needs of students with disabilities and those who are English-language learners [40]. Thus, comparable validity and test fairness do not necessarily require identical task conditions, but rather common construct-relevant processes with ignorable construct-irrelevant or ancillary processes that may be different across individuals and groups. Such accommodations must be justified with evidence that score meanings have not been eroded in the process.

The general issue of improving the accessibility of a test must be considered within an assessment design framework that can help the assessment planner maximize validity within the context of specific assessment purposes, resources, and other constraints. Evidenced-centered assessment design (ECD), which frames an assessment as embodying an evidentiary argument, has been suggested as a promising approach in this regard [21].

Detecting Bias at the Item Level. Statistical procedures to identify test items that might be biased have existed for many years [3, 29]. One approach to examining measurement bias at the item level is to perform a **differential item functioning (DIF)** analysis, focusing on the way that comparable or *matched* people in different groups perform on each test item. DIF procedures are empirical ways to determine if the item performance of comparable subgroups is different. DIF occurs when a statistically significant difference in performance on an individual test item occurs across two or more groups of examinees, *after* the examinees have been matched on total test/subtest scores [20, 22]. DIF methodologies and earlier methods share a common characteristic in that they detect unfairness of a single item relative to the test as a whole, rather than detecting pervasive unfairness. Thus, in the extremely unlikely case that all items were biased to exactly the same degree against

exactly the same groups, no items would be identified as unfair by current DIF methods [42].

DIF analysis is not appropriate in all testing situations. For example, it requires data from large samples and if only for this reason DIF analyses are more likely to be used in large-scale educational settings and are not likely to become a routine or expected part of the test development and validation process in employment settings.

DIF methods have been an integral part of test development procedures at several major educational test publishers since the 1980s [41]. Empirical research in domains where DIF analyses are common has rarely found sizable and replicable DIF effects [34]. However, it is worth noting that aggregations of DIF data have often led to generalizations about what kind of items to write and not to write. This in turn led to two outcomes: (a) fewer items with significant DIF values were found, because those with large DIF values were no longer being written; and (b) test fairness guidelines, which had previously been very distinct from empirical evaluations of test fairness, began to incorporate principles and procedures based on empirical (DIF) findings, rather than rely exclusively on the touchstone of social consensus on removing offensive or inappropriate test material.

Linked to the idea of measurement bias at the item level is the concept of an item sensitivity review. The fundamental purpose of fairness reviews is to implement the social value of not presenting test-takers with material that they are likely to find offensive or upsetting. In such a review, items are reviewed by individuals with diverse perspectives for language or content that might have differing meanings to members of various subgroups. Even though studies have noted a lack of correspondence between test items identified as possibly biased by statistical and by judgmental means [7, 37], major test publishers in the United States have instituted judgmental review processes designed to identify possibly unfair items and offensive, stereotyping, or alienating material (e.g., [16]).

Fairness Through the Test Design Process

It is important to realize that group differences can be affected intentionally or unintentionally by choices made in test design. Many choices made in the design of tests have implications for group differences. Such

differences are seldom completely avoidable. Fair test design should, however, provide examinees comparable opportunity, insofar as possible, to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test [39]. Given that in many cases there are a number of different ways to predict success at most complex activities (such as in school or on a job), test developers should carefully select the relevant constructs and the ways to operationalize the measurement of those constructs to provide the best opportunity for all subgroups to demonstrate their knowledge.

As guidance to the test-construction process, *ETS Standards for Quality and Fairness* state that: 'Fairness requires that construct-irrelevant personal characteristics of test-takers have no appreciable effect on test results or their interpretation' (p. 17) [16]. More specifically, ETS standards recommends adopting the following guidelines: (a) treat people with respect in test materials; (b) minimize the effects of construct-irrelevant knowledge or skills; (c) avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting; (d) use appropriate terminology to refer to people; (e) avoid stereotypes; and (f) represent diversity in depictions of people.

Fairness as Lack of Predictive Bias

During the mid-1960s, concurrent with the Civil Rights Movement, measurement professionals began to pay increasing attention to score differences on educational and psychological tests among groups (often referred to as *adverse impact*). Considerable attention has been given to developing fair selection models in the context of college admissions and job entry. These models put heavy emphasis on predictive validity, and in one way or another they all address the possibility of differences in the predictor-criterion relationship for different groups of interest.

Fair Selection Models – mid-1960s and 1970s.

Differential prediction (also called *predictive bias*; see AERA/APA/NCME *Standards*, 1999, for definition [2]) by race and gender in the use of assessment instruments for educational and personnel selection has been a long-standing concern. In trying to explicate the relation between prediction and selection, Willingham and Cole [39] and Cole [10] have pointed out that prediction has an obvious and important bearing on selection, but it is not the same thing –

prediction involves an expected level of criterion performance given a particular test score; selection involves the use of that score in decision-making. Cleary's [8] model stipulated that no predictive bias exists if a common regression line can describe the predictive relationship in the two groups being compared. Other selection models [9, 14, 26, 38] were developed in the 1970s taking into account the fact that any imperfect predictor will fail to select members of a lower-scoring group in proportion to their rate of criterion success. The models all require setting different standards of acceptance for individuals in different groups to achieve group equity as defined by the model. Petersen and Novick [32] pointed out that the various models are fundamentally incompatible with one another at the level of their goals and assumptions and lead to contradictory recommendations, unless there is perfect prediction – and even given perfect prediction, these are competing selection models based on differing selection goals, and will thus always be mutually contradictory. As a result of the continuous public and professional discussion around test bias there was a growing recognition that fair selection is related to fundamental value differences. Several utility models were developed that go beyond the above selection models in that they require specific value positions to be articulated [13, 18, 32, 35]. In this way, social values are explicitly incorporated into the measurement involved in selection models. Such models were seldom utilized in practice, as they require data that are difficult to obtain. More importantly, perhaps, they require application of dichotomous definitions of success and failure that are themselves methodologically and conceptually rather arbitrary and problematic.

After the intense burst of fairness research in the late 1960s and early 1970s described above, Flaugher [17] noted that there was no generally accepted definition of the concept 'fair' with respect to testing. Willingham and Cole [39] concluded that the effort to determine which model, among the variety of such models proposed, best represented fair selection was perhaps the most important policy debate of the 1970s among measurement specialists.

Current View of Test Bias

The contemporary professional view of bias is that it is an aspect of validity. A commonly used definition of test bias is based on a lack of predictive

bias. The AERA/APA/NCME Standards [2] defines predictive bias as ‘the systematic under- or overprediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance’ (p 179). This perspective (consistent with the Cleary model) views predictor use as unbiased if a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest. If the predictive relationship differs in terms of either slopes or intercepts, bias exists because systematic errors of prediction would be made on the basis of group membership [11, 19, 25, 27, 28, 31]. Whether or not subgroup differences on the predictor are found, predictive bias analyses should be undertaken when there are questions about whether a predictor and a criterion are related in the same way for relevant subgroups.

Several technical concerns need to be considered when trying to quantify predictive bias:

1. Analysis of predictive bias requires an unbiased criterion. It is possible for the same bias to exist in the predictor and the criterion, or it may be that there are different forms of bias in the criterion itself across groups, and that these biases might, however unlikely that may be, would cancel each other out.
2. The issue of statistical power to detect slope and intercept differences should be taken into account. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor or criterion unreliability are factors contributing to low power [1].
3. When discussing bias against a particular group in the admissions process, the entire set of variables used in selection should be taken into consideration, not just parts of it. The inclusion of additional variables may dramatically change the conclusions about predictive bias [26, 33].

Differential prediction by race has been widely investigated in the domain of cognitive ability. For White-African American and White-Hispanic comparisons, slope differences are rarely found. While intercept differences are not uncommon, they typically take the form of overprediction of minority group performance [4, 15, 23, 24, 36]. Similar results were found by Beller [5, 6] for the Psychometric Entrance Test (PET) used as one of two components for admissions to higher education in Israel.

There were few indications of bias in using PET, and when found, they were in *favor* of the minority groups. In other words, the use of a common regression line overpredicted the criterion scores for the minority groups.

Studies investigating sex bias in tests (e.g., [40]) found negligible opposing effects of under- and overprediction of criterion scores when using general scholastic test scores and achievement scores, respectively. These opposing effects tend to offset each other, and it is therefore not surprising that an actual admission score, which often consists of both general scholastic test and achievement scores, is generally unbiased (e.g., [5]).

Conclusion

Much of the criticism of psychological tests is derived from the observation that ethnic and gender groups differ extensively in test performance. Criticism is generally stronger if the groups that show relatively poor performance are also socioeconomically disadvantaged. Much of the polemic concerning test bias confuses several issues: (a) differences in test performance among groups are often regarded, in and of themselves, as an indication of test bias, ignoring performance on the external criterion that the test is designed to predict. Often, groups that perform poorly on tests also tend to perform poorly on measures of the criterion. Furthermore, analysis of the relationship between tests and criteria often reveals similar regression lines for the various groups; and (b) the issue of test bias is often confused with the possibility of bias in the content of some individual items included in a test. Group differences in test performance are attributed to specific item content, and rather than eliminating problematic items in a systematic way (i.e., checking all items for differential performance), this confusion has, in some cases, led to suggestions of a wholesale rejection of reliable and valid test batteries.

Nevertheless, there is a growing recognition in the measurement community that, even though score differences alone are not proof of bias, score differences may not all be related to the measured constructs; and that even valid differences may be misinterpreted to the detriment of the lower-scoring groups when scores are used for important purposes such as selection. The fundamental question remains the degree

to which the observed group differences reflect real, underlying psychological processes, and the degree to which group differences simply reflect the way the tests were constructed.

The current edition of the *Standards for Educational and Psychological Testing* [2] indicates that fairness 'is subject to different definitions and interpretations in different social and political circumstances'. With regard to test use, the core meaning of fairness is comparable validity: A fair test is one that yields comparable and valid scores from person to person, group to group, and setting to setting. However, fairness, like validity, is not just a psychometric issue. It also rests on social values. Thus alternative views about its essential features will persist.

Note

1. A construct is a set of knowledge, skills, abilities, or traits a test is intended to measure.

References

- [1] Aguinis, H. & Stone-Romero, E.F. (1997). Methodological artifacts in moderated multiple regression and their effects of statistical power, *Journal of Applied Psychology* **82**, 192–206.
- [2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.
- [3] Angoff, W.H. & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude, *Journal of Educational Measurement* **10**, 95–106.
- [4] Bartlett, C.J., Bobko, P., Mosier, S.B. & Hanna, R. (1978). Testing for fairness with a moderated multiple regression strategy, *Personnel Psychology* **31**, 233–242.
- [5] Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities, *Educational Measurement: Issues and Practice* **13**(2), 12–20.
- [6] Beller, M. (2001). Admission to higher education in Israel and the role of the psychometric entrance test: educational and political dilemmas, *Assessment in Education* **8**(3), 315–337.
- [7] Bond, L. (1993). Comments on the O'Neill and McPeck paper, in *Differential Item Functioning*, P. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale.
- [8] Cleary, T.A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges, *Journal of Educational Measurement* **5**, 115–124.
- [9] Cole, N.S. (1973). Bias in selection, *Journal of Educational Measurement* **10**, 237–255.
- [10] Cole, Nancy. (1981). Bias in testing, *American Psychologist* **36**, 1067–1077.
- [11] Cole, N.S. & Moss, P.A. (1989). Bias in test use, in *Educational Measurement*, 3rd Edition, R.L. Linn, ed., Macmillan Publishing, New York, pp. 201–219.
- [12] Cole, N.S. & Zieky, M. (2001). The new faces of fairness, *Journal of Educational Measurement* **38**, 369–382.
- [13] Cronbach, L.J. (1976). Equity in selection – where psychometrics and political philosophy meet, *Journal of Educational Measurement* **13**, 31–41.
- [14] Darlington, R.B. (1971). Another look at "cultural fairness.", *Journal of Educational Measurement* **8**, 71–82.
- [15] Dunbar, S. & Novick, M. (1988). On predicting success in training for men and women: examples from Marine Corps clerical specialties, *Journal of Applied Psychology* **73**, 545–550.
- [16] Educational Testing Service. (2002). *ETS Standards for Quality and Fairness*, ETS, Princeton.
- [17] Flaugher, R.L. (1978). The many definitions of test bias, *American Psychologist* **33**, 671–679.
- [18] Gross, A.L. & Su, W. (1975). Defining a "fair" or "unbiased" selection model: a question of utilities, *Journal of Applied Psychology* **60**, 345–351.
- [19] Gulliksen, H. & Wilks, S.S. (1950). Regression tests for several samples, *Psychometrika* **15**, 91–114.
- [20] Hambleton, R.K. & Jones, R.W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning, *Educational Research Quarterly* **18**, 21–36.
- [21] Hansen, E.G. & Mislevy, R.J. (2004). Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and English language learners, in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, San Diego.
- [22] Holland, P. & Wainer, H., eds (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale.
- [23] Houston, W.M. & Novick, M.R. (1987). Race-based differential prediction in air force technical training programs, *Journal of Educational Measurement* **24**, 309–320.
- [24] Hunter, J.E., Schmidt, F.L. & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing, in *Perspectives on Mental Testing*, C.R. Reynolds & R.T. Brown, eds, Plenum Press, New York, pp. 41–99.
- [25] Lautenschlager, G.J. & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction, *Applied Psychological Measurement* **10**, 133–139.
- [26] Linn, R.L. (1973). Fair test use in selection, *Review of Educational Research* **43**, 139–164.
- [27] Linn, R.L. (1978). Single group validity, differential validity and differential prediction, *Journal of Applied Psychology* **63**, 507–512.
- [28] Linn, R.L. (1984). Selection bias: multiple meanings, *Journal of Educational Measurement* **8**, 71–82.

- [29] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison Wesley, Reading.
- [30] Messick, S. (1989). Validity, in *Educational Measurement*, 3rd Edition, R.L. Linn, ed., Macmillan Publishing, New York, pp. 13–103.
- [31] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, McGraw-Hill, New York.
- [32] Petersen, N.S. & Novick, M.R. (1976). An evaluation of some models for culture fair selection, *Journal of Educational Measurement* **13**, 3–29.
- [33] Sackett, P.R., Laczko, R.M. & Lippe, Z.P. (2003). Differential prediction and the use of multiple predictors: the omitted variables problem, *Journal of Applied Psychology* **88**(6), 1046–1056.
- [34] Sackett, P.R., Schmitt, N., Ellingson, J.E. & Kabin, M.B. (2001). High stakes testing in employment, credentialing, and higher education: prospects in a post-affirmative action world, *American Psychologist* **56**, 302–318.
- [35] Sawyer, R.L., Cole, N.S. & Cole, J.W.L. (1976). Utilities and the issue of fairness in a decision theoretic model for selection, *Journal of Educational Measurement* **13**, 59–76.
- [36] Schmidt, F.L., Pearlman, K. & Hunter, J.E. (1981). The validity and fairness of employment and educational tests for Hispanic Americans: a review and analysis, *Personnel Psychology* **33**, 705–724.
- [37] Shepard, L.A. (1982). Definitions of bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., Johns Hopkins University Press, Baltimore, pp. 9–30.
- [38] Thorndike, R.L. (1971). Concepts of culture fairness, *Journal of Educational Measurement* **8**, 63–70.
- [39] Willingham, W.W. & Cole, N.S. (1997). *Gender and Fair Assessment*, Lawrence Erlbaum, Mahwah.
- [40] Willingham, W.W., Ragosta, M., Bennett, R.E., Braun, H., Rock, D.A. & Powers, D.E. (1988). *Testing Handicapped People*, Allyn & Bacon, Boston.
- [41] Zieky, M. (1993). Practical questions in the use of DIF statistics in test development, in *Differential Item Functioning*, P. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 337–347.
- [42] Zieky, M., & Carlton, S. (2004). In search of international principles for fairness review of assessments, in *Paper Presented at the IAEA*, Philadelphia.

Further Reading

ETS Fairness Review Guidelines. (2003). Princeton, NJ: ETS. Retrieved February, 10, 2005, from <http://ftp.ets.org/pub/corp/overview.pdf>

MICHAL BELLER

Test Construction

MARK J. GIERL

Volume 4, pp. 2007–2011

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Test Construction

Item response theory (IRT) (*see Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data*) provides an appealing conceptual framework for test construction due, in large part, to the item and test information function. The *item information function* provides a measure of how much psychometric information an item provides at a given ability level, θ . For dichotomously-scored items calibrated using the three-parameter logistic IRT model, the item information function for item i is calculated as:

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{(c_i + e^{Da_i(\theta - b_i)})(1 + e^{-Da_i(\theta - b_i)})^2}, \quad (1)$$

where $D = 1.7$, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-chance parameter [9]. (To illustrate key concepts, the three-parameter logistic IRT model is used because it often provides the best fit to data from multiple-choice tests. The item and test information function for select polytomous item response models are described in Chapters 2 through 9 of van der Linden and Hambleton [12]). For any given θ , the amount of information increases with larger values of a_i and decreases with larger values of c_i . That is, item discrimination reflects the amount of information an item provides assuming the pseudo-chance level is relatively small.

The *test information function* is an extension of the item information function. The test information function is the sum of the item information functions at a given θ :

$$I(\theta) = \sum_{i=1}^n I_i(\theta), \quad (2)$$

where $I_i(\theta)$ is the item information and n is the number of test items. This function defines the relationship between ability and the psychometric information provided by a test. The more information each item contributes, the higher the test information function. The test information function is also related directly to measurement precision because the amount of information a test provides at a given θ is inversely proportional to the precision with which

ability is estimated at that θ -value, meaning:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \quad (3)$$

where $SE(\theta)$ is the **standard error** of estimation. $SE(\theta)$ is the standard deviation of the asymptotically normal distribution for those examinees who have a maximum likelihood estimate of θ . The standard error of estimation can be used to compute a **confidence interval** for the corresponding θ -values across the score scale, which promotes a more accurate interpretation of the ability estimates. The standard error of estimation varies across ability level, unlike the standard error of measurement in classical test theory which is constant for all ability levels, because test information frequently varies across ability level.

Basic Approach to Test Construction Using Item and Test Information Functions

Both the item and test information functions are used in test construction. Lord [8] outlined the following four-step procedure, first suggested by Birnbaum [4], for designing a test using calibrated items from an item bank:

- Step 1: Decide on the shape desired for the test information function. The desired function is called the *target information function*. Lord [8] called the *target information function* a *target information curve*.
- Step 2: Select items with item information functions that will fill the hard-to-fill areas under the target information function.
- Step 3: Cumulatively add up the item information functions, obtaining at all times the information function for the part-test composed of items already selected.
- Step 4: Continue until the area under the target information function is filled up to a satisfactory approximation.

Steps 3 and 4 are easily understood, but steps 1 and 2 require more explanation.

The shape of the target information function, identified in Step 1, must be specified with the purpose of the test in mind. For example, a norm-referenced test designed to evaluate examinees across a broad

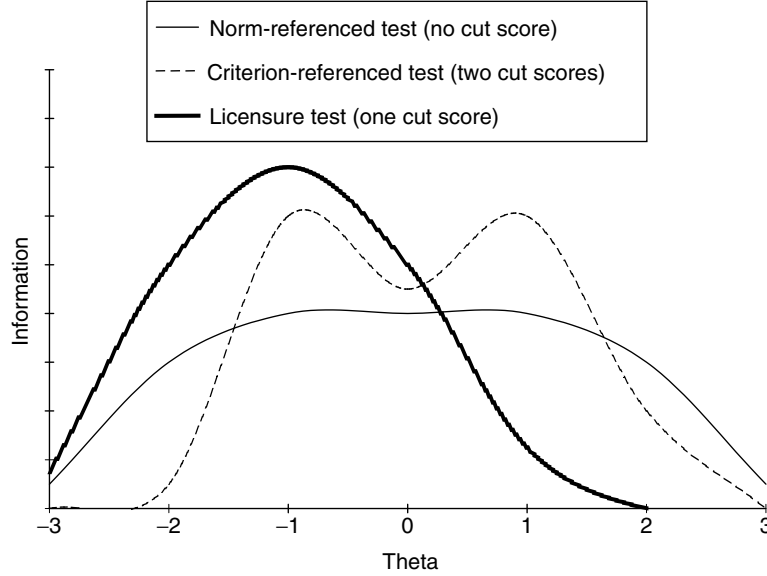


Figure 1 Target information function for three hypothetical tests designed for different purposes

range of ability levels would have a uniform target information function that spans much of the θ -scale. A criterion-referenced test designed to differentiate examinees at an ‘acceptable standard’ located at $\theta = -1.0$ and a ‘standard of excellence’ located at $\theta = 1.0$ would, by comparison, have a target information function with two information peaks near the θ cut scores associated with these two standards. A licensure test designed to identify minimally competent examinees, which could be operationally defined as a score above $\theta = -1.0$, would have a target information function with one peak near the θ cut score. These three hypothetical target information functions are illustrated in Figure 1.

Once the target information function is specified in step 1, then item selection in step 2 can be conducted using one of two statistically-based methods. The first item selection method is maximum information. It provides the maximum value of information for an item regardless of its location on the θ -scale. For the three-parameter logistic IRT model, maximum information is calculated as:

$$I_i(\theta)_{\text{MAX}} = \frac{D^2 a_i^2}{8(1 - c_i)^2} \times [1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{3/2}], \quad (4)$$

where $D = 1.7$, a_i is the discrimination parameter, and c_i is the pseudo-chance parameter [9]. Maximum information is often used in test construction because it provides a method for selecting the most discriminating items from an item bank. However, one problem that can arise when using the most discriminating items is estimation bias: Items with large expected a_i parameters are likely to overestimate their true a_i values because the correlation between the expected and true parameters is less than 1.0. Estimation bias is problematic when developing a test using the most discriminating items [i.e., items with the largest $I_i(\theta)_{\text{MAX}}$ values] because the a_i parameters will be inflated relative to their true values and, as a result, the test information function will be overestimated [6, 7]. This outcome could lead to overconfidence in the accuracy of the examinees’ ability estimates, given the test information function at a given θ is inversely proportional to the precision of measurement at that ability level.

The second item selection method is theta maximum. It provides the location on the θ -scale where an item has the most information. For the three-parameter model, theta maximum is calculated as:

$$\theta_i(I)_{\text{MAX}} = b_i + \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2}, \quad (5)$$

Table 1 Item parameters, maximum information, and theta maximum values for four example items

Item	a -parameter	b -parameter	c -parameter	$I_i(\Theta)_{\text{MAX}}$	$\theta_i(I)_{\text{MAX}}$
1	0.60	-1.10	0.20	0.18	-1.01
2	0.70	0.14	0.19	0.25	0.25
3	0.64	0.91	0.19	0.21	1.01
4	0.50	-2.87	0.19	0.13	-2.79

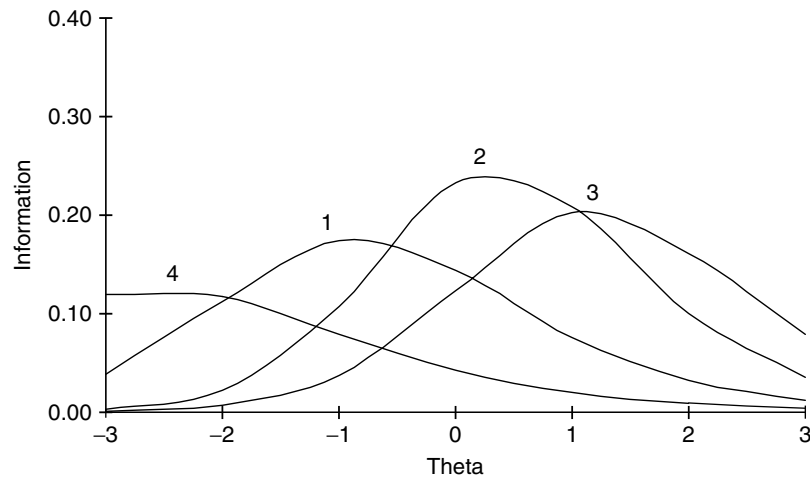
where $D = 1.7$, $\ln$ is the natural logarithm, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the pseudo-chance parameter [9]. Theta maximum is influenced primarily by the difficulty parameter because it reflects the location (i.e., b_i value) rather than the height (i.e., a_i value) of the item information function. Moreover, the b_i estimates tend to be more accurate than the a_i estimates. Therefore, theta maximum often contains less estimation bias than maximum information thereby producing a more consistent estimate of the test information function [5] which, in turn, yields a more reliable measure of θ .

A simple example helps illustrate the differences between $I_i(\theta)_{\text{MAX}}$ and $\theta_i(I)_{\text{MAX}}$. Table 1 contains the a -, b -, and c -parameter estimates for four items along with their $I_i(\theta)_{\text{MAX}}$ and $\theta_i(I)_{\text{MAX}}$ values. Figure 2 shows the information function for each item. If the goal was to select the most discriminating item from this set, then item 2 would be chosen because it has maximum information [i.e., $I_i(\theta)_{\text{MAX}} = 0.25$]. If, on the other hand, the goal was to select the

item that was most discriminating at $\theta = -1.0$, then item 1 would be chosen because it has maximum information around this point on the theta scale [i.e., $\theta_i(I)_{\text{MAX}} = -1.01$]. Notice that item 1 is not the most discriminating item, overall, but it does yield the most information at $\theta = -1.0$, relative to the other three items.

Developments in Test Construction Using Item and Test Information Functions

Rarely are tests created using item selection methods based on statistical criteria alone. Rather, tests must conform to complex specifications that include content, length, format, item type, cognitive levels, reading level, and item exposure, in addition to statistical criteria. These complex specifications, when combined with a large bank of test items, can make the test construction task, as outlined by Lord [8], a formidable one because items must be selected to meet a statistical target information function while at the same time

**Figure 2** Information functions for four items

satisfying a large number of test specifications (e.g., content) and constraints (e.g., length). Optimal test assembly procedures have been developed to meet this challenge [11]. Optimal test assembly requires the optimization of a test attribute (e.g., target information function) using a unique combination of items from a bank. The goal in optimal test assembly is to identify the set of feasible item combinations in the bank given the test specifications and constraints. The assembly task itself is conducted using an computer algorithm or heuristic. Different approaches have been developed to automate the item selection process in order to optimize the test attribute including 0–1 linear programming [1], heuristic-based test assembly [10], network-flow programming [2], and optimal design [3]. These advanced test construction procedures, which often use item and test information functions, now incorporate the complex specifications and constraints characteristic of modern test design and make use of computer technology. But they also maintain the logic inherent to Lord's [8] procedure demonstrating why the basic four-step approach is fundamental for *understanding* and *appreciating* developments and key principles in optimal test assembly.

References

- [1] Adema, J.J., Boekkooy-Timminga, E. & van der Linden, W.J. (1991). Achievement test construction using 0–1 linear programming, *European Journal of Operations Research* **55**, 103–111.
- [2] Armstrong, R.D., Jones, D.H. & Wang, Z. (1995). Network optimization in constrained standardized test construction, in *Applications of Management Science: Network applications*, Vol. 8, K. Lawrence & G.R. Reeves, eds, JAI Press, Greenwich, pp. 189–122.
- [3] Berger, M.P.F. (1998). Optimal design of tests with dichotomous and polytomous items, *Applied Psychological Measurement* **22**, 248–258.
- [4] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 397–479.
- [5] Gierl, M.J., Henderson, D., Jodoin, M. & Klinger, D. (2001). Minimizing the influence of item parameter estimation errors in test development: a comparison of three selection procedures, *Journal of Experimental Education* **69**, 261–279.
- [6] Hambleton, R.K. & Jones, R.W. (1994). Item parameter estimation errors and their influence on test information functions, *Applied Measurement in Education* **7**, 171–186.
- [7] Hambleton, R.K., Jones, R.W. & Rogers, H.J. (1993). Influence of item parameter estimation errors in test development, *Journal of Educational Measurement* **30**, 143–155.
- [8] Lord, F.M. (1977). Practical applications of item characteristic curve theory, *Journal of Educational Measurement* **14**, 117–138.
- [9] Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.
- [10] Luecht, R. (1998). Computer-assisted test assembly using optimization heuristics, *Applied Psychological Measurement* **22**, 224–236.
- [11] van der Linden, W.J., ed. (1998). Optimal test assembly [special issue], *Applied Psychological Measurement* **22**(3), 195–211.
- [12] van der Linden, W.J. & Hambleton, R.K., eds (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

MARK J. GIERL

Test Construction: Automated

BERNARD P. VELDKAMP

Volume 4, pp. 2011–2014

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Test Construction: Automated

In large-scale educational measurement, thousands of candidates have to be tested and the results of these tests might have major impact on candidate's lives. Imagine, for example, the consequences of failing on an admission test for someone's professional career. Because of this, great care is given to the process of testing. Careful psychometric planning of high-stakes tests consists of several steps.

First, decisions are made about the kind of abilities that have to be measured, and about the characteristics of the test. These decisions result in a test blueprint. In this blueprint, the test length is specified and some rules for content balancing and/or other characteristics are defined.

The next step is to write items for the test. Instead of writing and pretesting items for each single test form over and over again, the concept of item banking was introduced. Items are written and pretested on a continuous basis and the item characteristics and statistics are stored in an item bank. In most item banks for large-scale testing, the item statistics are calculated on the basis of **item response theory** (IRT) models. In IRT measurement models, item parameters and person parameters are modeled separately [2]. Apart from sampling variation, the item parameters do not depend on the population or on the other items in the test. Because of this property, items that are calibrated can be used for different group of candidates that belong to the same population. (For a more detailed introduction to IRT, see [2] and [3].)

The final step is to select items from the bank and to compose a test. Optimal test construction deals with the problem of how to select those items that meet the specifications in the best way. All kinds of smart decision rules have been developed to select the items. The main objective for most tests is to maximize measurement precision. When IRT models are applied, measurement precision is determined by the amount of information in the test [3]. Birnbaum [1] presented a rather general approach for test construction. His algorithm consisted of the following steps.

1. Decide on the shape of the desired test information function.

2. Select items from the pool with information functions to fill areas under the target-information function.
3. After each item is added to the test, calculate the test information function.
4. Continue selecting items until the test information function approximates the desired shape.

However, Birnbaum's approach does not take all kinds of realistic test characteristics into account. It just focuses the amount of information, that is, the measurement precision of the test. If more and more test characteristics have to be added to the construction problem, the approach becomes hard to adapt. Optimal test construction is a generalization of Birnbaum's approach that does take these realistic characteristics into account [7]. In the mid-1980s, the first methods for optimal test construction were developed. The observation was made that test construction is just one example of a selection problem. Other well-known selection problems are flight-scheduling, work-scheduling, human resource planning, inventory management, and the traveler-salesman problem. In order to solve optimal-test-construction problems, methods to solve these selection problems had to be translated and applied in the area of test development.

Multiple Objectives

Like most real-world selection problems, optimal test construction is a rather complex problem. For example, a test blueprint might prefer the selection of those items that simultaneously maximize the information in the test, minimize the exposure of the items, optimize item bank usage, balance content, minimize the number of gender-biased items, provide most information for diagnostic purposes, and so on. Besides, in test blueprints, some of these objectives are usually favored above others. To make the problem even more complicated, most mathematical programming algorithms can only handle single-objective selection problems instead of multiple-objective ones.

Three different strategies have been developed to solve multiple-objective selection problems [10]. These strategies are classified as methods based on (a) prior, (b) progressive, and (c) posterior weighting of the importance of different objectives. It should be mentioned that the names of the groups of methods might be a little bit confusing because these names

have different meaning in terminology of **Bayesian Statistics**.

For prior weighting, an inventory of preferences of objectives is made first. On the basis of this order, a sequence of single-objective selection problems is formulated and solved. For progressive methods, a number of solutions are presented to the test assembler. On the basis of his/her preference, a new single-objective selection problem is formulated, and this process is repeated until an acceptable solution is found. For posterior methods, all different kinds of priority orderings of objectives are taken into account. The solutions that belong to all different priority orderings are presented to the test assembler. From all these solutions, the preferred one is chosen. Methods in all groups do have in common that a multiple-objective problem is reformulated into one or a series of single-objective problems. They just differ in the way the priority of different objectives is implemented in the method.

The Methods

For optimal test construction, it is important that the methods are easy to interpret and easy to handle. Two methods from the class of prior weighting methods are generally used to solve optimal-test-construction problems. When the 0–1 Linear Programming (LP) [8] is applied, a target is formulated for the amount of information, the deviation from the target is minimized, and bounds on the other objectives are imposed. For the weighted deviation method [5], targets are formulated for all objectives, and a weighted deviation from these targets is minimized.

The general 0–1 LP model for optimal construction of a single test can be formulated as:

$$\min y \quad (1)$$

subject to:

$$\left| \sum_{i=1}^I I_i(\theta_k) x_i - T(\theta_k) \right| \leq y \quad \forall k, \quad (\text{target information}) \quad (2)$$

$$\sum_{i=1}^I a_c \cdot x_i \leq n_c \quad \forall c, \quad (\text{generic constraint}) \quad (3)$$

$$\sum_{i=1}^I x_i = n, \quad (\text{total test length}) \quad (4)$$

$$x_i \in \{0, 1\}. \quad (\text{decision variables}) \quad (5)$$

In (1) and (2), the deviation between the target-information curve and the information in the test is minimized for several points $\theta_k, k = 1, \dots, K$, on the ability scale. The generic constraint (3) denotes the possibility to include specifications for all kinds of item and test characteristics in the model. Constraints can be included to deal with item content, item type, the word count of the item, the time needed to answer the item, or even constraints to deal with inter-item dependencies. (For an overview of test-construction models, see [7].) The test length is defined in (4), and (5) is a technical constraint that makes sure that items are either in the test ($x_i = 1$) or not ($x_i = 0$).

When the weighted deviation model (WDM) is applied, the test blueprint is used to formulate goals for all item and test characteristics. These goals are seen as desirable properties. During the test-construction process, items that minimize the deviations from these goals are selected. If it is possible, the goals will be met. Otherwise, the deviations of the goal are as small as possible. For some characteristics, it is more important that the goals are met than for others. By assigning different weights to the characteristics, the impacts of the deviations differ. In this way, an attempt is being made to guarantee that the most important goals will be met when the item bank contains enough good-quality items.

The WDM model can be formulated as:

$$\min \sum_j w_j d_j \quad (\text{minimize weighted deviation}) \quad (6)$$

subject to:

$$\left| \sum_{i=1}^I I_i(\theta_k) x_i - T(\theta_k) \right| \leq d_k \quad \forall k, \quad (\text{target information}) \quad (7)$$

$$\sum_{i=1}^I a_c \cdot x_i - n_c \leq d_c \quad \forall c, \quad (\text{generic constraint}) \quad (8)$$

$$x_i \in \{0, 1\} \quad d_j \geq 0. \quad (\text{decision variables}) \quad (9)$$

Where the variables d_j denote the deviations and w_j denotes the weight of deviation j .

Both the 0–1 LP model and the WDM have been successfully applied to solve the optimal-test-construction problems. Which method to prefer

Table 1 Overview of different test-construction models

Parallel test forms	For security reasons, or when a test is administered in several testing windows, parallel tests have to be assembled. Several definitions of parallelness exist, but the concept of weakly parallel tests is most often applied. This means that the same set of constraints is met by the tests and the test information functions are identical. To assemble parallel tests, item variables x_i are replaced by variables $x_{ij} \in \{0, 1\}$ that indicate whether item i is selected for test j . The number of decision variables grows linearly, but the complexity of the problem grows exponentially.
Item sets	Some items in the pool may be grouped around a common stimulus. When a test is administered, all items for this stimulus have to be presented consecutively and sometimes a minimum or maximum number of items from this set need to be selected. To assemble a test with item sets, a model can be extended with decision variables $z_s \in \{0, 1\}$ that denote whether set s is selected for the test.
Classical test construction	Besides IRT, classical test theory (CTT) is still often applied to assemble tests. One of the drawbacks of CTT is that classical item parameters depend on the population and the other items in the test. When the assumption can be made that the population of the examinees hardly changes, optimal test construction might also be possible for classical test forms. A common objective function is to optimize Cronbach's alpha, a lower bound to the reliability of the test.
Test for multiple abilities	For many tests, several abilities are involved in answering the items correctly. In some cases, all abilities are intentional, but in other cases, some of them are considered nuisance. When only one dominant ability is present, the others might be ignored; otherwise, the easiest approach is to optimize Kullback–Leibler information instead of Fisher Information, because its multidimensional form still is a linear function of the items in the test.
Computerized adaptive testing	In Computerized adaptive testing, the items are selected sequentially during test administration. Difficulty of the items is adapted to the estimated ability level of the examinee. For each new item, a test assembly problem has to be solved. Since most test characteristics are defined at test level and item selection happens at item level, somehow these characteristics at test level have to be built in the lower level model of selecting the next item.
Multi-stage testing	In multi-stage testing, a test consists of a network of smaller tests. After finishing the first small test, an examinee is directed to the next test on the basis of his/her ability level. In this way, the difficulty level of the small tests is adapted to the estimated ability level of the examinee. For assembling such a network of small tests, sets of small tests that only differ in difficulty level have to be first assembled. Then these tests have to be assigned to a stage in the network. Optimal multi-stage test assembly is very complicated because all routes through the network have to result in tests that meet the test characteristics.

depends on the way the item and test characteristics are described in the test blueprint. When a very strict formulation of characteristics is used in the test blueprint, the 0–1 LP model seems preferable, because it guarantees that the resulting test meets all constraints. The WDM model is more flexible. It also gives the test assembler more opportunities to prioritize some characteristics above others.

Several Test-Construction Models

In equations 1 to 5 and 6 to 9, general models are given for optimal construction of a single test. Many different models are available for different test-construction problems [7]. An overview of different

models is given in Table 1. In this table, the special features of the models are described.

Current Developments

In the past twenty years, many optimal-test-construction models have been developed and algorithms for test construction have been fine-tuned. Although tests are assembled to be optimal, this does not imply that their quality is perfect. An upper bound to the quality of the test is defined by the quality of items in the pool. The next step in optimization, therefore, is to optimize composition and usage of the item pool.

Optimum item pool design [9] focuses on item pool development in management. An optimal blueprint is developed to guide the item writing process.

This blueprint is not only based on test characteristics, but also takes features of the item selection algorithm into account. The goal of these design methods is to develop an item pool with minimal costs and optimal item usage.

Besides, exposure-control methods have been developed that can be used to optimize the usage of item banks. These methods are applicable for testing programs that use an item pool over a period of time. In optimal test construction, the best items are selected for a test. As a consequence, a small portion of the items in the pool is selected for the majority of the tests. This problem became most obvious when computerized adaptive testing was introduced. To deal with problems of unequal usage of items in the pool, several exposure-control methods are available, both to deal with overexposure of the popular items (e.g., Simpson–Hetter method [6]) or to deal with underexposure (e.g., progressive method [4]).

References

- [1] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 397–479.
- [2] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer Academic Publishers, Boston.
- [3] Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale.
- [4] Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing, *Journal of Educational Measurement* **35**, 311–327.
- [5] Stocking, M.L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing, *Applied Psychological Measurement* **17**, 277–292.
- [6] Simpson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing, *Proceedings of the 27th Annual Meeting of the Military Testing Association*, Navy Personnel Research and Development Center, San Diego, pp. 973–977.
- [7] van der Linden, W.J. (2005). *Optimal Test Assembly*, Springer-Verlag, New York.
- [8] van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test assembly with practical constraints, *Psychometrika* **54**, 237–247.
- [9] van der Linden, W.J., Veldkamp, B.P. & Reese, L.M. (2000). An integer programming approach to item bank design, *Applied Psychological Measurement* **24**, 139–150.
- [10] Veldkamp, B.P. (1999). Multiple objectives test assembly problems, *Journal of Educational Measurement* **36**, 253–266.

BERNARD P. VELDKAMP

Test Dimensionality: Assessment of

MARC E. GESSAROLI AND ANDRE F. DE CHAMPLAIN

Volume 4, pp. 2014–2021

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Test Dimensionality: Assessment of

Test scores are viewed as representations of a theoretical construct, or set of constructs, hypothesized to underlie examinee performances. A test score is typically interpreted as the manifestation of one or more latent traits (*see Latent Variable*) that the test is designed to measure. An examination of the *structural aspect* of construct validity [33] involves an assessment of test dimensionality. That is, the degree to which the dimensional structure of the response matrix is consistent with the domain(s) hypothesized to underlie performance must be examined. Assessing the dimensional structure of a response matrix therefore constitutes a central psychometric activity in that the primary concern of any organization involved in assessment is to ensure that scores and decisions reported to examinees are accurate and valid reflections of the proficiencies that were intended to be targeted by the examination. As well, understanding the dimensional structure of a test is important for other psychometric activities such as equating and calibration.

The term *test dimensionality* is, in some sense, misleading because it does not refer only to the particular set of items comprising a test. Rather, the dimensionality of a test is a function of the interaction between the set of items and the group of examinees responding to those items. An examination of the responses of examinee populations responding to the same set of items might quite possibly result in different conclusions regarding the dimensions underlying the trait. The differences among different examinee populations on variables that might be important in responding to the test items such as curriculum, prior experience, age, and so on, must be carefully considered before any generalizations are made.

What is meant by test dimensionality has been debated in the literature and is still unclear. Early work considered test dimensionality to be related to test homogeneity and reliability. Current definitions relate test dimensionality to some form of the principle of local independence (*LI*).

Definitions of Dimensionality

Dimensionality Based on Local Independence

The principle of local independence is achieved when for fixed levels of a vector of latent traits (θ) the responses for an examinee are statistically independent. More formally, if the item responses for p items are represented by random variables $\mathbf{Y} = Y_1, Y_2, \dots, Y_p$, then the responses to the p are locally independent when, for m latent traits Θ having fixed values θ ,

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_p = y_p | \Theta = \theta) = \prod_{j=1}^p P(Y_j = y_j | \Theta = \theta). \quad (1)$$

The definition of local independence in (1) involves all 2^p higher-order interactions among the items and is known as *strong* local independence (*SLI*). A less stringent version of local independence considers only the second-order interactions among the items. Here, local independence holds if for all item responses $Y_j = y_j$ and $Y_k = y_k$ ($j \neq k; j, k = 1, \dots, p$),

$$P(Y_j = y_j, Y_k = y_k | \Theta = \theta) = P(Y_j = y_j | \Theta = \theta) \times P(Y_k = y_k | \Theta = \theta). \quad (2)$$

In practice, (2) is usually assessed by

$$\sum_{j=1}^p \sum_{\substack{k=1 \\ j \neq k}}^p \text{cov}(Y_j, Y_k | \Theta = \theta) = 0. \quad (3)$$

This definition of LI, known as *weak* local independence (*WLI*), only requires that the covariances between all pairs of items be zero for fixed values of the latent traits.

The existence of *SLI* implies *WLI*. As well, under multivariate normality, *SLI* holds if *WLI* is valid [29]. The very little research comparing analyses on the basis of the two forms of *LI* has found no practical differences between them [24].

McDonald [27] and McDonald and Mok [32] assert that the principle of local independence provides the definition of a latent trait. More formally, they state that $\Theta_1, \Theta_2, \dots, \Theta_m$ are the m latent traits underlying the item responses if and only if, for $\Theta = \theta$, *SLI* holds. Now, in practice, this definition

has been relaxed to require *WLI* instead of *SLI*. Using this definition, the number of dimensions underlying test responses (d_{LI}) is equal to m .

Definitions of latent traits and test dimensionality using either *SLI* or *WLI* are based on precise theoretical requirements of statistical independence. As such, these two principles are reflective of the more general principle of *strict* local independence.

This is a mathematical definition of test dimensionality based on latent traits that, some have argued [11, 12], sometimes fails to capture all the dependencies among the item responses. Furthermore, it is possible to satisfy the mathematical definition yet not fully account for all of the psychological variables affecting the item responses. For example, suppose examinees, in a test of reading comprehension, were asked to read a passage relating to Greek Theatre and then answer a series of questions relating to the passage. It is quite possible that a single trait would account for the mathematical dependencies among the items. However, although local independence is satisfied by a single mathematical latent trait, there might be two psychological traits that influence the responses to the items. An examinee's response to an item might be influenced by some level of proficiency related to general reading comprehension as well as some specific knowledge of Greek Theatre. These two traits are confounded in the single latent trait that results in local independence [10].

Other researchers (e.g., [38, 42, 43]) state that the mathematical definition based on strict local independence is too stringent because it considers both major and minor dimensions. They argue that minor dimensions, although present mathematically, do not have an important influence on the item responses and probably are unimportant in a description of the dimensionality underlying the item responses [38]. This notion is the basis for a definition of *essential dimensionality* described next.

Dimensionality Based on Essential Independence

Using the same notation as above, Stout [43] states that a response vector, $\mathbf{y}$, of p items is said to be *essentially independent* (*EI*) with regard to the latent variables Θ , if, for every pair of responses $y_j, y_k (j, k = 1, \dots, p)$,

$$\frac{2}{p(p-1)} \sum_{1 \leq j < k \leq p} |\text{cov}(y_j, y_k)| \Theta = \theta|$$

$$\rightarrow 0 \text{ as } p \rightarrow \infty. \quad (4)$$

EI is similar to *WLI* in that it only considers pairwise dependencies among the items. However, unlike *WLI*, *EI* does not require that conditional item covariances be equal to zero. Rather, *EI* requires that the mean $|\text{cov}(y_j, y_k)| \Theta = \theta|$ across item pairs is small (approaches 0) as test length p increases. As a result, *EI* only considers dominant dimensions while *WLI* in theory requires all dimensions, however minor, to satisfy (2) or (3). The *essential dimensionality* (d_{EI}) can subsequently be defined as the smallest number of dimensions required for *EI* to hold. From the above definitions, it is clear that $d_{EI} \leq d_{LI}$. In the instance where $d_{EI} = 1$, the item response matrix is said to be *essentially unidimensional*.

Methods to Assess Dimensionality

Owing to the increasing popularity of item response theory (IRT) (*see Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data*), most early work in the assessment of test dimensionality focused specifically on the assessment of unidimensionality. Hattie [15, 16], in a comprehensive review of such techniques, identified indices that purported to assess whether a test was unidimensional. He found that most of the indices used (e.g., those based on reliability, homogeneity, principal components) were *ad hoc* in nature and were not based on any formal definition of dimensionality. There appeared to be confusion regarding the concepts of homogeneity, internal consistency, and dimensionality. For example, although the degree of reliability of a test was thought to be related to its dimensionality (i.e., higher reliability was indicative of a more unidimensional test), Green, Lissitz, and Mulaik [13] showed that it is possible for coefficient alpha to be large with a five-dimensional test.

Today's methods to assess dimensionality are more theoretically sound because they are based on either local independence or essential independence [8]. The methods, in some way, are related to the principles of local and essential independence because they provide indices measuring the degree to which the data are not conditionally independent. Some methods provide global indices while others assess dimensionality by assessing the amount of conditional dependence present between pairs of items.

Methods Based on Local Independence

Factor analysis is the regression of an observed variable on one or more unobserved variables. In dichotomous item scoring, Y is an observed binary variable having a value of 0 for an incorrect response and 1 for a correct response. The unobserved variable(s) are the latent traits needed to correctly answer these items.

Original factor analytic work described a linear relationship between the binary responses and the latent traits, and parameters were estimated by fitting a phi correlation matrix. There are two important weaknesses to this approach. First, predicted values of the dependent variable may be either greater than 1 or less than 0 where, in fact, the item scores are bounded by 0 and 1. This model misspecification leads, in some cases, to additional spurious factors described initially as ‘difficulty factors.’ McDonald and Ahlawat [31] clarified the issue by suggesting that these factors are attributable to the misfit of the model at the upper and lower extremes of the item response function where the relationship between the trait and item responses is nonlinear.

The fitting of a **tetrachoric correlation** matrix was suggested as a possible alternative to the fitting of the phi correlation matrix. While having a sound theoretical basis, there are practical issues associated with the calculation of a tetrachoric correlation matrix that do not allow for the general recommendation of this approach. First, non-Gramian matrices and Heywood cases have been reported. Also, the fitting of tetrachoric matrices has been found to yield poor results in the presence of guessing in the item responses. This is not surprising because the underlying latent distribution of each item assumed in calculating the tetrachoric correlations are the equivalent of a two-parameter normal ogive function, not the three-parameter function that would include a parameter for guessing [25].

Over the past few decades, a number of weighted least-squares estimation methods have been proposed within the context of **structural equation modeling** for use in linear **confirmatory factor analytic** models with dichotomous variables [3, 5, 21, 35]. The fit of a given m -dimensional factor model is typically assessed using a robust chi-square statistic. One limitation of using such methods for assessing dimensionality is that the recommended sample size is typically

prohibitive in practical applications. Recently, diagonally weighted least-squares estimation methods [36] implemented in the software packages Mplus [37] and PRELIS/LISREL [22] have proven more useful with smaller sample sizes (*see Structural Equation Modeling: Software*). However, neither package currently offers an adjustment for guessing or smoothing when analyzing a tetrachoric correlation matrix. More research is needed to assess the utility of these approaches for assessing dimensionality.

McDonald [28] showed that the nonlinear relationship between the probability of correctly answering an item and the underlying trait(s) could be approximated by a third-order polynomial model. He classified this as a model that is linear in its coefficients but nonlinear in the traits [30]. McDonald’s polynomial approximation is implemented in the computer program NOHARM [7]. Parameters are estimated by fitting the joint proportions among the items by unweighted least squares (*see Least Squares Estimation*). McDonald states that the magnitudes and patterns of the residual joint proportions (i.e., the difference between the observed and predicted values of the off-diagonal components of the matrix of joint proportions) provides evidence of the degree to which the fitted model achieves local independence. The sum or mean of the absolute residuals have been found to be generally related to the number of dimensions underlying the item response matrix [14, 16, 24]. As well as providing the residual joint proportions, NOHARM currently provides Tanaka’s goodness of fit index [47] as a measure of model fit. This index has been used in structural equation modeling but little research has been carried out as to its utility in this context.

Another statistic using the residual joint proportions from NOHARM is the Approximate χ^2 statistic [9]. The Approximate χ^2 statistic was proposed as an *ad hoc* method to be used to aid practitioners when other, more theoretically sound procedures, are inappropriate. Although the authors acknowledge limitations of the statistic, the Approximate χ^2 statistic has performed quite well in identifying the correct dimensional structure with simulated data based on compensatory item response models [9, 48]. As with any significance test, the null hypothesis will always be rejected with large enough samples and caution should always be used when interpreting the results.

The common item responses models based on the normal ogive or logistic functions have been shown

to be special cases of a more general nonlinear factor analytic model [28, 46]. McDonald [30] places these IRT models (*see* **Item Response Theory (IRT) Models for Polytomous Response Data; Item Response Theory (IRT) Models for Rating Scale Data**) into a third classification of factor analytic models – models that are nonlinear in both their coefficients and traits.

The full-information factor analysis (FIFA) methods proposed by Bock, Gibbons, and Muraki [1] and implemented in TESTFACT 4.0 [2] yield parameter estimates and fit statistics for a m -dimensional item response model. FIFA is theoretically sound because it uses information from the 2^p response patterns instead of only pairwise relationships to estimate the model parameters and thus uses the strong principle of local independence.

A measure of model misfit in TESTFACT is given by the likelihood ratio χ^2 test. Mislevy [34] indicates that this statistic might poorly approximate the theoretical chi-square distribution due to the incomplete cells in the 2^p response-pattern table. The authors of TESTFACT suggest the use of the difference between two likelihood ratio χ^2 statistics from nested models to test whether a higher dimensional model yields a significantly better fit to the data.

Little research has been carried out to investigate the performance of these statistics. However, Knol and Berger [24] found that the chi-square difference test was unable to correctly identify the number of dimensions in simulated data. De Champlain and Gessaroli [6], in a study investigating the effects of small samples and short test lengths, reported inflated rejections of the assumption of unidimensionality for unidimensionally simulated data. More research under a wider variety of conditions is warranted.

Other methods have been specifically developed to assess the amount of conditional dependence present in pairs of items assuming a single trait underlying the responses. The Q_3 statistic is a measure of conditional correlation between two items [49]. Chen and Thissen [4] assessed the performance of four **measures of association** – (a) Pearson χ^2 , (b) Likelihood Ratio G^2 , (c) Standardized ϕ Coefficient Difference, and (d) Standardized Log-Odds Ratio Difference (τ) – to test for conditional association between binary items in their two-way joint responses. These statistics are computed by comparing the cells in the 2×2 tables of observed and expected joint frequencies where the expected joint frequencies are obtained from an IRT model. Much more detail is available in

Chen and Thissen [4]. Some results of investigations of the performance of these indices can be found in Chen and Thissen [4], Yen [49], and Zwick [51].

The utility of parametric models in assessing test dimensionality is dependent upon the item response model specified and the distributional assumptions of the latent trait(s). Incorrect assumptions and/or model specification might lead to inaccurate conclusions regarding the test dimensionality.

Holland and Rosenbaum [19] discuss an approach to assess the conditional independence among pairs of items without assuming an underlying item response model. Their procedure is based on the work of Holland [18] and Rosenbaum [39]. Conditional association for each pair of items is tested with the **Mantel–Haenszel statistic** [26]. Hattie [16] suggested that this approach was promising because of its link to local independence. However, relatively little research has been carried out studying the effectiveness of the procedure. See Ip [20] for a discussion of some issues.

Ip [20] states that little research has been given to controlling the Type I **error rate** associated with multiple tests inherent in testing the many item pairs for local independence. He proposes a method using a step-down Mantel–Haenszel approach to control for family-wise error rates. Ip suggests that this method might be used to provide more detailed information after a global test has concluded that a response matrix is not unidimensional.

Methods Based on Essential Independence

Over the past 15 to 20 years, Stout and his colleagues have carried out considerable work developing nonparametric methods based on the principle of essential independence described earlier. This work, while still being developed and refined, has resulted in three primary methods that aim to identify the nature and amount of multidimensionality in a test: (a) DIMTEST [42, 45] tests the null hypothesis that a response matrix is essentially unidimensional; (b) DETECT [23, 50] provides an index quantifying the amount of dimensionality present given a particular clustering or partitioning of the items in a test; and, (c) HCA/CCPROX [40], using agglomerative hierarchical cluster analysis, attempts to identify clusters of items that reflect the true (approximate) dimensionality underlying a set of item responses.

DIMTEST is the most popular and widely used procedure. Quite generally, DIMTEST tests the null hypothesis of essential unidimensionality by comparing the magnitudes of the conditional covariances of a subset of items, AT1, chosen to be (a) as unidimensional as possible, and (b) as dimensionally dissimilar to the other items on the test, with the conditional covariances of another subset of items on the test (PT). In the original version of DIMTEST, another subset of the test items, AT2, having the same number of items and similar item difficulties as AT1, is used to adjust the T -statistic for preasymptotic bias. Nandakumar [38] provides a more detailed explanation of DIMTEST.

DIMTEST generally has performed well with simulated data; it has high power in rejecting unidimensionality with multidimensional data and maintains a Type I error rate close to nominal values when the simulated data are unidimensional. However, Type I error rates are inflated in cases where AT1 has high-item discriminations relative to the other items on the test. In this case, AT2 often fails to adequately correct the T -statistic. As well, DIMTEST is not recommended for short test lengths (<20 items) because the necessary partitioning of items into AT1, AT2, and PT results in too few items in each subgroup for a proper analysis. Seraphine [41] concluded that DIMTEST had poorer performance in simulation studies where the data were based on noncompensatory or partially noncompensatory models. A comprehensive investigation of the performance of DIMTEST is provided by Hattie, Krakowski, Rogers, and Swaminathan [17].

Research into DIMTEST (and DETECT and HCA/CCPROX) is continuing. Recent research has suggested the use of nonparametric IRT parametric bootstrap method to correct for the preasymptotic bias in the T -statistic in DIMTEST. A method to use DIMTEST, HCA/CCPROX, and DETECT together to investigate multidimensionality is outlined by Stout, Habing, Douglas, Kim, Roussos, and Zhang [44].

Other Issues

The assessment of test dimensionality is complex. Disagreement exists in several areas such as whether dimensions should be defined and interpreted using psychological or statistical criteria, whether

dimensions should be interpreted as only being examinee characteristics that influence the responses or whether item characteristics (such as item dependencies within a testlet) should be also be considered, and whether only dominant dimensions should be included in the definition and interpretation. Added to this are the different approaches to dimensionality assessment, including methods to investigate whether a test is unidimensional, others that attempt to find the number of dimensions, and yet others whose purpose is to determine the dimensional structure underlying the responses.

Although not discussed here, methods such as those based on factor analysis enable a more detailed analysis of the dimensional structure including the relative strengths of each dimension and the relative strengths of each dimension on individual items.

Essential versus Local Independence in Practice

Essential independence as defined by Stout [42, 43] requires that the conditional covariances among pairs of items are approximately zero. In this way, it uses the *weak* principle of local independence as its basis but, in theory, differs from *WLI* because it does not require that the conditional covariances be equal to zero. In practice, both methods evaluate whether the conditional covariances are sufficiently small to conclude that the proposed dimensional structure explains the item covariances. No distinction is made as to whether the conditional covariances are due to sampling error or the effect of additional or minor nuisance dimensions. In practice, the conclusions reached by the approaches associated with the two principles are usually quite similar. Further discussions are found in Gessaroli [8] and McDonald and Mok [32].

Dimensional Structure

The nature of the dimensionality underlying tests can be complex. *Simple structure* occurs when each trait influences independent clusters of items on the test. That is, each item is related to only one trait. In simulation studies, the amount of multidimensionality is often manipulated by either increasing or decreasing the magnitudes of the intertrait correlations. Higher intertrait correlations result in less multidimensionality, in the sense that the magnitudes of the conditional are smaller when fitting a unidimensional model. In the extreme case, perfectly correlated traits result in

a unidimensional model. Items may also be related to more than one trait. This, combined with correlations among the traits, leads to complex multidimensional structures that are often difficult to identify with the dimensionality assessment methods described above. In general, the dimensionality assessment methods function best when the data have *simple structure*.

Confirmatory Analyses

The dimensionality assessment methods described earlier are exploratory. However, knowledge of the examinees, curriculum, test blueprint, and so on might lead a researcher to ask whether a particular model fits the data. In such cases, the researcher has a priori expectations about the traits underlying the performance as well as the relationship of the items to these traits. For example, an expectation of a model having *simple structure* might be tested. Factor analysis provides a logical basis for performing these confirmatory analyses. As of now, NOHARM provides the only analytical program for confirmatory factor analysis. However, TESTFACT 4.0 estimates parameters for a bifactor model. DIMTEST can be used in a confirmatory mode by allowing the researcher to choose items to be placed in the AT1 subtest. Methods to use DIMTEST, HCA/CCPROX, and DETECT in a confirmatory way are outlined in Stout, et al. [44].

Polytomous Item Responses

The discussion so far has been limited to the dimensionality assessment of binary-coded item responses. The relatively little work carried out for polytomous data has largely been extensions of existing methods for the binary case. Poly-DIMTEST is an extension of DIMTEST and can be used to test the assumption of essential unidimensionality. The structural equation modeling literature provides many extensions of binary to polytomous data [21]. Ip [20] discusses an extension of the assessment of conditional association between pairs of items to polytomous data.

References

- [1] Bock, R.D., Gibbons, R. & Muraki, E. (1988). Full information item factor analysis, *Applied Psychological Measurement* **12**, 261–280.
- [2] Bock, R.D., Gibbons, R., Schilling, S.G., Muraki, E., Wilson, D.T. & Wood, R. (2003). *TESTFACT 4.0 [Computer Software]*, Scientific Software International, Lincolnwood.
- [3] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.
- [4] Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory, *Journal of Educational and Behavioral Statistics* **22**, 265–289.
- [5] Christofferson, A. (1975). Factor analysis of dichotomized variables, *Psychometrika* **40**, 5–32.
- [6] De Champlain, A.F. & Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths, *Applied Measurement in Education* **11**, 231–253.
- [7] Fraser, C. & McDonald, R.P. (2003). NOHARM Version 3.0: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. [Computer software and manual] <http://www.niagarac.on.ca/~cfraser/download/noharmdl.html>.
- [8] Gessaroli, M.E. (1994). The assessment of dimensionality via local and essential independence. A comparison in theory and practice, in *Modern Theories of Measurement: Problems and Issues*, D. Laveault, B.D. Zumbo, M.E. Gessaroli & M.W. Boss, eds, University of Ottawa, Ottawa, pp. 93–104.
- [9] Gessaroli, M.E. & De Champlain, A.F. (1996). Using an approximate χ^2 statistic to test the number of dimensions underlying the responses to a set of items, *Journal of Educational Measurement* **33**, 157–179.
- [10] Gessaroli, M.E. & Folske, J.C. (2002). Generalizing the reliability of tests comprised of testlets, *International Journal of Testing* **2**, 277–295.
- [11] Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models, *British Journal of Mathematical and Statistical Psychology* **33**, 234–246.
- [12] Goldstein, H. & Wood, R. (1989). Five decades of item response modelling, *British Journal of Mathematical and Statistical Psychology* **42**, 139–167.
- [13] Green, S.B., Lissitz, R.W. & Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of test unidimensionality, *Educational and Psychological Measurement* **37**, 827–838.
- [14] Hambleton, R.K. & Rovinelli, R. (1986). Assessing the dimensionality of a set of test items, *Applied Psychological Measurement* **10**, 287–302.
- [15] Hattie, J. (1984). An empirical study of various indices for determining unidimensionality, *Multivariate Behavioral Research* **19**, 49–78.
- [16] Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9**, 139–164.
- [17] Hattie, J., Krakowski, K., Rogers, H.J. & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality, *Applied Psychological Measurement* **20**, 1–14.

- [18] Holland, P.W. (1981). When are item response models consistent with observed data? *Psychometrika* **46**, 79–92.
- [19] Holland, P.W. & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models, *The Annals of Statistics* **14**, 1523–1543.
- [20] Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models, *Psychometrika* **66**, 109–132.
- [21] Jöreskog, K.G. (1990). New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares, *Quality and Quantity* **24**, 387–404.
- [22] Jöreskog, K.G. & Sörbom, D. (2003). *LISREL 8.54 [Computer Software]*, Scientific Software International, Lincolnwood.
- [23] Kim, H.R. (1994). New techniques for the dimensionality assessment of standardized test data, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.
- [24] Knol, D.L. & Berger, M.P.F. (1991). Empirical comparison between factor analysis and multidimensional item response models, *Multivariate Behavioral Research* **26**, 456–477.
- [25] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.
- [26] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease, *Journal of the National Cancer Institute* **22**, 719–748.
- [27] McDonald, R.P. (1962). A note on the derivation of the latent class model, *Psychometrika* **27**, 203–206.
- [28] McDonald, R.P. (1967). *Nonlinear Factor Analysis*, Psychometric Monographs No. 15, The Psychometric Society.
- [29] McDonald, R.P. (1981). The dimensionality of tests and items, *British Journal of Mathematical and Statistical Psychology* **34**, 100–117.
- [30] McDonald, R.P. (1982). Linear versus nonlinear models in item response theory, *Applied Psychological Measurement* **6**, 379–396.
- [31] McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty factors in binary data, *British Journal of Mathematical and Statistical Psychology* **27**, 82–99.
- [32] McDonald, R.P. & Mok, M.M.-C. (1995). Goodness of fit in item response models, *Multivariate Behavioral Research* **30**, 23–40.
- [33] Messick, S. (1995). Standards of validity and the validity of standards in performance assessment, *Educational Measurement: Issues and Practice* **14**, 5–8.
- [34] Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables, *Journal of Educational Statistics* **11**, 3–31.
- [35] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.
- [36] Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage, Newbury Park, pp. 205–243.
- [37] Muthén, B. & Muthén, L. (2004). *Mplus Version 3 [Computer Software]*, Muthén & Muthén, Los Angeles.
- [38] Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality, *Journal of Educational Measurement* **28**, 99–117.
- [39] Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory, *Psychometrika* **49**, 425–435.
- [40] Roussos, L.A., Stout, W.F. & Marden, J.I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality, *Journal of Educational Measurement* **35**, 1–30.
- [41] Seraphine, A.E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ, *Applied Psychological Measurement* **24**, 82–94.
- [42] Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika* **52**, 589–617.
- [43] Stout, W.F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* **55**, 293–325.
- [44] Stout, W.F., Habing, B., Douglas, J., Kim, H.R., Roussos, L. & Zhang, J. (1996). Conditional covariance-based multidimensionality assessment, *Applied Psychological Measurement* **20**, 331–354.
- [45] Stout, W.F., Nandakumar, R., Junker, B., Chang, H. & Steidinger, D. (1992). DIMTEST: a Fortran program for assessing dimensionality of binary items, *Applied Psychological Measurement* **16**, 236.
- [46] Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables, *Psychometrika* **52**, 393–408.
- [47] Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage, Newbury Park, pp. 11–39.
- [48] Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items, *Applied Psychological Measurement* **27**, 159–203.
- [49] Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, *Applied Psychological Measurement* **30**, 187–213.
- [50] Zhang, J. & Stout, W.F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure, *Psychometrika* **64**, 213–249.
- [51] Zwirk, R. (1987). Assessing the dimensionality of NAEP reading data, *Journal of Educational Measurement* **24**, 293–308.

MARC E. GESSAROLI AND ANDRE
F. DE CHAMPLAIN

Test Translation

SCOTT L. HERSHBERGER AND DENNIS G. FISHER

Volume 4, pp. 2021–2026

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Test Translation

‘When we sometimes despair about the use of language as a tool for measuring or at least uncovering awareness, attitude, percepts and belief systems, it is mainly because we do not yet know *why* questions that look so similar actually produce such very different results, or how we can predict contextual effects on a question, or in what ways we can ensure that the respondents will all use the same frame of reference in answering an attitude question’ [19, p. 49].

The Problem

Translating a test from the language in which it was originally written (the ‘source’ language) to a new language (the ‘target’ language) is not a simple process. Problems in test translation can be thought of as problems of test inequality – before we accept the value of a translated test, we should have evidence of its semantic and conceptual equivalence to the original scale. In the following, we define semantic and conceptual equivalence, and suggest strategies for maximizing the equivalence of a translated test and the source test, as well as ways of identifying inequality when it is present.

Semantic Equivalence

Definition

Two tests are semantically equivalent if the identification of words in the target language test has identical or similar meanings to those used in the source language scale. As an example of how semantic equivalence go awry, we cite a Spanish translation of the Readiness to Change Questionnaire (RCQ), which is used to assess stages of change among substance abusers [20]. This Spanish translation of the test has been criticized as having several items that do adequately capture the ideas or meanings expressed by the corresponding English items [7]. In item seven of the RCQ, the English version appears as ‘Anyone can talk about wanting to do something about drinking, but I am actually doing something about it.’ This was translated to ‘Cualquiera puede manifestarar [*sic*] su intención de hacer algo en relación con la bebida,

pero yo ya estoy haciéndolo.’ The word ‘manifestar’ (which was misspelled as ‘manifestarar’) means ‘to manifest’, and is defined as ‘to make evident or certain by showing or displaying’ [17]. However, this definition did not relate to the original English version, which states the idea ‘to talk about wanting’. Consequently, using the action verb ‘manifestar’ in the translated Spanish version did not convey, nor interpret, the original English action ‘to talk about wanting’. Further, in the English version, a cognitive process is being described, whereas in the translated Spanish version, an action is being described. In the following section, we describe five methods for confirming the semantic equivalence of source and target language versions of a scale.

Confirming Semantic Equivalence

Direct Translation. In direct translation, the source language test is translated into the target language, presumably by someone who is fluent in both languages. This is the extent of the translation process. Obviously, as a method for developing semantically equivalent scales, direct translation leaves much to be desired. No external checks on the fidelity of the translation are made; semantic equivalence is taken on faith.

Translation/Back-translation. This is perhaps the most common method for developing semantically equivalent scales. Translation/back-translation is a cyclical process in which the source language test is translated into the target language scale. Then, a second translator attempts to translate the target language test back into the original source language scale. If this translation is not judged sufficiently close to the original source language, another target language translation is attempted that tries to eliminate discrepancies. This process continues until the target language test satisfactorily captures the wording and meaning of the original source language scale.

Ultimate Test. The ultimate test is a two-step process [5]. In the first step, a respondent is asked to perform a behavior using the instructions from the target language version of the scale. Presumably, if the respondent performs the correct behavior, we are sure that at least the instructions are semantically equivalent.

In the second step, bilingual respondents are assigned randomly to four groups. The first group is administered the test in the source language, the second group is administered the test in the target language, the third group is administered the first half of the test in the source language and the second half in the target language, and the fourth group is administered the first half of the test in the target language and the second half in the source language. Using the four-group design, semantic equivalence is indicated if the response distributions of the four groups do not statistically differ, and if the correlation between the two halves of the test in the third and fourth groups is statistically significant.

Parallel Blind Technique. In this technique, two target language versions of the test are independently created, after which the versions are compared, with any differences being reconciled for a third and final version [23].

Random Probe Technique. Here, the target language test is administered to target language speakers. In addition to responding to the items, the respondents provide explanations for their responses. These explanations should uncover any misconceptions about the meaning of the items [8].

Solving Problems of Semantic Equivalence

Although the five procedures discussed above will help identify problems in semantic equivalence, they all, by definition, occur after the initial test translation. Potentially less effort and time will be spent if potential problems in semantic equivalence are addressed during the initial construction of the target language version of the scale. We offer three suggestions for how to maximize the probability of semantic equivalence when the source language version of the test is first constructed.

Write with Translation in Mind. Behling and Law [3] strongly advise researchers to ‘write with translation in mind’. That is, the source language version of the test should be written using words, phrases, sentence structures, and grammatical forms that will facilitate the later translation of the scale. They provide a number of useful suggestions: (a) write short sentences

(b) where possible, write sentences in the active rather than passive voice, (c) repeat nouns rather than substitute with ambiguous pronouns, (d) do not use colloquialisms, metaphors, slang, out-dated expressions or unusual words, (e) avoid conditional verbs such as ‘could’, ‘would’, and ‘should’, (f) avoid subjective qualifiers such as ‘usually’, ‘somewhat’, ‘a bit’, and so on, (g) do not use interrogatively worded sentences, nor double negatives.

Decentering. Decentering follows the same iterative sequence as translation/back-translation with one critical difference: Both the target language version *and* the source language version of the test can be revised during this process [5]. Specifically, following the identification of discrepancies between the back-translated test and the source language scale, decisions are made as to whether the flaws in the target language or source language versions of the test are responsible. Thus, either version can be revised. Once revisions have been made, the translation/back-translation cycle begins anew, continuing until the versions are judged to be sufficiently equivalent semantically.

Multicultural Team Approach. In this approach, a bilingual team constructs the two-language versions of the test in tandem. Some applications of this approach result in two scales that, while (presumably) measuring the same construct, comprise different items in order to capture cultural differences in how the construct is expressed.

Conceptual Equivalence

Definition

Conceptually equivalent scales measure the same construct. Problems associated with conceptually inequivalent scales have to do with the operationalization of the construct in the source language version of the test – whether the test items represent an adequate sampling of behaviors reflecting the construct in both the source *and* target language versions of the scale. Conceptual inequality often occurs from differences in the true dimensionality of a construct between cultures. The dimensionality of a construct refers to how many factors or aspects there are to the construct; for example, the construct of intelligence is sometimes described by two factors or

dimensions, fluid and crystallized intelligence. While the source language test may be sufficient to capture the hypothesized dimensionality of the construct, the target language version may not because the construct has a different dimensionality across cultures. For example, Smith et al. [21] evaluated whether the three factors used to describe responses in the United States to a circadian rhythm test – the construct here being a preference for morning or evening activities – was adequate to describe responses in Japan. The researchers found that one of the three factors was poorly represented in the data from a Japanese language version of the scale. In the following section, we describe four statistical approaches to evaluating the conceptual equality of different language versions of a scale.

Statistically Confirming Conceptual Equivalence

Correlations with Other Scales. Correlations between the target language version of the test and other scales should be similar to correlations between the source language version and the same scales. For example, in English-speaking samples, we would expect a relatively high correlation between the RCQ and the SOCRATES, because both measure a substance abuser's readiness to change. We should also expect a similarly high correlation between the two scales in Spanish-speaking samples. Finding a comparable correlation does not provide strong evidence for conceptual equivalence, but neither does finding an incomparable correlation, because differences in correlation could be due to flaws in translation (a lack of semantic equality).

Exploratory Factor Analysis. Exploratory factor analysis (see **Factor Analysis: Exploratory**) is a technique for identifying **latent variables** (factors) by using a variety of measured variables. The analysis is considered exploratory when the concern is with determining how many factors are necessary to explain the relationships among the measured variables. Similarity of the factor structures found in the source and target language samples is evidence of conceptual equivalence. Although confirmatory factor analysis (see **Factor Analysis: Confirmatory**) (see below) is generally more useful for evaluating conceptual equivalence, exploratory factor analysis can still be of assistance, especially if the source and

target language versions of the scales are being constructed simultaneously, as in the multicultural team approach described earlier.

There are several methods available for comparing the similarity of factors found in the source and target language samples. The *congruence coefficient* [10] measures the correlation between two factors. Another measure, the *salient variable similarity index* [6], evaluates the similarity of factors based on how many of the same items load 'significantly' (saliently) on both. *Configurative matching* [13] determines factor similarity by determining the correlations between factors from two samples that are rotated simultaneously.

Confirmatory Factor Analysis. In contrast to the raw empiricism involved in 'discovering' factors in exploratory factor analysis, confirmatory factor analysis (CFA) is a 'hypotheticist procedure designed to test a hypothesis about the relationship of certain hypothetical common factor variables, whose number and interpretation are given in advance' [18, p. 265]. In general terms, CFA is not concerned with discovering a factor structure, but with confirming the existence of a specific factor structure. Therein lies its relevance to evaluating the conceptual equivalence of source language and target language scales. Assuming we have confirmed our hypothesized factor structure for the source language scale, we want to determine whether the target language test has the same factor structure. In order for the two factor structures to be considered conceptually equivalent, a confirmatory factor model is specified that simultaneously tests equality in five areas: (a) number of factors (i.e., the dimensionality of the factor structure), (b) magnitude of the item loadings on the factors, (c) correlations among the factors, (d) error variances of the items, and (e) factor means.

Most CFA software programs allow one to test whether the same model fits across multiple samples. Space prohibits a detailed description of the procedures involved in CFA. Kline [14] provides an excellent introduction, and Kojima et al. [15] is a clear example of an application of CFA to equivalence in test translation. Typically, we begin by specifying that all parameters of the model (i.e., factor loadings, factor correlations, error variances, and factor means) are identical for the source and target language test samples. This model of *factor invariance*

(identical factor structures) is evaluated by examining one or more indices of the model's fit to the data. One of the most widely used indices for assessing the fit of a model is the χ^2 (chi-square) goodness-of-fit statistic. When evaluating factor invariance and testing factorial invariance, researchers are interested in a nonsignificant χ^2 goodness-of-fit test. A nonsignificant χ^2 indicates that the two scales are factor-invariant, that there is conceptual equivalence between the source and the target language scales. On the other hand, a significant χ^2 indicates that the two scales are *not* factor-invariant, but differ in one or more of the five ways noted above. The idea that the two scales are well described by identical factor structures is rejected by the data.

If the χ^2 goodness-of-fit statistic indicates that factor structures differ between samples, we usually proceed to discover why they differ. The word 'discover' is important because now we are no longer interested in *confirming* the equality of the two scales, but in *exploring* why they are unequal. Although the procedures involved in this exploration differ from those used in traditional exploratory factor analysis, philosophically and scientifically, we have returned to the raw empiricism of exploratory factor analysis. Various models are fit to the two sample data, allowing the two samples to differ in one of the five ways noted above, and equating across the other four. For example, we might specify that the factor correlations differ between the two samples, but specify that the number of factors, the factor loadings, the item error variances, and the factor means be the same. Each of the revised models (M_1) is associated with its own χ^2 goodness-of-fit statistic, which is compared to the χ^2 statistic associated with the factor invariance model (M_0), the model that specified equality between the source and target language scales on all parameters. The χ^2 associated with M_1 must always be no smaller than the χ^2 associated with M_0 . If the difference in the two model chi-squares, which is also chi-square distributed, is significant, M_1 fits the data better than M_0 . Our interpretation at this point would be that the factor correlations differ between the two scales. The process of model modification could take any number of paths at this point; opinion varies as to the 'best' sequence of model modifications [4]. We could choose to specify, one at a time, test inequality on either the number of factors, factor loadings, item error variances, or factor means while specifying equality on the other parameters. Alternatively,

we could specify a model that retains test inequality for the factor correlations, and add, one at a time, test inequality for the number of factors, item error variances, or factor means. Let us call either of these models M_2 . We would then evaluate the significance of $\chi^2_{M_2} - \chi^2_{M_1}$, and use that result to determine subsequent model modifications, until, finally, we find a model that fits the data well and whose fit cannot be improved by any other subsequent model modifications. The researcher must always bear in mind, however, that the validity of the final model is quite tentative, and should be cross-validated on new samples. In any event, unless the model of factor invariance, M_0 , fits the data well, the source and target language scales are not conceptually equivalent.

Item Response Theory. Item response theory (IRT) is another powerful method for testing conceptual equivalence (*see Item Response Theory (IRT) Models for Dichotomous Data*). Again, space limitations prohibit us from discussing IRT in detail; [9] provides an excellent introduction and [2] is a clear example of an application of IRT to equivalence in test translation.

In IRT, the probability of a correct response to an item is modeled using one or more parameters. A typical set of item parameters is item difficulty and item discrimination. For two tests to be conceptually equivalent, corresponding items on the tests must have identical item difficulties and item discriminations. In IRT, an item is said to have **differential item functioning** (DIF) when the item difficulty and item discrimination of the same item on two versions of the test differ significantly. DIF can be tested statistically using a number of methods. The *parameter equating method* [16] tests differences between two tests' item parameters using a chi-square statistics. The **Mantel-Haenszel method** [11] (*see Item Bias Detection: Classical Approaches*) also uses a chi-square statistic to test whether the observed frequencies of responses to one test version differs significantly from the frequencies expected if there were no differences in the item parameters. When all items within a scale are assumed to have the same discrimination, a *t* Test, the *item difficulty shift statistic* [12], is used to test for differences in an item's difficulty between two tests. Yet another measure of item DIF is the *mean square residual* [24], which involves analyzing the fit of each item in the source language test group and the target language test group. An item

should either fit or fail to fit in a similar manner for both samples; thus it relates to the concept being measured in the same way for each group. Thissen et al. [22] describe a model comparison method for testing DIF that is analogous to the model comparison method used to test for factor invariance in CFA. Item parameter estimates are obtained from a model that equates the parameters between the two groups, and item parameter estimates are obtained from a second model that allows one or more of the item parameters to differ between the two groups. If the difference between the two models' chi-squares is significant, the item parameters differ between the groups.

Some methods for detecting DIF do not require IRT estimation of item parameters. For example, the *delta-plot method* [1] involves plotting pairs of item difficulties (deltas) for the same item from the two groups. If the two tests are conceptually equivalent, we would expect that the item difficulties would order themselves in the same way in the two groups, thus forming a 45° line from the origin when plotted.

Conclusion

Although the methods involved in confirming semantic and conceptual equivalence have been described separately, in practice, it is frequently difficult to determine whether apparent flaws in test translation are attributable to the one or the other or both. For example, in [21] the researchers attributed factor structure differences not to true cultural differences in the nature of the construct (conceptual inequivalence), but to flaws in the Japanese translation (semantic inequivalence). Although the researchers were certain that semantic inequivalence was the culprit, the evidence they present, at least to our understanding, does not argue strongly for the lack of either type of equivalence. No doubt, in this study, and in many other studies, test translation problems will arise from a bit of each.

References

- [1] Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., The John Hopkins University Press, Baltimore, pp. 96–116.
- [2] Beck, C.T. & Gable, R.K. (2003). Postpartum depression screening scale: Spanish version, *Nursing Research* **52**, 296–306.
- [3] Behling, O. & Law, K.S. (2000). *Translating Questionnaires and Other Research Instruments: Problems and Solutions*, Sage Publications, Thousand Oaks.
- [4] Bollen, K.A. (2000). Modeling strategies: in search of the holy grail, *Structural Equation Modeling* **7**, 74–81.
- [5] Brislin, R.W. (1973). Questionnaire wording and translation, in *Cross-Cultural Research Methods*, R.W. Brislin, W.J. Lonner & R.M. Thorndike, eds, John Wiley & Sons, New York, pp. 32–58.
- [6] Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*, Plenum Press, New York.
- [7] Fisher, D.G., Rainof, A., Rodríguez, J., Archuleta, E. & Muñiz, J.F. (2002). Translation problems of the Spanish version of the readiness to change questionnaire, *Alcohol & Alcoholism* **37**, 100–101.
- [8] Guthery, D. & Lowe, B.A. (1992). Translation problems in international marketing research, *Journal of Language for International Business* **4**, 1–14.
- [9] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.
- [10] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Edition, University of Chicago Press, Chicago.
- [11] Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure, in *Testing Validity*, H. Wainer & H.I. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 129–146.
- [12] Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., The John Hopkins University Press, Baltimore, pp. 117–160.
- [13] Kaiser, H., Hunka, S. & Bianchini, J. (1971). Relating factors between studies based upon different individuals, *Multivariate Behavior Research* **6**, 409–422.
- [14] Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, The Guilford Press, New York.
- [15] Kojima, M., Furukawa, T.A., Takahashi, H., Kawaii, M., Nagaya, T. & Tokudome, S. (2002). Cross-cultural validation of the beck depression inventory-II in Japan, *Psychiatry Research* **110**, 291–299.
- [16] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.
- [17] Mish, F.C., ed. (2000). *Merriam-Webster's Collegiate Dictionary*, Springfield Merriam-Webster.
- [18] Mulaik, S.A. (1988). Confirmatory factor analysis, in *Handbook of Multivariate Experimental Psychology*, 2nd Edition, J.R. Nesselroade & R.B. Cattell, eds, Plenum Press, New York, pp. 259–288.
- [19] Oppenheim, A.N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*, Continuum International, London.
- [20] Rodríguez-Martos, A., Rubio, G., Auba, J., Santo-Domingo, J., Torralba, L.I. & Campillo, M. (2000). Readiness to change questionnaire: reliability study

- of its Spanish version, *Alcohol & Alcoholism* **35**, 270–275.
- [21] Smith, C.S., Tisak, J., Bauman, T. & Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: implications for between- and within-population assessments, *Journal of Applied Psychology* **76**, 628–636.
- [22] Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models, in *Differential Item Functioning: Theory and Practice*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 147–169.
- [23] Werner, O. & Campbell, T.D. (1970). Translating, working through interpreters, and problems of decentering, in *A Handbook of Method in Cultural Anthropology*, R. Noll & R. Cohen, eds, Columbia University Press, New York, pp. 398–420.
- [24] Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis*, MESA Press, Chicago.

SCOTT L. HERSHBERGER AND DENNIS
G. FISHER

Tetrachoric Correlation

SCOTT L. HERSHBERGER

Volume 4, pp. 2027–2028

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tetrachoric Correlation

The *tetrachoric correlation* is used to correlate two artificially dichotomized variables, X and Y , which have a *bivariate-normal distribution* (see **Catalogue of Probability Density Functions**) [8]. In a bivariate-normal distribution, the distribution of Y is normal at fixed values of X , and the distribution of X is normal at fixed values of Y . Two variables that are bivariate-normally distributed must each be normally distributed. The calculation of the tetrachoric correlation involves corrections that approximate what the **Pearson product-moment correlation** would have been if the data had been continuous. If instead of the tetrachoric correlation, the Pearson product-moment correlation formula is directly applied to the data (we are not willing to assume that X and Y are truly bivariate-normally distributed), the resulting correlation is referred to as a *phi correlation*.

Tetrachoric correlations are often useful in behavioral genetic research (see **Correlation Issues in Genetics Research**). For example, we might have twin concordance data for a psychiatric illness such as schizophrenia; either both twins are diagnosed as schizophrenic (concordant twins) or only one of the co-twins is diagnosed as schizophrenic (discordant twins) [4]. Instead of assuming that there is true phenotypic discontinuity in schizophrenia due to a single gene of large effect, we assume that the discontinuity is an arbitrary result of classifying people by kind rather than by degree. In the latter case, the phenotype is truly continuous, the result of many independent genetic factors, leading to a continuous distribution of genotypes for schizophrenia. Models that assume that a continuous distribution of genotypes underlie an artificially dichotomized phenotype are referred to as *continuous liability models* [7].

The genetic analysis of continuous liability models (see **Liability Threshold Models**) assumes that the liability for the phenotype (e.g., schizophrenia) in pairs of twins is bivariate-normal with zero mean vector and a correlation ρ between the liabilities of twin pairs. This is the tetrachoric correlation. If both twins are above a genetic *threshold* t , then both twins will be diagnosed as schizophrenic. If both are below the genetic threshold, then neither twin will be diagnosed as schizophrenic. If one twin is above the threshold and the other below, then the first will be diagnosed as schizophrenic and the other will not.

The probability that both twins will be diagnosed as schizophrenic is thus

$$p_{11} = \int_t^\infty \int_t^\infty \theta(x, y, \rho) dy dx, \quad (1)$$

where $\theta(x, y, \rho)$ is the bivariate-normal probability function. Similarly, the probability that the first twin will be diagnosed as schizophrenic and the second will not is

$$p_{10} = \int_t^\infty \int_{-\infty}^t \theta(x, y, \rho) dy dx. \quad (2)$$

Similar expressions follow for the other categories of twin diagnosis, p_{01} and p_{00} .

In order to obtain values for t , the genetic threshold, and ρ , the tetrachoric correlation between twins, the bivariate-normal probability integrals are evaluated numerically for selected values of t and ρ values that maximize the likelihood of the data [3]. The observed data are the number of twin pairs in each of the four cells of the twin **contingency table** for schizophrenia. Thus, the log-likelihood to be maximized for a given contingency table is

$$L = C + \sum_i \sum_j N_{ij} \ln p_{ij}, \quad (3)$$

where C is some constant. Estimates of t and ρ that produce the largest value of L for a contingency table are **maximum likelihood estimates**.

As an example of computing the tetrachoric correlation for a dichotomous phenotype, consider the monozygotic twin concordance data ($N = 56$ pairs) for a male homosexual orientation from [2]:

		Twin1	
		Yes	No
Twin 2	Yes	29	14
	No	13	0

Assuming a model in which a homosexual orientation is due to additive genetic and random environmental effects, the tetrachoric correlation between monozygotic twins is 0.50. The phi correlation between the twins' observed dichotomized phenotypes is 0.32.

2 Tetrachoric Correlation

The tetrachoric correlation is also often used in the factor analysis of dichotomous item data for the following reason. In the context of cognitive testing, where an item is scored ‘right’ or ‘wrong’, a measure of the difficulty of an item is the proportion of the sample that passes the item [1]. Factor analyzing binary item data on the basis of phi correlations may lead to the extraction of spurious factors known as *difficulty factors*, because there may be a tendency to yield factors defined solely by items of similar difficulty. Why difficulty factors are produced by factoring phi correlations has been attributed to the restricted range of phi correlations; $-.8 \leq 0 \leq .8$ [5], and to the erroneous application of the linear common factor model to the inherently nonlinear relationship that may exist between item responses and latent factors [6]. In either case, use of the tetrachoric correlation instead of the phi correlation may prevent the occurrence of spurious difficulty factors.

References

- [1] Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*, Wadsworth, Monterey.
- [2] Bailey, J.M. & Pillard, R.C. (1991). A genetic study of male sexual orientation, *Archives of General Psychiatry* **48**, 1089–1096.
- [3] Eaves, L.J., Last, K.A., Young, P.A. & Martin, P.A. (1978). Model-fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.
- [4] Gottesman, I.I. & Shields, J. (1982). *Schizophrenia: The Epigenetic Perspective*, Cambridge University Press, New York.
- [5] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Edition, University of Chicago Press, Chicago.
- [6] McDonald, R.P. & Ahlwat, K.S. (1974). Difficulty factors in binary data, *British Journal of Mathematical and Statistical Psychology* **27**, 82–99.
- [7] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Press, Boston.
- [8] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, McGraw-Hill, New York.

SCOTT L. HERSHBERGER

Theil Slope Estimate

CLIFFORD E. LUNNEBORG

Volume 4, pp. 2028–2030

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Theil Slope Estimate

The Linear Regression Model

In simple linear regression, the expected or mean value of a response variable, Y , is modeled as a linear function of the value of an explanatory variable, X :

$$E(Y|X = x_i) = \alpha + \beta x_i, \quad (1)$$

(see **Multiple Linear Regression**); that is, for each value of X of interest to the researcher, the values of Y are distributed about this mean value. A randomly chosen observation, y_i , from the distribution of responses for $X = x_i$ is modeled as

$$y_i = \alpha + \beta x_i + e_i, \quad (2)$$

where e_i is a random draw from a distribution of deviations. The task in regression is to estimate the unknown regression constants, α and β , and, often, to test hypotheses about their values.

Estimation and hypothesis testing depend upon a sample of n paired observations, (x_i, y_i) , $i = 1, 2, \dots, n$. The n values of the explanatory variable are treated as fixed constants, values specified by the researcher.

The Least Squares Estimates of the Regression Parameters

The most common estimates of the regression parameters are those based on minimizing the average squared discrepancy between the sampled values of Y and the estimated means of the distributions from which they were sampled (see **Least Squares Estimation**); that is, we choose as estimates of the linear model intercept and slope those values, $\hat{\alpha}$ and $\hat{\beta}$, that minimize the mean squared deviation:

$$MSD = \frac{1}{n} \sum_{i=1}^n [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2. \quad (3)$$

The resulting estimates can be expressed as functions of the sample means, standard deviations, and correlation: $\hat{\beta} = r_{xy}[SD(y)/SD(x)]$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

The Normal Linear Regression Model

The normal regression model assumes that response observations are sampled independently and that the deviations, the e_i s, are distributed as the normal random variable with a mean of zero and a variance, σ^2 , that does not depend upon the value of x_i . Under this model, the least squares slope and intercept estimates are unbiased and with sampling distributions that are normal with variances that are a function of σ^2 and the mean and variance of the x_i s. As σ^2 is estimated by $(n/n - 2)MSD$, this leads to hypothesis tests and confidence-bound estimates based on one of Student's t distributions.

Nonparametric Estimation and Hypothesis Testing

The least squares slope estimate can be influenced strongly by one or a few outlying observations thus providing a misleading summary of the general strength of relation of the response observations to the level of the explanatory variable. A more robust estimator has been proposed by Theil [4].

Order the paired observations, (x_i, y_i) , in terms of the sizes of the x_i s, letting $(x_{[1]}, y_1)$ designate the pair with the smallest value of X and $(x_{[n]}, y_n)$ the pair with the largest value of X .

For every pair of explanatory variable scores for which $x_{[j]} < x_{[k]}$, we can compute the two-point slope

$$S_{jk} = \frac{(y_k - y_j)}{(x_{[k]} - x_{[j]})}. \quad (4)$$

If there are no ties in the values of X , there will be $(n - 1)n/2$ such slopes; if there are ties, the number will be smaller.

The Theil slope estimator is the median of these two-point slopes,

$$\hat{\beta}_T = \text{Mdn}(S_{jk}). \quad (5)$$

Outlying observations will be accompanied by two-point slopes that are either unusually small or unusually large. These are discounted in the Theil estimator.

Conover [1] describes how the ordered set of two-point slopes, the S_{jk} s, can be used as well to find a **confidence interval** for β .

An intercept estimator related to the Theil slope estimate has been proposed by Hettmansperger,

2 Theil Slope Estimate

McKean & Sheather [2]. If we use the Theil slope estimate to compute a set of differences of the form, $a_i = y_i - \hat{\beta}_T x_i$, then the regression intercept can be estimated robustly by the median of these differences,

$$\hat{\alpha}_H = \text{Mdn}(a_i). \quad (6)$$

To carry out Theil's nonparametric test of the hypothesis that $\beta = \beta_0$, we first compute the n differences, $D_i = y_i - \beta_0 x_i$. If β has been correctly described by the null hypothesis, any linear dependence of the y_i s on the x_i s will have been accounted for. As a result, the D_i s will be uncorrelated with the x_i s. Hollander and Wolf [3] propose carrying out Theil's test by computing **Kendall's rank correlation** between the D_i s and the x_i s, $\tau(D, x)$, and testing whether τ differs significantly from zero. Conover [1] suggests using the **Spearman rank correlation**, $\rho(D, x)$, for the same test. Hollander and Wolf [3] provide, as well, an alternative derivation of the Theil test.

Example

As an example of the influence of an outlier on the estimation of regression parameters, [5] gives the artificial example in Table 1.

The y value of 12.74 clearly is out of line with respect to the other observations. The least squares estimates of the regression parameters are $\hat{\alpha} = 3.002$ and $\hat{\beta} = 0.500$. As there are no ties among the x values, there are 55 two-point slopes. Their median provides the Theil estimate of the slope parameter, $\hat{\beta}_T = 0.346$. The corresponding Hettmansperger intercept estimate is $\hat{\alpha}_H = 4.004$. Both differ considerably from the least squares estimates.

Table 1 An Anscombe influence example

x	y
4	5.39
5	5.73
6	6.08
7	6.42
8	6.77
9	7.11
10	7.46
11	7.81
12	8.15
13	12.74
14	8.84

Following the procedure outlined in [1], the 95% confidence interval for β is given by the 17th and 39th of the ordered two-point slopes: [0.345, 0.347]. When one ignores the aberrant two-point slopes associated with the point (13, 12.74), the other two-point slopes are in remarkable agreement for this example.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [2] Hettmansperger, T.P., McKean, J.W. & Sheather, S.J. (1997). Rank-based analyses of linear models, in *Handbook of Statistics*, Vol. 15, Elsevier-Science, Amsterdam.
- [3] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.
- [4] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, *Indagationes Mathematicae* **12**, 85–91.
- [5] Weisberg, S. (1985). *Applied Linear Regression*, 2nd Edition, Wiley, New York.

CLIFFORD E. LUNNEBORG

Thomson, Godfrey Hilton

PAT LOVIE

Volume 4, pp. 2030–2031

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Thomson, Godfrey Hilton

Born: March 27, 1881, in Carlisle, England.

Died: February 9, 1955, in Edinburgh, Scotland.

Godfrey Thomson's life did not start propitiously. His parents separated when he was an infant and his mother took him to live with her own mother and sisters in her native village of Felling, on industrial south Tyneside. There he attended local schools until the age of 13, narrowly avoiding being placed into work as an engineering patternmaker by winning a scholarship to Rutherford College in Newcastle-upon-Tyne. Three years later, he was taken on as a pupil-teacher (essentially, an apprenticeship) in his old elementary school in Felling, but had to attend Rutherford College's science classes in the evenings and weekends. His performance in the London University Intermediate B.Sc. examinations earned him a Queen's Scholarship in 1900, which allowed him to embark on a full-time science course and also train as a teacher at the Durham College of Science (later renamed Armstrong College) in Newcastle, at that time part of the University of Durham. He graduated in 1903, with a distinction in Mathematics and Physics, the year after obtaining his teaching certificate. With the aid of a Pemberton Fellowship from Durham, he set off to study at the University of Strasburg in 1903, gaining a Ph.D. for research on Hertzian waves three years later.

The Queen's Scholarship, however, had strings attached that required the beneficiary to teach for a certain period in 'in an elementary school, the army, the navy, or the workhouse'. Apparently, an assistant lectureship at Armstrong College qualified under these headings! As teaching educational psychology was one of Thomson's duties, he felt obliged to learn something of psychology generally, and was mildly surprised to find that he enjoyed the experience. However, it was during a summer vacation visit to C. S. Myers's laboratory in Cambridge in 1911 that his interest was caught by **William Brown's** book *The Essentials of Mental Measurement* [1]. Although Thomson's initial foray was on the psychophysical side (and led to publications that earned him a D.Sc. in 1913), he was also intrigued by Brown's criticisms of **Charles Spearman's** two-factor theory of human ability. According to Thomson [7], sitting at his fire-side and armed with only a dice, a house slipper,

and a notepad, he was able to generate sets of artificial scores with a correlational structure consistent with Spearman's theory, but without needing its cornerstone, the single underlying general factor g . The publication of this finding in 1916 [3] marked both the start of Thomson's many significant contributions to the debates on intelligence and the newly emerging method of **factor analysis**, and also of a long running, and often bitter, quarrel with Spearman.

Thomson's own notion of intelligence evolved into his 'sampling hypothesis' in which the mind was assumed to consist of numerous connections or 'bonds' and that, inevitably, tests of different mental abilities would call upon overlapping samples of these bonds. In Thomson's view, therefore, the correlational structure that resulted suggested a statistical rather than a mental phenomenon. *The Factorial Analysis of Human Ability* [4], published in five editions between 1939 and 1951, was Thomson's major work on factor analysis. He also coauthored with Brown several further editions in 1921, 1925, and 1940, of the book, *The Essentials of Mental Measurement*, that had so fired him in 1911.

However, it is for his work on devising mental tests that Thomson is best remembered. This began in 1920 for the newly promoted Professor Thomson with a commission from Northumberland County Council for tests that could be used in less-privileged schools to select pupils who merited the opportunity of a secondary education – as Thomson himself had benefited many years earlier. By 1925, after a year spent in the United States with E. L. Thorndike, he had accepted a chair of education at Edinburgh University, with the associated post of Director of Moray House teacher training college, and had begun to formulate what became known as the Moray House tests; these tests would be widely used in schools throughout the United Kingdom and many other countries. Thomson and his many collaborators were also involved in a large-scale study of how the test results from schoolchildren related to various social factors, including family size and father's occupation, and to geographical region.

Thomson received many honors from learned societies and academies abroad. He was knighted in 1949. Even after retiring in 1951, he was still writing and working diligently on data from a longitudinal Scottish study. Thomson's final book, *The Geometry of Mental Measurement* [6], was published in 1954.

2 Thomson, Godfrey Hilton

‘God Thom’, as he was nicknamed (though not wholly affectionately), by his students in Edinburgh, died from cancer in 1955 at the age of 73.

Further material on Thomson’s life and work can be found in [2, 5, 7].

References

- [1] Brown, W. (1911). *The Essentials of Mental Measurement*, Cambridge University Press, London.
- [2] Sharp, S. (1997). “Much more at home with 3.999 pupils than with four”: the contributions to psychometrics of Sir Godfrey Thomson, *British Journal of Mathematical and Statistical Psychology* **50**, 163–174.
- [3] Thomson, G.H. (1916). A hierarchy without a general factor, *British Journal of Psychology* **8**, 271–281.
- [4] Thomson, G. (1939). *The Factorial Analysis of Human Ability*, London University Press, London.
- [5] Thomson, G. (1952). Godfrey Thomson, in *A History of Psychology in Autobiography*, Vol. 4, E.G. Bor-ing, H.S. Langfeld, H. Werner & R.M. Yerkes, eds, Clark University Press, Worcester, pp. 279–294.
- [6] Thomson, G.H. (1954). *The Geometry of Mental Measurement*, London University Press, London.
- [7] Thomson, G. (1969). *The Education of an Englishman*, Moray House College of Education, Edinburgh.

PAT LOVIE

Three Dimensional (3D) Scatterplots

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

Volume 4, pp. 2031–2032

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Three Dimensional (3D) Scatterplots

A standard **scatterplot** is appropriate to display the relationship between two continuous variables. But what happens if there are three variables? If the third variable is categorical, it is customary to print different symbols for the different points. If the variable is metric, then many packages allow the user to set the size of the data points to represent the value on the third variable and the result is the **bubble plot**. This is what is normally recommended when graphing three continuous variables and the sample size is small. Another possibility is to make a series of two-variable scatterplots for each bivariate comparison, sometimes called a **scatterplot matrix**. But if it is the three-way relationships that is of interest, these bivariate scatterplots are not appropriate.

An alternative procedure available in many graphics packages is to plot the data points for three variables in a three dimensional space. Like the standard scatterplot, the data points are placed at their appropriate location within a coordinate space, but, this time, the space is three dimensional. Because paper and computer screens are two dimensional, it is important to use some of the available features, such as rotation of axes, so that the all the dimensions are clear.

Figure 1(a) shows data on the baseline scores for working memory span using three tests: digit span, visual spatial span, and Corsi block span [1]. These were expected to be moderately correlated, and they are. While this plot can help in understanding the patterns in the data, it can still be difficult to make sense of the data. The Corsi task is more complex than the other two tasks, and the researchers were interested in how well the digit and visual spatial tasks could predict the Corsi scores. The resulting regression plane has been added to Figure 1(b). This helps to show the general pattern of the cloud of data points. Other planes could be used instead (for example, from more robust methods, polynomials, etc.).

There is a sense in which the three-dimensional scatterplots attempt to make the two-dimensional page into a three-dimensional object, and this can never be wholly satisfactory. Using size, contours, and colors to show values on other dimensions within a two-dimensional space are often easier for the reader to interpret.

Reference

- [1] Wright, D.B. & Osborne, J.E. (in press). Dissociation, cognitive failures, and working memory, *American Journal of Psychology*.

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

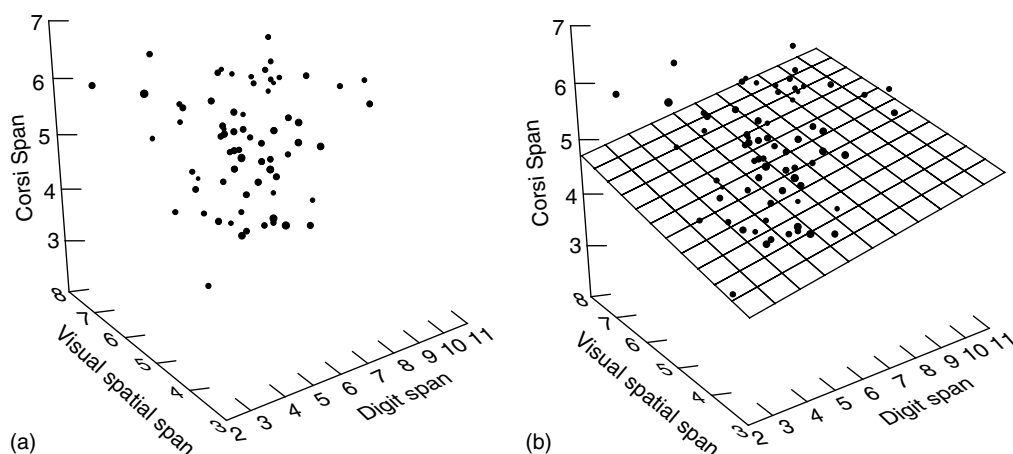


Figure 1 Showing three dimensions in a scatterplot. Figure 1a shows only the data points. Figure 1b includes the linear regression plane

Three-mode Component and Scaling Methods

PIETER M. KROONENBERG

Volume 4, pp. 2032–2044

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Three-mode Component and Scaling Methods

Introduction

What is Three-mode Analysis?

Most statistical methods are used to analyze the scores of objects (subjects, groups, etc.) on a number of variables, and the data can be arranged in a two-way *matrix*, that is, a rectangular arrangement of rows (objects) and columns (variables) (*see Multivariate Analysis: Overview*). However, data are often far more complex than this, and one such complexity is that data have been collected under several conditions or at several time points; such data are referred to as *three-way data* (see Figure 1). Thus, for each condition there is a matrix, and the set of matrices for all conditions can be arranged next to each other to form a broad matrix of subjects by variables times conditions. Alternatively, one may create a tall matrix of subjects times conditions by variables. The third possibility is to arrange the set of matrices in a three-dimensional block or three-way *array*, so that metaphorically the data now fit into a box. The collection of techniques that attempt to analyze such data boxes are referred to as *three-mode methods*, and making sense of such data is the art of *three-mode analysis*. Thus three-mode analysis is the analysis of data that fit into boxes.

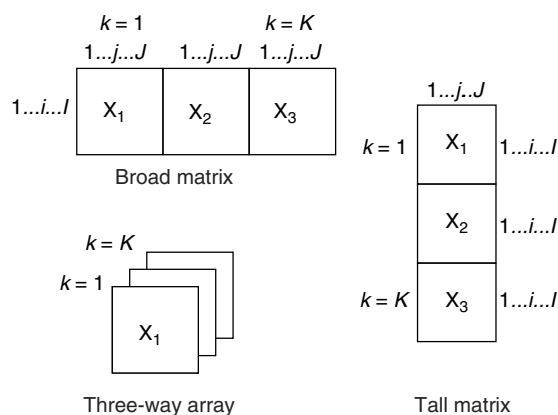


Figure 1 Two-way matrices and a three-way array

Usually a distinction is made between *three-way* and *three-mode*. The word *way* is more general and points to the three-dimensional arrangement irrespective of the content of the data, while the word *mode* is more specific and refers to the content of each of the ways. Thus objects, variables, and conditions can be the three modes of a data array. When the same entities occur twice, as is the case in a correlation matrix, and we have correlation matrices for the same variables measured in several samples, one often speaks of a two-mode three-way data array, where the variables and the samples are the two modes. However, to avoid confusion and wordiness, we generally refer to three-way data and three-way arrays, and to three-mode methods and three-mode analysis, with the exception of the well-established name of three-way analysis of variance. The word *two-mode analysis* is then reserved for the analysis of two-way matrices.

Why Three-mode Analysis?

Given that there are so many statistical and data-analytic techniques for two-way data, why are these not sufficient for three-way data? The simplest answer is that two-mode methods do not respect the three-way design of the data. Such disrespect is not unusual as, for instance, time series data are often analyzed as if the time mode was an unordered mode and the time sequence is only used in interpretation.

Three-way data are supposedly collected because all three modes were necessary to answer the pertinent research questions. Such research questions can be facetiously summarized as: ‘Who does what to whom and when?’, or more specifically: ‘Which groups of subjects behave differently on which variables under which conditions?’ or in an agricultural setting ‘Which plant varieties behave in a specific way in which locations on which attributes?’. Such questions cannot be answered with two-mode methods, because there are no separate parameters for all three modes. When analyzing three-way data with two-mode methods, one has to rearrange the data as in Figure 1, and this means that either the subjects and conditions are combined to a single mode (‘tall matrix’) or the variables and conditions are so combined (‘broad matrix’). Thus, two of the modes are always confounded and no independent parameters for these modes are present in the model itself.

2 Three-mode Component and Scaling Methods

In general, a three-mode model is much more parsimonious for three-way data than an appropriate two-mode model. To what extent this is true depends very much on the specific model used. In some three-mode component models, low-dimensional representations are defined for all three modes, which can lead to enormous reductions in parameters. Unfortunately, it cannot be said that this means that automatically the results of a three-mode analysis are always easier to interpret. Again this depends on the questions asked and the data and models used.

An important aspect of three-mode models, especially in the social and behavioral sciences, is that they allow the analysis of individual differences. The subjects from whom the data have been collected do not disappear in sufficient statistics for distributions, such as means, (co)variances or correlations, and possibly higher-order moments such as the **kurtosis** and **skewness**, but they are examined in their own right. This implies that often the data at hand are taken as is, and not necessarily as a random sample from a larger population in which the subjects are in principle exchangeable. Naturally, this affects the generalizability but that is considered inevitable. At the same time, however, the subjects are recognized as the ‘data generators’ and are awarded a special status, for instance, when statistical stability is determined via **bootstrap** or **jackknife** procedures. Furthermore, it is nearly always the contention of the researcher that similar samples are or may become available, so that at least part of the results are valid outside the context of the specific sample.

Three-way Data

As mentioned above, three-way data fit into three-dimensional boxes, which in the social and behavioral sciences often take the form of subjects by variables by conditions.

The first way (subjects) has index i running along the vertical axis, the second way or variables index j runs along the horizontal axis, and the third way or conditions index k runs along the ‘depth’ axis of the box. The number of levels in each way is I , J , and K . The $I \times J \times K$ three-way data matrix $\underline{\mathbf{X}}$ is thus defined as the collection of elements, x_{ijk} with the indices $i = 1, \dots, I$; $j = 1, \dots, J$; $k = 1, \dots, K$.

A three-way array can also be seen as a collection ‘normal’ (=two-way) matrices or *slices*. There are three different arrangements for this, as is shown in

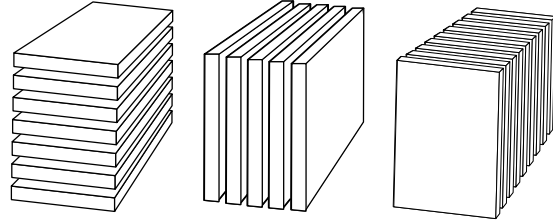


Figure 2 Slices of a three-mode data array; horizontal, lateral and frontal, respectively

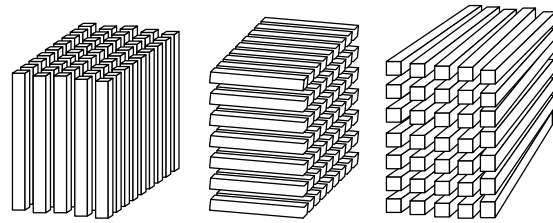


Figure 3 Fibers of a three-mode data array; columns, rows, and tubes, respectively

Figure 2. Furthermore, one can break up a three-way matrix into one-way submatrices (or vectors), called *fibers* (see Figure 3). The slices are referred to as *frontal slices*, *horizontal slices*, and *lateral slices*. The fibers are referred to as *rows*, *columns*, and *tubes*. The prime reference paper for three-mode terminology is [35].

A Brief Example: Abstract Paintings

Consider the situation in which a number of persons (Mode 1) have rated twenty abstract paintings (Mode 2) using some 10 different rating scales which measure the feelings these paintings elicit (Mode 3). Suppose a researcher wants to know if there is a common structure underlying the usage of the rating scales with respect to the paintings, how the various subjects perceive this common structure, and/or whether subjects can be seen as types or combination of types in their use of the rating scales to describe their emotions. Although all subjects might agree on the kinds (or dimensions) of feelings elicited by the paintings, for some subjects some of these dimensions might be more important and/or more correlated than for other subjects, and one could imagine that different types of subjects evaluate the paintings in different ways. One can gain insight into

such problems by constructing graphs or clusters not only for the paintings, and for the rating scales, but also for the subjects, and find ways to combine the information from all three modes into a coherent story about the ways people look at and feel about abstract paintings.

Models and Methods for Three-way Data

Three-mode Component Models

Three-mode analysis as an approach towards analyzing three-way data started with Ledyard Tucker's publications [53, 54, 55]). By early 2004, his work on three-mode analysis was cited about 500 to 600 times in the journal literature. He called his main model *three-mode factor analysis*, but it is now generally referred to as *three-mode component analysis*, or more specifically the *Tucker3 model*. Before his series papers, various authors had investigated ways to deal with sets of matrices, especially from a purely linear algebra point of view, but studying three-way

data really started with Tucker's seminal work. In the earlier papers, Tucker formulated two models (the principal component model and a common factor model) and several computational procedures. He also wrote and collaborated on about 10 applications, not all of them published. Levin, one of his Ph.D. students, wrote an expository paper on applying the technique in psychology [44]. After that, the number of applications and theoretical papers gradually, but slowly, increased. A second major step was taken when Kroonenberg and De Leeuw [38] presented an improved, least-squares solution for the original component model as well as a computer program to carry out the analysis. Kroonenberg [37] also presented an overview of the then state of the art with respect to Tucker's component model, as well as an annotated bibliography [36].

Table 1 gives an overview of the major three-mode models which are being used in practice. Without going into detail here, a perusal of the table will make clear how these models are related to one another by imposing restrictions or adding extensions.

Table 1 Major component and scaling models for three-way data

Model	Sum notation	Matrix and vector notation
SVD	$x_{ij} \approx \sum_{s=1}^S w_{ss} (a_{is} b_{js})$	$\mathbf{X} = \mathbf{A} \mathbf{W} \mathbf{B}' = \sum_{s=1}^S w_{ss} (\mathbf{a}_s \otimes \mathbf{b}_s)$
Tucker2	$x_{ijk} \approx \sum_{p=1}^P \sum_{q=1}^Q h_{pqk} (a_{ip} b_{jq})$	$\mathbf{X}_k \approx \mathbf{A} \mathbf{H}_k \mathbf{B}'$
Tucker3	$x_{ijk} \approx \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} (a_{ip} b_{jq} c_{kr})$	$\underline{\mathbf{X}} \approx \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} (\mathbf{a}_p \otimes \mathbf{b}_q \otimes \mathbf{c}_r)$
Parafac	$x_{ijk} \approx \sum_{s=1}^S \tilde{w}_{kss} (\tilde{a}_{is} \tilde{b}_{js}) [\tilde{w}_{kss} = g_{sss} \tilde{c}_{ks}]$ $x_{ijk} \approx \sum_{s=1}^S g_{sss} (\tilde{a}_{is} \tilde{b}_{js} \tilde{c}_{ks})$	$\mathbf{X}_k \approx \tilde{\mathbf{A}} \tilde{\mathbf{W}}_k \tilde{\mathbf{B}}'$ $\underline{\mathbf{X}} \approx \sum_{s=1}^S g_{sss} (\tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{b}}_s \otimes \tilde{\mathbf{c}}_s)$
IDIOSCAL	$x_{ii'k} \approx \sum_{p=1}^P \sum_{p'=1}^P h_{pp'k} (a_{ip} a_{i'p'})$	$\mathbf{X}_k \approx \mathbf{A} \mathbf{H}_k \mathbf{A}'$
INDSCAL	$x_{ii'k} \approx \sum_{s=1}^S \tilde{w}_{kss} (\tilde{a}_{is} \tilde{a}_{i's}) [\tilde{w}_{kss} = g_{sss} \tilde{c}_{ks}]$ $x_{ijk} \approx \sum_{s=1}^S g_{sss} (\tilde{a}_{is} \tilde{a}_{i's} \tilde{c}_{ks})$	$\mathbf{X}_k \approx \tilde{\mathbf{A}} \tilde{\mathbf{W}}_k \tilde{\mathbf{A}}'$ $\underline{\mathbf{X}} \approx \sum_{s=1}^S g_{sss} (\tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{c}}_s)$

Notes: $\underline{\mathbf{G}} = (g_{pqr})$ is a full $P \times Q \times R$ core array, $\mathbf{H}_k$ is a full $P \times Q$ slice of the extended core array; $\mathbf{W}$ and $\mathbf{W}_k$ are diagonal $S \times S$ matrices. Unless adorned with a tilde ($\sim$), $\mathbf{A} = (a_{ip})$, $\mathbf{B} = (b_{jq})$, and $\mathbf{C} = (c_{kr})$ are orthonormal. The scaling models are presented in their inner-product form.

4 Three-mode Component and Scaling Methods

These relationships are explained in some detail in [3, 33, 34], and [37].

Three-mode Factor Models

A stochastic version of Tucker's common three-mode factor model was first proposed by Bloxom [8], and this was further developed by Bentler and coworkers ([5, 6, 7], and [43]). Bloxom [9] discussed Tucker's factor models in term of higher-order composition rules. Much later the model was treated in extenso with many additional features by Oort ([46, 47]), while Kroonenberg and Oort [39] discuss the link between stochastic three-mode factor models and three-mode component models for what they call multimode covariance matrices. Multimode PLS models were developed by Lohmöller [45].

Parallel Factor Models

Parallel to the development of three-mode component analysis, Harshman ([25, 26]) conceived a component model which he called the *parallel factors model* – (PARAFAC). He conceived this model as an extension of regular component analysis and, using the parallel proportional profiles principle proposed by Cattell [18], he showed that the model solved the rotational indeterminacy of ordinary two-mode **principal component analysis**.

At the same time, Carroll and Chang [15] proposed the same model, calling it canonical decomposition (CANDECOMP). However, their development was primarily related to individual differences scaling, and their main contribution was to algorithmic aspects of the model without further developing the full potential for the analysis of 'standard' three-way arrays. This is the main reason why in this article the model is consistently referred to as the Parafac model.

A full-blown exposé of the model and some extensions is contained in [27] and [28], a more applied survey can be found in [29], and a tutorial with a chemical slant in [11]. The Parafac model has seen a large upsurge in both theoretical development and applications, when it was realized in (analytical) chemistry that physical models of the same form were encountered frequently and the parameter estimation of these models could be solved by the Parafac/Candecomp algorithm; for details see the book by Smilde, Geladi, and Bro [52].

Three-mode Models for Categorical Data

Van Mechelen and coworkers have developed a completely new paradigm for tackling binary three-mode data using Boolean algebra to construct models and express relations between parameters, see especially [19]. Another important approach to handling categorical data was presented by Sands and Young [50] who used optimal scaling of the categorical variables in conjunction with the Parafac model. Large three-way **contingency tables** have been tackled with three-mode **correspondence analysis** [13] and association models [2].

Hierarchies of Three-mode Component Models

Hierarchies of three-way models from both the French and Anglo-Saxon literature, which include the Tucker2 and Tucker3 models respectively, have been presented by Kiers ([33, 34]).

Individual Differences Scaling Models

The work of Carroll and Chang [15] on individual differences multidimensional scaling, or the INDSCAL model, formed a milestone in three-way analysis, and by early 2004 their paper was cited around 1000 times in the journal literature. They extended existing procedures for two-way data, mostly symmetric summed similarity matrices, to three-way data, building upon less far-reaching earlier work of Horan [30]. Over the years, various relatives of this model have been developed, and important from this article's point of view are the IDIOSCAL model [16] a less restricted variant of INDSCAL, the Parafac model which can be interpreted as an asymmetric INDSCAL model and the Tucker2 model which also belongs to the class of individual differences models of which it is the most general representative. The INDSCAL model has also been called a '*generalized subjective metrics model*' [50]. Other similar models have been developed within the context of **multidimensional scaling**, and general discussions of individual differences models and their interrelationships can, for instance, be found in [3] and [56] and their references.

Three-mode Cluster Models

Carroll and Arabie [14] developed a clustering version of the INDSCAL model with a set of common

clusters with each sample or subject in the third mode having individual weights associated with these clusters. The procedure was called *individual differences clustering*—INDCLUS and is applied to sets of similarity matrices. Within that tradition, several further models were suggested, including some (ultrametric) **tree models** (see, for example, [17] and [22]).

Sato and Sato [51] presented a **fuzzy clustering method** for three-mode data by treating the problem as a multicriteria optimization problem and searching for a Pareto efficient solution. Coppi [21] is another contribution to this area.

On the basis of multivariate modeling using **maximum likelihood estimation** Basford [4] developed a mixture method approach to clustering three-mode continuous data and this approach has seen considerable application in agriculture. Extensions to categorical data can be found in Hunt and Basford [31], while further contributions in this vein have been made by Rocci and Vichi [48].

Other Three-way and Three-mode Models and Techniques

In several other fields, three-mode and three-way developments have taken place such as **unfolding**, block models, **longitudinal data**, clustering trajectories, conjoint analysis, PLS modeling, and so on, but an overview will not be given here. In France, several techniques for three-mode analysis have been developed, especially STATIS [41] and AFM [23] which are being used in francophone and Mediterranean countries, but not much elsewhere. An extensive bibliography on the website of the Three-Mode Company contains references to most of the papers dealing with three-mode and three-way issues (<http://three-mode.leidenuniv.nl/>).

Multiway and Multimode Models

Several extensions now exist generalizing three-mode techniques to multiway data. The earliest references are probably [15] for multiway CANDECOMP, [40] for generalizing Tucker's nonleast squares solution to the Tucker4 model, and [32] for generalizing the least-squares solution to analyze the Tucker n model. The book [52] contains the references to the many developments that have taken place especially in chemometrics with respect to multiway modeling.

Detailed Examples

In this section, we will present two examples of the analysis of three-way data: one application of the Tucker3 model and one application of individual differences scaling.

A Tucker3 Component Analysis: Stress and Coping at School

Coping Data: Description. The small example data set to demonstrate three-mode data consists of the ratings of 14 selected Dutch primary school children (mean age 9.8 years). On the basis of the results of a preliminary three-mode analysis, these 14 children were selected from a larger group of 390 children so that they were maximally different from each other and had relatively clear structures in the analysis (for a full account, see [49]).

The children were presented with a questionnaire describing six situations: Restricted in class by the teacher (*TeacherNo*), restricted at home by the mother (*MotherNo*), too much work in class (*WorkLoad*), class work was too difficult (*TooDifficult*), being bullied at school (*Bullied*) and not being allowed to participate in play with the other children (*NotParticipate*). For each of these situations, the children had to indicate what they felt (*Emotions*: Sad, Angry, Annoyed, Afraid) and how they generally dealt with the situation (*Strategies*: Avoidance Coping, Approach Coping, Seeking Social Support, Aggression). The data set has the form of 6 situations by 8 emotions and strategies by 13 children.

The specific interest of the present study is whether children have a different organization of the interrelations between situations, and emotions & strategies. In particular, we assume that there is a single configuration of situations and a single configuration of emotions & strategies, but not all children use the same strategies and do not have the same emotions when confronted with the various situations. In terms of the analysis model, we assume that the children may rotate and stretch or shrink the two configurations before they are combined. In other words, the combined configuration of the situations and the emotions & strategies may look different from one child to another.

The Tucker3 model and fit to the data. The Tucker3 model (see Table 1) has component matrices

6 Three-mode Component and Scaling Methods

for each of the three modes **A**, **B**, and **C**, and a core array **G** which contains information about the strength of the relationships of the components from the three modes. In particular, in the standard form of the model, g_{pqr}^2 indicates the explained variability of the p th components of the children, the q th component of the emotions & strategies and the r th component of the situations (see Figure 6, Panel 1).

The model chosen for this example was the $3 \times 3 \times 2$ -model with 3 children components, 3 emotion & strategy components, and 2 situation components with a relative fit of 0.45. The components of the modes account for 0.27 0.11, and 0.07 (children), 0.26, 0.12, and 0.07 (emotions & strategies), and 0.29 and 0.16 (situations); all situations fit reasonably (i.e., about average = 0.45) except for restricted by the mother (*MotherNo*, relative fit = 0.18). The standardized residuals of the emotions and strategies were about equal, but the variability in Afraid, Social Support, and Aggression was not well fitted compared to the average of 0.45 (0.17, 0.10, and 0.19, respectively). The relative fit of most children was around the average, except that children 6 and 14 fitted rather badly (0.14 and 0.08, respectively).

Interpretation. *Individual differences between children.* The 14 children were children were selected to show individual differences and these differences are evident from the graphs of the subject space: Component 1 versus 2 (Figure 4) and Components 1 versus 3 (Figure 5). To get a handle on the kind of individual differences the scores of the children on the components were correlated with background variables available. The highest correlations for the first component were with ‘Happy in school’, ‘Quality of relationship with the teacher’, ‘Having a good time at school’ (average about 0.65). On the whole for most children except for 1 and 6 who have negative values on the first component, their scores on the first component go together with positive scores on general satisfaction with school. The second component correlated with ‘Not ill versus ill’ (however, only 10 and 1 were ill at the time) and Emotional support by the teacher (about 0.55 with 3,6,7 low scores and 1,8,13 high ones). Finally, the third component correlated 0.70 with Internalizing problem behavior (a CBCL scale—see [1]), but only 8 of the 14 have valid scores on this variable, so that this cannot be taken too seriously. The height of the correlations

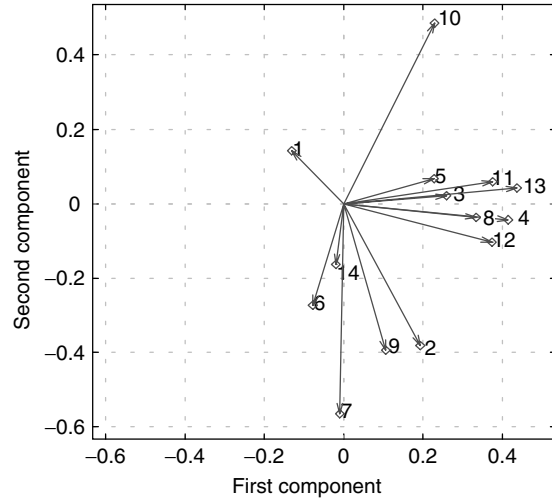


Figure 4 Coping data: three-dimensional children space: 1st versus 2nd Component

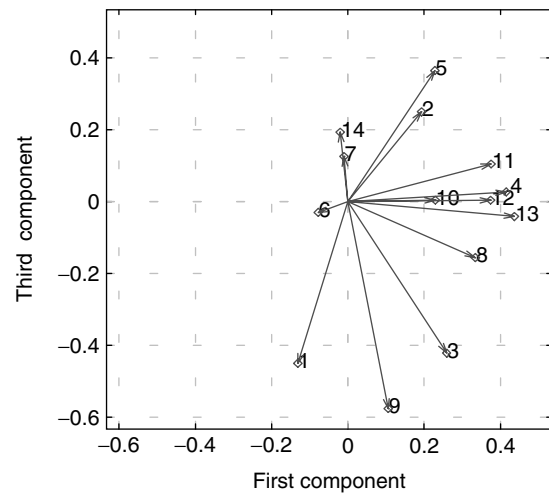


Figure 5 Coping data: three-dimensional children space: 1st versus 3rd Component

makes that we can use these variables for interpretation of the biplots as shown below, but correlations based on 14 scores with missing values on some of them do not make for very strong statements.

Even though the correlations are all significant, one should not read too much into them with respect to generalizing to the sample (and population) from

which they were drawn, as the children were a highly selected set from the total group, but it serves to illustrate that appropriate background information can be used to enhance the interpretability of the subject space.

How children react differently in different situations. The whole purpose of the analysis is to see whether and how the children use different coping strategies and have different emotions in the situation presented to them. To evaluate this, we may look at joint biplots (see [37], Chap. 6). Figure 6 gives a brief summary of their construction. For, say, the r th component of the third mode C , the corresponding core slice G_r is divided between the other two modes A and B to construct the coordinates for the joint **biplot**. In particular, A^* and B^* are computed using a singular value decomposition of the core slice $G_r = U_r \Lambda_r V_r$.

Evaluating several possibilities, it was decided to use the Situation mode as reference mode (i.e., mode C in Figure 6), so that the Children and Emotions & Strategies appear in the joint biplot. For the first situation component the three joint biplot dimensions accounted for 21.8%, 7.2%, and 0.4% respectively so that only the first two biplot dimensions needed to be portrayed, and the same was true for the second situation component where the three joint biplot dimensions accounted for 10.3%, 5.2%, and 0.0001%, respectively.

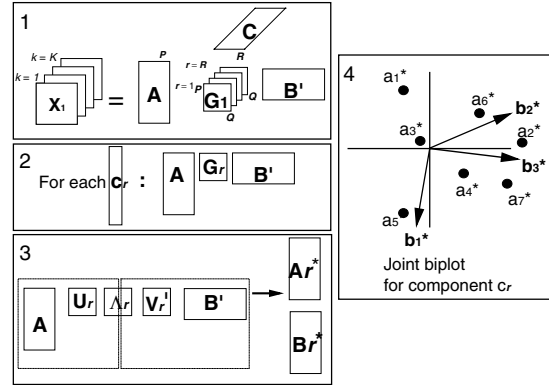


Figure 6 Joint biplot construction for the r th component of the third mode C

To facilitate interpretation, two joint biplots are presented for the *first situation* component. One plot (Figure 7) for the situations loading positively on the component (Being bullied and Not being allowed to participate), and one plot (Figure 8) for the situations loading negatively on the component (Class work too difficult and Too much work in class). This can be achieved by mirroring one of the modes around the origin. Here, this was done for the emotions & strategies. The origin in this graph is the estimated mean scores for all emotions & strategies and a zero

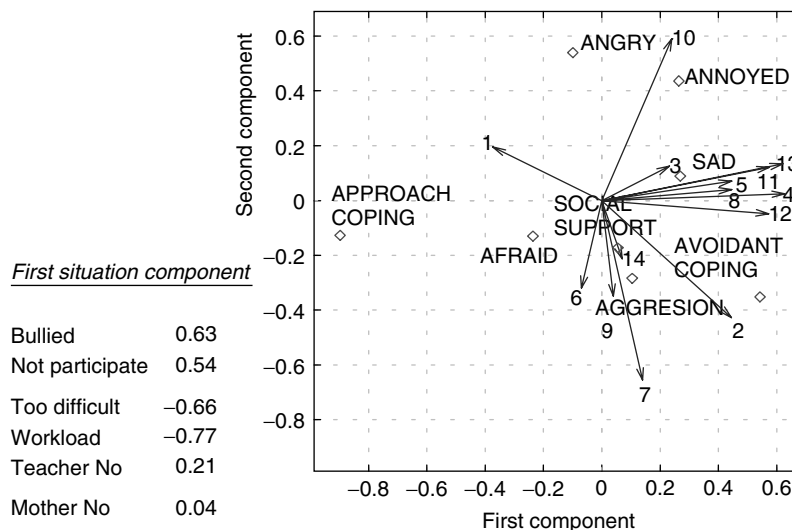


Figure 7 Coping Data: Joint biplot plot for bullying and not being allowed to participate (situations with positive scores on first situation component)

8 Three-mode Component and Scaling Methods

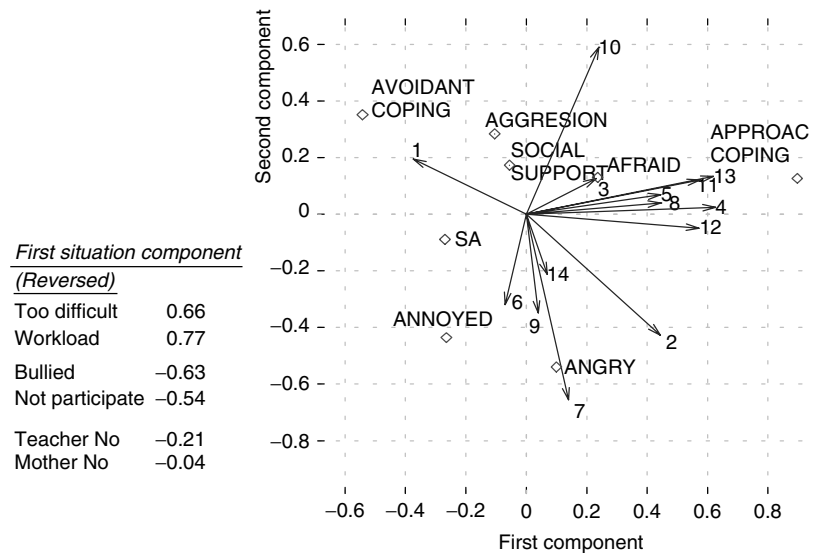


Figure 8 Coping data: joint biplot plot for class work too difficult and too much work in class (Situations with negative scores on first situation component)

value on the component of the situations represent the estimated mean of that situation. To interpret the plot, the projections of children on the emotions & strategies are used, with high positive values indicating that the child has comparatively high scores for such an emotion (strategy). The situations weight the values of these projections, so that it can be established whether the child uses a particular emotion in a particular situation relative to other situations.

Bullying and not being allowed to participate (1st situation component). Most children, especially 5, 8, 9, 11, 12, and 13, use an avoidant coping strategy comparatively more often than an approach coping strategy when bullied or being left out (Figure 7). From the external variables we know that these are typically the children who are happy at school and who do well. Only child 1 (which is not so happy at school and does not do so well) seems to do the reverse, that is, using an approach coping rather than an avoidant coping strategy. Child 10 who according to the external variables is more ill than the others, is particularly angry and annoyed in such situations, while 2 and 7 resort towards aggressive behavior. Large differences with respect to social support are not evident.

Class work too difficult and too much work in class (1st situation component). When faced with too difficult or too much class work, the well-adjusted children, especially 5, 8, 9, 11, 12, and 13, use an approach coping strategy comparatively more often than an avoidant coping strategy, and only 1 seems to do the reverse, that is, using an avoidant rather than an approach coping strategy (Figure 8). Child 10 who is more ill than the others, is somewhat aggressive in such situations, while the more robust children 2 and 7 are particularly angry and rather annoyed. Large differences with respect to social support are not evident.

Restricted by the teacher and the mother (2nd situation component). These two situations have high loadings on the second situation components and not on the first, so that the joint plot associated with this component should be inspected. For the children who are happy at school and do well (i.e., 3, 4, 8, 12, 13), being restricted by the teacher and to a lesser extent by the mother is particularly met with avoidance coping, 10 and 11 are comparatively angry as well, 5 seeks some social support and is relatively angry, afraid, and sad (Figure 9). Child 14 is fairly unique in that it uses more approach coping, but it is comparatively sad and annoyed as well. Children 2, 6, and 7 are primarily more

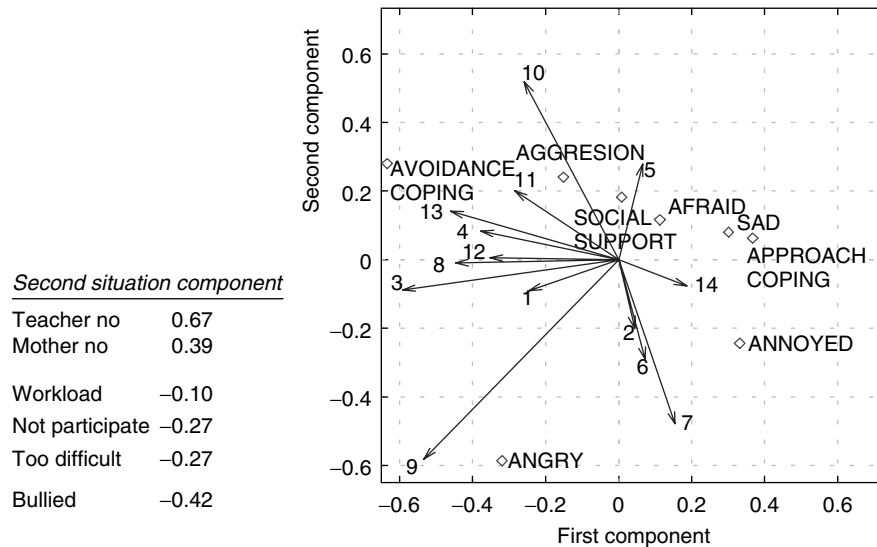


Figure 9 Coping data: Joint biplot plot for class work too difficult and Too much work in class. (Situations with positive scores on second situation component)

angry and annoyed but do not favor one particular strategy over another, and finally child 9's reaction is primarily one of anger over any other emotions.

Note that being bullied also has sizable (negative) loading on the second situation dimension, indicating that being bullied is more complex than is shown in Figure 7 and the reverse pattern from the one discussed for being restricted by the teacher and the mother is true for bullying. To get a complete picture for bullying the information of the two joint plots should be combined, which is not easy to do. In a paper especially devoted to the substantive interpretation of the data, one would probably search for a rotation of the situation space such that bullying loads only on one component and interpret the associated joint plot especially for bullying; however, this will not be pursued here.

INDSCAL and IDIOSCAL: Typology of Pain

Pain data: Description. In this example, subjects were requested to indicate the similarities between certain pain sensations. The question is whether the subjects perceived pain in a similar manner, and in which way and to what extent pain sensations were considered as being similar.

The similarities were converted to dissimilarities to make them comparable to distances, but we will treat the dissimilarities as *squared* distances. As is shown in the MDS-literature, double-centering squared distances gives scalar products which can be analyzed by scalar-product models, such as INDSCAL and IDIOSCAL (see [15] and also Table 1). The INDSCAL model assumes that there exists a common stimulus configuration, which is shared by all judges (subjects), and that this configuration has the same structure for all judges, except that they may attach different importance (salience) to each of the (fixed) axes of the configuration. This results in some judges having configurations which are stretched out more along one of the axes. The IDIOSCAL model is similar except that each judge may rotate the axes of the common configuration over an angle before stretching.

Apart from the investigation into pain perception, we were also interested in examining for this data set Arabie, Carroll, and De Sarbo's [3] claim that IDIOSCAL '[...] has empirically yielded disappointing results in general' (p. 45) by comparing the IDIOSCAL and INDSCAL models.

Results. On the basis of a preliminary analysis, 16 of the 41 subjects were chosen for this example on the basis of the fit of the IDIOSCAL model to their data. The analysis reported here is a two-component

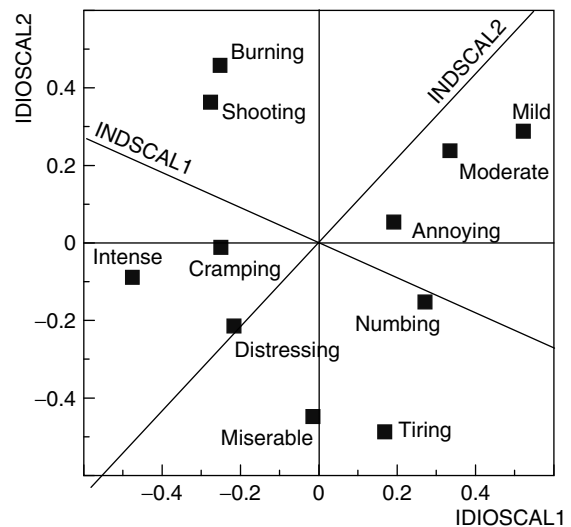
Table 2 Pain Data: IDIOSCAL subject weights and cosines and INDSCAL subject weights (sorted with respect to INDSCAL weights)

Type of subject	IDIOSCAL subject weights			IDIOSCAL cosines	INDSCAL subject weights	
	(1,1)	(2,2)	(1,2)		(1,1)	(2,2)
Control	0.71	0.15	-0.27	-0.82	0.45	0.03
Chronic Pain	0.59	0.20	-0.22	-0.64	0.38	0.05
Chronic Pain	0.66	0.25	-0.18	-0.46	0.37	0.13
Control	0.53	0.22	-0.15	-0.44	0.31	0.09
Control	0.66	0.09	-0.03	-0.13	0.29	0.19
Chronic Pain	0.41	0.23	-0.24	-0.79	0.28	0.02
Control	0.50	0.14	-0.06	-0.22	0.24	0.14
RSI Pain	0.42	0.32	0.27	0.75	0.07	0.42
RSI Pain	0.53	0.32	0.19	0.45	0.16	0.38
Chronic Pain	0.34	0.45	0.17	0.44	0.09	0.36
RSI Pain	0.52	0.24	0.15	0.44	0.16	0.33
RSI Pain	0.45	0.29	0.12	0.34	0.15	0.31
Chronic Pain	0.35	0.47	0.08	0.19	0.14	0.30
Control	0.51	0.30	0.08	0.22	0.19	0.30
Control	0.25	0.32	0.08	0.32	0.08	0.23
RSI Pain	0.52	0.09	0.04	0.16	0.22	0.19

IDIOSCAL solution with a fit of 39.2%, indicating that the data are very noisy. In Figure 10 violent pains, such as shooting, burning, cramping, intense pain are in the same region of the space, as are the less dramatic ones, such as mild, moderate, and annoying, and also tiring, miserable, and distressing form a group.

The subject weights are shown in the left-hand panel of Table 2. The table provides the weights allocated to the first dimension and to the second dimension as well as the ‘individual orientation’ expressed as a cosine between the two dimensions. Clearly the subjects fall in two groups, those that put the dimensions under an acute angle and those that put them at an obtuse angle. Proof enough for an individual differences of orientation scaling, it seems. However, one problem is that for identifiability of the model, it has to be assumed that the stimulus space is orthogonal. To check whether this was problematic we also performed an INDSCAL analysis. This analysis provided a fit of 38.3%, hardly worse than the previous analysis, and given that its interpretation is more straightforward it is clearly to be preferred. The additional complexity of the IDIOSCAL model was only apparent in this case and the results support the conclusion in [3].

In Figure 10, we have drawn the orientation of the two INDSCAL axes, which have an inner product

**Figure 10** Grigg pain data: two-dimensional IDIOSCAL stimulus space with the best fitting INDSCAL axes

of -0.31 and thus make an angle of 108 degrees. In the right hand panel of Table 2, we see the INDSCAL subject weights, which also show the two groups found earlier. Staying with the basic INDSCAL interpretation, we see that one group of subjects (1,2,3,10,11,12,13,14) tends to emphasize the axis of

burning, shooting, intense, cramping pain in contrast with mild, numbing, and tiring. The other group of subjects (5,6,7,8,9,15,16) contrast mild and moderate pain with intense, tiring, distressing, and miserable, and place burning and shooting somewhere in the middle.

In the original design, the subjects consisted of three groups: chronic pain sufferers, repetitive-strain-injury sufferers, and a control group. If the information available is correct, then the empirical division into two groups runs right through two of the design groups, but all of the RSI sufferers are in the second group. In drawing conclusions, we have to take into consideration that the subjects in this example are a special selection from the real sample.

Further Reading

The earliest book length treatment of three-mode component models is [37], followed by [12], which emphasizes chemistry application, as does the recent book [52]. Multiway scaling models were extensively discussed in [3], and a recent comprehensive treatment of multidimensional scaling, which contains chapters on three-way scaling methods, is [10], while also [24] is an attractive recent book on the analysis of proximity data which also pays attention to three-way scaling.

Two collections of research papers on three-mode analysis have been published, which contain contributions of many people who were working in the field at the time. The first collection, [42], contains full-length overviews including the most extensive treatment of the Parafac model [27] and [28]. The second collection, [20], consists of papers presented at the 1988 conference on Multiway analysis in Rome. Finally, several special issues on three-mode analysis have appeared over the years: *Computational Analysis & Data Analysis*, 18(1) in 1994, *Journal of Chemometrics*, 14(3) in 2000, and *Journal of Chemometrics*, 18(1) in 2004.

Acknowledgments

This contribution was written while Fellow-in-Residence at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS). My thanks go to Irma Röder (Coping data) and Lynn Grigg (University of Queensland, Brisbane, Australia; Pain data) for their

permission to use their data sets as examples. The last example is a reworked version of an earlier paper and is reproduced by permission of Springer Verlag, Berlin.

References

- [1] Achenbach, T.M. & Edelbrock, C. (1983). *Manual for the Child Behavior Check List and revised child behavior profile*, Department of Psychiatry, University of Vermont, Burlington.
- [2] Anderson, C.J. (1996). The analysis of three-way contingency tables by three-mode association models, *Psychometrika* **61**, 465–483.
- [3] Arabie, P., Carroll, J.D. & DeSarbo, W.S. (1987). *Three-way scaling and clustering*, Sage Publications, Beverly Hills.
- [4] Basford, K.E. & McLachlan, G.J. (1985). The mixture method of clustering applied to three-way data, *Journal of Classification* **2**, 109–125.
- [5] Bentler, P.M. & Lee, S.-Y. (1978). Statistical aspects of a three-mode factor analysis model, *Psychometrika* **43**, 343–352.
- [6] Bentler, P.M. & Lee, S.-Y. (1979). A statistical development of three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **32**, 87–104.
- [7] Bentler, P.M., Poon, W.-Y. & Lee, S.-Y. (1988). Generalized multimode latent variable models: Implementation by standard programs, *Computational Statistics and Data Analysis* **6**, 107–118.
- [8] Bloxom, B. (1968). A note on invariance in three-mode factor analysis, *Psychometrika* **33**, 347–350.
- [9] Bloxom, B. (1984). Tucker's three-mode factor analysis model, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie, & R.P. McDonald, eds, Praeger, New York, pp. 104–121.
- [10] Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York. (Chapters 20 and 21).
- [11] Bro, R. (1997). PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171.
- [12] Bro, R. (1998). Multi-way analysis in the food industry. Models, algorithms, and applications, PhD thesis, University of Amsterdam, Amsterdam.
- [13] Carlier, A. & Kroonenberg, P.M. (1996). Decompositions and biplots in three-way correspondence analysis, *Psychometrika* **61**, 355–373.
- [14] Carroll, J.D. & Arabie, P. (1983). INDCLUS: an individual differences generalization of the ADCLUS model and the MAPCLUS algorithm, *Psychometrika* **48**, 157–169.
- [15] Carroll, J.D. & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition, *Psychometrika* **35**, 283–319.

- [16] Carroll, J.D. & Chang, J.J. (1972). IDIOSCAL: A generalization of INDSCAL allowing IDIOSyncratic reference system as well as an analytic approximation to INDSCAL, in *Paper Presented at the Spring Meeting of the Classification Society of North America*, Princeton, NJ.
- [17] Carroll, J.D., Clark, L.A. & DeSarbo, W.S. (1984). The representation of three-way proximity data by single and multiple tree structure models, *Journal of Classification* **1**, 25–74.
- [18] Cattell, R.B. & Cattell, A.K.S. (1955). Factor rotation for proportional profiles: analytical solution and an example, *The British Journal of Statistical Psychology* **8**, 83–92.
- [19] Ceulemans, E., Van Mechelen, I. & Leenen, I. (2003). Tucker3 hierarchical classes analysis, *Psychometrika* **68**, 413–433.
- [20] Coppi, R. & Bolasco, S., eds (1989). *Multway Data Analysis*, Elsevier Biomedical, Amsterdam.
- [21] Coppi, R. & D'Urso, P. (2003). Three-way fuzzy clustering models for LR fuzzy time trajectories, *Computational Statistics & Data Analysis* **43**, 149–177.
- [22] De Soete, G. & Carroll, J.D. (1989). Ultrametric tree representations of three-way three-mode data, in *Multway Data Analysis*, R., Coppi & S., Bolasco, eds, Elsevier Biomedical, Amsterdam, pp. 415–426.
- [23] Escofier, B. & Pagès, J. (1994). Multiple factor analysis (AFMULT package), *Computational Statistics and Data Analysis* **18**, 121–140.
- [24] Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*, Arnold, London, (Chapters 5 and 6).
- [25] Harshman, R.A. (1970). Foundations of the PARAFAC procedure: models and conditions for an “explanatory” multi-modal factor analysis, *UCLA Working Papers in Phonetics* **16**, 1–84.
- [26] Harshman, R.A. (1972). Determination and proof of minimum uniqueness conditions for PARAFAC1, *UCLA Working Papers in Phonetics* **22**, 111–117.
- [27] Harshman, R.A. & Lundy, M.E. (1984a). The PARAFAC model for three-way factor analysis and multidimensional scaling, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie & R.P. McDonald, eds, Praeger, New York, pp. 122–215.
- [28] Harshman, R.A. & Lundy, M.E. (1984b). Data preprocessing and the extended PARAFAC model, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie & R.P. McDonald, eds, Praeger, New York, pp. 216–284.
- [29] Harshman, R.A. & Lundy, M.E. (1994). PARAFAC: Parallel factor analysis, *Computational Statistics and Data Analysis* **18**, 39–72.
- [30] Horan, C.B. (1969). Multidimensional scaling: combining observations when individuals have different perceptual structures, *Psychometrika* **34**, 139–165.
- [31] Hunt, L.A. & Basford, K.E. (2001). Fitting a mixture model to three-mode three-way data with categorical and continuous variables, *Journal of Classification* **16**, 283–296.
- [32] Kapteyn, A., Neudecker, H. & Wansbeek, T. (1986). An approach to n -mode components analysis, *Psychometrika* **51**, 269–275.
- [33] Kiers, H.A.L. (1988). Comparison of “Anglo-Saxon” and “French” three-mode methods, *Statistique et Analyse Des Données* **13**, 14–32.
- [34] Kiers, H.A.L. (1991). Hierarchical relations among three-way methods, *Psychometrika* **56**, 449–470.
- [35] Kiers, H.A.L. (2000). Towards a standardized notation and terminology in multiway analysis, *Journal of Chemometrics* **14**, 105–122.
- [36] Kroonenberg, P.M. (1983). Annotated bibliography of three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **36**, 81–113.
- [37] Kroonenberg, P.M. (1983). *Three-Mode Principal Component Analysis: Theory and Applications*, (Errata, 1989; available from the author). DSWO Press, Leiden.
- [38] Kroonenberg, P.M. & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms, *Psychometrika* **45**, 69–97.
- [39] Kroonenberg, P.M. & Oort, F.J. (2003). Three-mode analysis of multi-mode covariance matrices, *British Journal of Mathematical and Statistical Psychology* **56**, 305–336.
- [40] Lastovicka, J.L. (1981). The extension of component analysis to four-mode matrices, *Psychometrika* **46**, 47–57.
- [41] Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). The ACT (STATIS method), *Computational Statistics and Data Analysis* **18**, 97–119.
- [42] Law, H.G. Snyder Jr, C.W., Hattie, J.A. & McDonald, R.P., eds (1984). *Research Methods for Multimode Data Analysis*, Praeger, New York.
- [43] Lee, S.-Y. & Fong, W.-K. (1983). A scale invariant model for three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **36**, 217–223.
- [44] Levin, J. (1965). Three-mode factor analysis, *Psychological Bulletin* **64**, 442–452.
- [45] Lohmöller, J.-B. (1989). *Latent Variable Path Modeling with Partial Least Squares*, Physica, Heidelberg.
- [46] Oort, F.J. (1999). Stochastic three-mode models for mean and covariance structures, *British Journal of Mathematical and Statistical Psychology* **52**, 243–272.
- [47] Oort, F.J. (2001). Three-mode models for multivariate longitudinal data, *British Journal of Mathematical and Statistical Psychology* **54**, 49–78.
- [48] Rocci, R. & Vichi, M. (2003). Simultaneous component and cluster analysis: the between and within approaches, in *Paper presented at the International Meeting of the Psychometric Society*, Chia Laguna (Cagliari), Italy, 7–10 July.
- [49] Röder, I. (2000). Stress in children with asthma. Coping and social support in the school context. PhD thesis, Department of Education, Leiden University, Leiden.
- [50] Sands, R. & Young, F.W. (1980). Component models for three-way data: ALSCOMP3, an alternating least

- squares algorithm with optimal scaling features, *Psychometrika* **45**, 39–67.
- [51] Sato, M. & Sato, Y. (1994). On a multicriteria fuzzy clustering method for 3-way data, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2**, 127–142.
- [52] Smilde, A.K., Geladi, P. & Bro, R. (2004). *Multiway Analysis in Chemistry*, Wiley, Chichester.
- [53] Tucker, L.R. (1963). Implications of factor analysis of three-way matrices for measurement of change, in *Problems in Measuring Change*, C.W., Harris, ed., University of Wisconsin Press, Madison, pp. 122–137.
- [54] Tucker, L.R. (1964). The extension of factor analysis to three-dimensional matrices, in *Contributions to Mathematical Psychology*, H. Gulliksen & N. Frederiksen, eds, Holt, Rinehart and Winston, New York, pp. 110–127.
- [55] Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis, *Psychometrika* **31**, 279–311.
- [56] Young, F.W. & Hamer, R.M. (1987). *Multidimensional Scaling: History, Theory, and Applications*, Lawrence Erlbaum, Hillsdale. (reprinted 1994).

PIETER M. KROONENBERG

Thurstone, Louis Leon

RANDALL D. WIGHT AND PHILIP A. GABLE

Volume 4, pp. 2045–2046

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Thurstone, Louis Leon

Born: May 29, 1887, in Chicago.

Died: September 19, 1955, near Rapid City, Michigan.

Born in 1887 in Chicago to Swedish immigrants, Louis Leon Thunström began school in Berwyn, IL, moved with his parents to Centerville, MS, to Stockholm, Sweden, and back to Jamestown, NY – all before turning 14. Refusing to return to the United States without his books, Leon personally carried his three favorites onboard ship, among them Euclid. At 18, his name appeared in print for the first time in a *Scientific American* letter, suggesting how to manage the tension between Niagara Fall’s tourists and its energy output. Shortly thereafter, to ease assimilation, his parents changed the spelling of the family name to Thurstone. After high school, Thurstone entered Cornell and studied engineering. As one of his undergraduate projects, Thurstone built – and later patented – a motion picture camera and projector that eliminated flicker. Those designs attracted the attention of Thomas Edison, who invited Thurstone to spend the summer following his 1912 master of engineering degree as an assistant in Edison’s lab.

Thurstone became an engineering instructor at the University of Minnesota in the fall of 1912. While teaching, Thurstone pursued an interest in learning and enrolled in undergraduate experimental psychology courses. His interest and the inspired instruction he received prompted him to seek graduate study in psychology. In 1914 at age 27, he entered the University of Chicago. In 1915 and 1916, Thurstone accepted a graduate assistantship in applied psychology from Walter Bingham at the Carnegie Institute of Technology. In 1917, Thurstone received a Ph.D. from the University of Chicago, apparently without being in residence for at least two years. His dissertation, published in 1919, examined the learning curve equation. Thurstone joined the Carnegie faculty and advanced from assistant to full professor and to department chair. Between 1919 and 1923, Thurstone created psychometric instruments assessing aptitude, clerical skill, ingenuity, and intelligence.

Carnegie closed its applied psychology program in 1923, and Thurstone moved to Washington, D.C., to work for the Institute for Government Research, a

foundation trying to improve civil service examinations. The American Council on Education (ACE) was located in the same Dupont Circle building as the foundation’s office. Thurstone engaged the ACE’s staff in conversation centering on creating college admission examinations, and in 1924, the ACE began to financially support Thurstone in that endeavor. The tests Thurstone developed in this initiative during the following years included linguistic and quantitative subscores and evolved into the Scholastic Aptitude Test, or SAT. The year 1924 also saw Thurstone’s marriage to Thelma Gwinn and his accepting an offer to join the University of Chicago faculty.

Measurement theory drew Thurstone’s attention during the early years at Chicago, 1924 through 1928. In contrast to psychophysical scales that related stimulation to experience, Thurstone developed theory and techniques for scaling psychological dimensions without physical referents, deriving accounts using dispersion as a unit of psychological measure [5]. Beginning in the late 1920s, this work grew into procedures that explored the structure of **latent variables** underlying the response patterns he found using his scaled instruments. Thurstone’s influential concept of ‘simple structure’ (see **History of Factor Analysis: A Statistical Perspective**) guided the use of factor analysis in describing psychologically meaningful constructs. The development of multiple **factor analysis** is among his most widely known achievements. Notably, he reinterpreted ‘g’ in **Spearman’s** theory of general intelligence as a special case of a multidimensional factor structure [3, 4].

In the coming years, Leon and Thelma employed factor analytic techniques to improve college entrance exams and to measure primary mental abilities. Their work bore much fruit, including providing the University of Chicago with the country’s first credit by examination. Colleagues also recognized Thurstone’s accomplishments. He was elected president of the American Psychological Association in 1932, elected charter president of the Psychometric Society in 1936, and elected to the National Academy of Sciences (NAS) in 1938. Shortly after the NAS election, Ernest Hilgard reports being a dinner guest in Thurstone’s home and remembers Thurstone express surprise that as the son of immigrants, he (Thurstone) could successfully follow such a circuitous route to academic success.

During much of the twentieth century, Thurstone was a leading figure in psychometric and psychophysical theory and practice and in the investigation of attitudes, intelligence, skills, and values. His commitment to the scientific process is perhaps best seen in the weekly Wednesday evening seminars he conducted in his home – chalkboard and all – over the course of 30 years, often hosting 30 people at a time, conversing primarily over work in progress. After retiring from Chicago in 1952, he accepted a position at the University of North Carolina-Chapel Hill, where he established the L. L. Thurstone Psychometrics Laboratory. The home he and Thelma built near the campus included a seminar room with a built-in chalkboard where the seminar tradition continued unabated. On leave from Chapel Hill in the spring of 1954, Thurstone returned to Sweden as a visiting professor at the University of Stockholm, lecturing there and at other universities in northern Europe. This would be his last trip to the continent. In September 1955, Thurstone died at his summer home on Elk Lake in Michigan's upper peninsula (see [1], [2], [6] for more details of his life and work).

References

- [1] Adkins, D. (1964). Louis Leon Thurston: creative thinker, dedicated teacher, eminent psychologist, in *Contributions to Mathematical Psychology*, N. Frederiksen & H. Gulliksen, eds, Holt Rinehart Winston, New York, pp. 1–39.
- [2] Jones, L.V. (1998). L. L. Thurstone's vision of psychology as a quantitative rational science, in *Portraits of Pioneers in Psychology*, Vol. 3, G.A. Kimble & M. Wertheimer, eds, American Psychological Association, Washington.
- [3] Thurstone, L.L. (1938). *Primary Mental Abilities*, University of Chicago Press, Chicago.
- [4] Thurstone, L.L. (1947). *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind*, University of Chicago Press, Chicago.
- [5] Thurstone, L.L. & Chave, E.J. (1929). *The Measurement of Attitude*, University of Chicago Press, Chicago.
- [6] Thurstone, L.L. (1952). In *A History of Psychology in Autobiography*, Vol. 4, E.G. Boring, H.S. Langfeld, H. Werner & R.M. Yerkes, eds, Clark University Press, Worcester, pp. 295–321.

RANDALL D. WIGHT AND PHILIP A. GABLE

Time Series Analysis

P.J. BROCKWELL

Volume 4, pp. 2046–2055

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Time Series Analysis

Time Series Data

A time series is a set of observations x_t , each one associated with a particular time t , and usually displayed in a *time series plot* of x_t as a function of t . The set of times T at which observations are recorded may be a discrete set, as is the case when the observations are recorded at uniformly spaced times (e.g., daily rainfall, hourly temperature, annual income, etc.), or it may be a continuous interval, as when the data is recorded continuously (e.g., by a seismograph or electrocardiograph). A very large number of practical problems involve observations that are made at uniformly spaced times. For this reason the, present article focuses on this case, indicating briefly how missing values and irregularly-spaced data can be handled.

Example 1 Figure 1 shows the number of accidental deaths recorded monthly in the US. for the years 1973 through 1978. The graph strongly suggests (as is usually the case for monthly data) the presence of a periodic component with period 12 corresponding to the seasonal cycle of 12 months, as well as a smooth trend accounting for the relatively slow change in level of the series, and a random component accounting for irregular deviations from a deterministic model involving trend and seasonal components only.

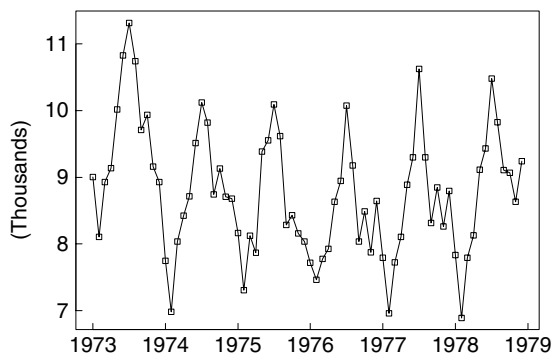


Figure 1 Monthly accidental deaths in the US from 1973 through 1978

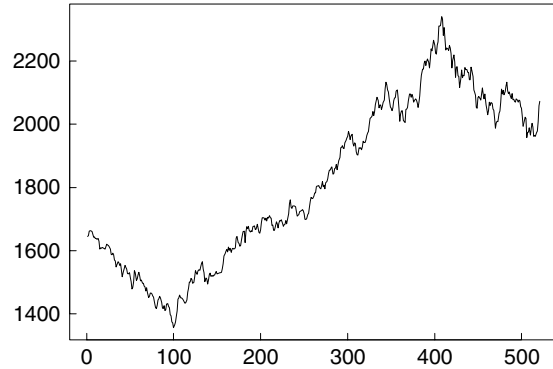


Figure 2 The Australian All-Ordinaries Index of Stock Prices on 521 successive trading days up to July 18, 1994

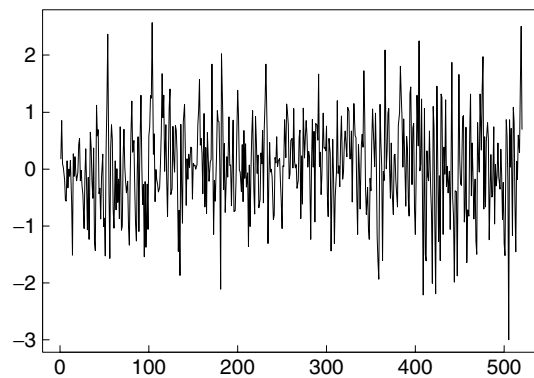


Figure 3 The daily percentage changes in the all-ordinaries index over the same period

Example 2 Figure 2 shows the closing value in Australian dollars of the Australian All-Ordinaries index (an average of 100 stocks sold on the Australian Stock Exchange) on 521 successive trading days, ending on July 18, 1994. It displays irregular variation around a rather strong trend. Figure 3 shows the daily percentage changes in closing value of the index for each of the 520 days ending on July 18, 1994. The trend apparent in Figure 2 has virtually disappeared, and the series appears to be varying randomly around a mean value close to zero.

Objectives

The objectives of time series analysis are many and varied, depending on the particular field of

2 Time Series Analysis

application. From the observations $x_1, \dots, x_n$, we may wish to make inferences about the way in which the data are generated, to predict future values of the series, to detect a ‘signal’ hidden in noisy data, or simply to find a compact description of the available observations.

In order to achieve these goals, it is necessary to postulate a mathematical model (or family of models), according to which we suppose that the data is generated. Once an appropriate family has been selected, we then select a *specific* model by estimating model parameters and checking the resulting model for goodness of fit to the data. Once we are satisfied that the selected model provides a good representation of the data, we use it to address questions of interest. It is rarely the case that there is a ‘true’ mathematical model underlying empirical data, however, systematic procedures have been developed for selecting the best model, according to clearly specified criteria, within a broad class of candidates.

Time Series Models

As indicated above, the graph of the accidental deaths series in Figure 1 suggests representing x_t as the sum of a slowly varying *trend* component, a period-12 *seasonal* component, and a *random* component that accounts for the irregular deviations from the sum of the other two components. In order to take account of the randomness, we suppose that for each t , the observation x_t is just one of many possible values of a random variable X_t that we *might* have observed. This leads to the following *classical decomposition model* for the accidental deaths data,

$$X_t = m_t + s_t + Y_t, \quad t = 1, 2, 3, \dots, \quad (1)$$

where the sequence $\{m_t\}$ is the *trend* component describing the long-term movement in the level of the series, $\{s_t\}$ is a *seasonal* component with known period (in this case, 12), and $\{Y_t\}$ is a sequence of random variables with mean zero, referred to as the *random* component. If we can characterize m_t , s_t , and Y_t in simple terms and in such a way that the model (1) provides a good representation of the data, then we can proceed to use the model to make forecasts or to address other questions related to the series. Completing the specification of the model by estimating the trend and seasonal components and characterizing the random component is a major part

of time series analysis. The model (1) is sometimes referred to as an *additive* decomposition model (see **Additive Models**). Provided the observations are all positive, the *multiplicative* model,

$$X_t = m_t s_t Y_t, \quad (2)$$

can be reduced to an additive model by taking logarithms of each side to get an additive model for the logarithms of the data.

The general form of the additive model (1) supposes that the seasonal component s_t has known period d (12 for monthly data, 4 for quarterly data, etc.), and satisfies the conditions

$$s_{t+d} = s_t \text{ and } \sum_{t=1}^d s_t = 0, \quad (3)$$

while $\{Y_t\}$ is a *weakly stationary sequence of random variables*, that is, a sequence of random variables satisfying the conditions,

$$\begin{aligned} E(Y_t) &= \mu, \quad E(Y_t^2) < \infty \text{ and} \\ \text{Cov}(Y_{t+h}, Y_t) &= \gamma(h) \text{ for all } t, \end{aligned} \quad (4)$$

with $\mu = 0$. The function γ is called the *autocovariance function* of the sequence $\{Y_t\}$, and the value $\gamma(h)$ is the *autocovariance at lag h* . In the special case when the random variables Y_t are independent and identically distributed, the model (1) is a classical regression model and $\gamma(h) = 0$ for all $h \neq 0$. However, in time series analysis, it is the dependence between Y_{t+h} and Y_t that is of special interest, and which allows the possibility of using past observations to obtain forecasts of future values that are better in some average sense than just using the expected value of the series. A measure of this dependence is provided by the autocovariance function. A more convenient measure (since it is independent of the origin and scale of measurement of Y_t) is the *autocorrelation function*,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \quad (5)$$

From observed values $y_1, \dots, y_n$ of a weakly stationary sequence of random variables $\{Y_t\}$, good estimators of the mean $\mu = E(Y_t)$ and the autocovariance

function $\gamma(h)$ are the *sample mean* and *sample autocovariance function*,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i \quad (6)$$

and

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (y_{i+|h|} - \hat{\mu})(y_i - \hat{\mu}), \quad -n < h < n, \quad (7)$$

respectively. The autocorrelation function of $\{Y_t\}$ is estimated by the *sample autocorrelation function*,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \quad (8)$$

Elementary techniques for estimating m_t and s_t can be found in many texts on time series analysis (e.g., [6]). More sophisticated techniques are employed in the packages X-11 and the updated version X-12 described in [12], and used by the US Census Bureau. Once estimators $\hat{m}_t$ of m_t and $\hat{s}_t$ of s_t have been obtained, they can be subtracted from the observations to yield the **residuals**,

$$y_t = x_t - \hat{m}_t - \hat{s}_t. \quad (9)$$

A stationary time series model can then be fitted to the residual series to complete the specification of the model. The model is usually chosen from the class of *autoregressive moving average (or ARMA) processes*, defined below in ARMA Processes.

Instead of estimating and subtracting off the trend and seasonal components to generate a sequence of residuals, an alternative approach, developed by Box and Jenkins [4], is to apply *difference operators* to the original series to remove trend and seasonality. The *backward shift operator* B is an operator that, when applied to X_t , gives X_{t-1} . Thus,

$$BX_t = X_{t-1}, \quad B^j X_t = X_{t-j}, \quad j = 2, 3, \dots$$

The *lag-1 difference operator* is the operator $\nabla = (1 - B)$. Thus,

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}. \quad (10)$$

When applied to a polynomial trend of degree p , the operator ∇ reduces it to a polynomial of degree $p - 1$. The operator ∇^p , denoting p successive

applications of ∇ , therefore, reduces any polynomial trend of degree p to a constant. Usually, a small number of applications of ∇ is sufficient to eliminate trends encountered in practice. Application of the *lag-d difference operator*, $\nabla_d = (1 - B^d)$ (not to be confused with ∇^d) to X_t gives

$$\nabla_d X_t = (1 - B^d)X_t = X_t - X_{t-d}, \quad (11)$$

eliminating any seasonal component with period d . In the Box–Jenkins approach to time series modeling, the operators ∇ and ∇_d are applied as many times as is necessary to eliminate trend and seasonality, and the sample mean of the differenced data subtracted to generate a sequence of residuals y_t , which are then modeled with a suitably chosen ARMA process in the same way as the residuals (9).

Figure 4 shows the effect of applying the operator ∇_{12} to the accidental deaths series of Figure 1. The seasonal component is no longer apparent, but there is still an approximately linear trend. Further application of the operator ∇ yields the series shown in Figure 5 with relatively constant level. This new series is a good candidate for representation by a stationary time series model.

The daily percentage returns on the Australian All-Ordinaries Index shown in Figure 2 already show no sign of trend or Seasonality, and can be modeled as a stationary sequence without preliminary detrending or deseasonalizing.

In cases where the variability of the observed data appears to change with the level of the data, a preliminary transformation, prior to detrending and deseasonalizing, may be required to stabilize

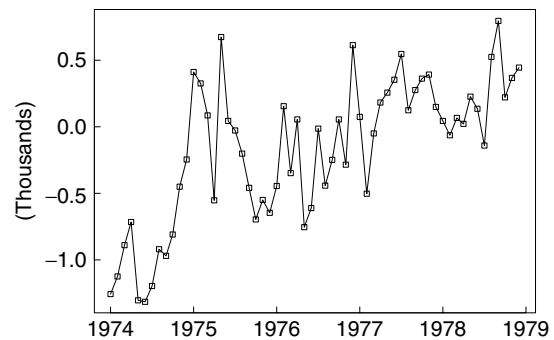


Figure 4 The differenced series $\{\nabla_{12}x_t, t = 13, \dots, 72\}$ derived from the monthly accidental deaths $\{x_1, \dots, x_{72}\}$ shown in Figure 1

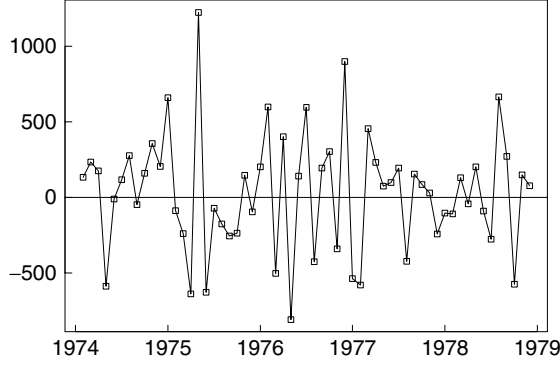


Figure 5 The differenced series $\{\nabla\nabla_{12}x_t, t = 14, \dots, 72\}$ derived from the monthly accidental deaths $\{x_1, \dots, x_{72}\}$ shown in Figure 1

the variability. For this purpose, a member of the family of *Box–Cox transformations* (see e.g., [6]) is frequently used.

ARMA Processes

For modeling the residuals $\{y_t\}$ (found as described above), a very useful parametric family of zero-mean stationary sequences is furnished by the *autoregressive moving average (or ARMA) processes*. The $\text{ARMA}(p, q)$ process $\{Y_t\}$ with autoregressive coefficients, $\phi_1, \dots, \phi_p$, moving average coefficients, $\theta_1, \dots, \theta_q$, and white noise variance σ^2 , is defined as a weakly stationary solution of the difference equations,

$$\begin{aligned} (1 - \phi_1 B - \dots - \phi_p B^p)Y_t \\ = (1 + \theta_1 + \dots + \theta_q B^q)Z_t, \quad t = 0, \pm 1, \pm 2, \dots, \end{aligned} \quad (12)$$

where B is the backward shift operator, the polynomials $\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$ have no common factors, and $\{Z_t\}$ is a sequence of uncorrelated random variables with mean zero and variance σ^2 . Such a sequence $\{Z_t\}$ is said to be *white noise with mean 0 and variance σ^2* , indicated more concisely by writing $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

The equations (12) have a unique stationary solution if, and only if, the equation $\phi(z) = 0$ has no root with $|z| = 1$, however, the possible values of $\phi_1, \dots, \phi_p$ are usually assumed to satisfy the stronger

restriction,

$$\phi(z) \neq 0 \text{ for all complex } z \text{ such that } |z| \leq 1. \quad (13)$$

The unique weakly stationary solution of equation (12) is then

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad (14)$$

where ψ_j is the coefficient of z^j in the power-series expansion,

$$\frac{\theta(z)}{\phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j, \quad |z| \leq 1.$$

Since Y_t in (14) is a function only of $Z_s, s \leq t$, the series $\{Y_t\}$ is said to be a *causal function* of $\{Z_t\}$, and the condition (13) is called the *causality condition* for the process (12). (Condition (13) is also frequently referred to as a *stability condition*.)

In the causal case, simple recursions are available for the numerical calculation of the sequence $\{\psi_j\}$ from the autoregressive and moving average coefficients $\phi_1, \dots, \phi_p$, and $\theta_1, \dots, \theta_q$ (see e.g., [6]). To every noncausal ARMA process, there is a causal ARMA process with the same autocovariance function, and, *under the assumption that all of the joint distributions of the process are multivariate normal*, with the same joint distributions. This is one reason for restricting attention to causal models. Another practical reason is that if $\{Y_t\}$ is to be simulated sequentially from the sequence $\{Z_t\}$, the causal representation (14) of Y_t does not involve *future* values $Z_{t+h}, h > 0$.

A key feature of ARMA processes for modeling dependence in a sequence of observations is the extraordinarily large range of autocorrelation functions exhibited by ARMA processes of different orders (p, q) as the coefficients $\phi_1, \dots, \phi_p$, and $\theta_1, \dots, \theta_q$, are varied. For example, if we take any set of observations, $x_1, \dots, x_n$ and compute the sample autocorrelations, $\hat{\rho}(0)(= 1), \hat{\rho}(1), \dots, \hat{\rho}(n-1)$, then for any $k < n$, it is always possible to find a causal AR(k) process with autocorrelations $\rho(j)$ satisfying $\rho(j) = \hat{\rho}(j)$ for every $j \leq k$.

The mean of the ARMA process defined by (12) is $E(Y_t) = 0$, and the autocovariance function can

be found from the equations obtained by multiplying each side of (12) by Y_{t-j} , $j = 0, 1, 2, \dots$, taking expectations and solving for $\gamma(j) = E(Y_t Y_{t-j})$. Details can be found in [6].

Example 3 The causal ARMA(1,0) or AR(1) process is a stationary solution of the equations

$$Y_t = \phi Y_{t-1} + Z_t, \quad |\phi| < 1, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

The autocovariance function of $\{Y_t\}$ is $\rho(h) = \sigma^2 \phi^{|h|} / (1 - \phi^2)$.

Example 4 The ARMA(0, q) or MA(q) process is the stationary series defined by

$$Y_t = \sum_{j=0}^q \theta_j Z_{t-j}, \quad \theta_0 = 1, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

The autocovariance function of $\{Y_t\}$ is $\gamma(h) = \sigma^2 \sum_{j=0}^{|h|} \theta_j \theta_{j+|h|}$ if $h \leq q$ and $\gamma(h) = 0$ otherwise.

A process $\{X_t\}$ is said to be an *ARMA process with mean μ* if $\{Y_t = X_t - \mu\}$ is an ARMA process as defined by equations of the form (12)).

In the following section, we consider the problem of fitting an ARMA model of the form (12), that is, $\phi(B)Y_t = \theta(B)Z_t$, to the residual series $y_1, y_2, \dots$, generated as described in Time Series Models. If the residuals were obtained by applying the differencing operator $(1 - B)^d$ to the original observations $x_1, x_2, \dots$, then we are effectively fitting the model

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2) \quad (15)$$

to the original data. If the order of the polynomials $\phi(B)$ and $\theta(B)$ are p and q respectively, then the model (15) is called an *ARIMA(p, d, q) model* for $\{X_t\}$. If the residuals y_t had been generated by differencing also at some lag greater than 1 to eliminate seasonality, say, for example, $y_t = (1 - B^{12})(1 - B)^d x_t$, and the ARMA model (12) were then fitted to y_t , then the model for the original data would be a more general ARIMA model of the form,

$$\begin{aligned} \phi(B)(1 - B^{12})(1 - B)^d X_t &= \theta(B)Z_t, \\ \{Z_t\} &\sim \text{WN}(0, \sigma^2). \end{aligned} \quad (16)$$

Selecting and Fitting a Model to Data

In Time Series Models, we discussed two methods for eliminating trend and seasonality with the aim of transforming the original series to a series of residuals suitable for modeling as a zero-mean stationary series. In ARMA Processes, we introduced the class of ARMA models with a wide range of autocorrelation functions. By suitable choice of ARMA parameters, it is possible to find an ARMA process whose autocorrelations match the sample autocorrelations of the residuals $y_1, y_2, \dots$, up to any specified lag. This is an intuitively appealing and natural approach to the problem of fitting a stationary time series model. However, except when the residuals are truly generated by a purely autoregressive model (i.e., an ARMA($p, 0$) model), this method turns out to give greater large-sample mean-squared errors than the method of *maximum Gaussian likelihood* described in the following paragraph.

Suppose for the moment that we know the orders p and q of the ARMA process (12) that is to be fitted to the residuals $y_1, \dots, y_n$, and suppose that $\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \sigma^2)$, is the vector of parameters to be estimated. The Gaussian likelihood $L(\beta; y_1, \dots, y_n)$, is the likelihood computed under the assumption that the joint distribution from which $y_1, \dots, y_n$ are drawn is multivariate normal (see **Maximum Likelihood Estimation**). Thus,

$$\begin{aligned} L(\beta; y_1, \dots, y_n) &= (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \\ &\times \exp \left(-\frac{1}{2} \mathbf{y}_n \Gamma_n^{-1} \mathbf{y}_n \right), \end{aligned} \quad (17)$$

where $\mathbf{y}_n = (y_1, \dots, y_n)'$, Γ_n is the matrix of autocovariances $[\gamma(i - j)]_{i,j=1}^n$, and $\gamma(h)$ is the autocovariance function of the model defined by (12). Although direct calculation of L is a daunting task, L can be reexpressed in the *innovations form* of Schweppe [22], which is readily calculated from the minimum mean-squared error one-step linear predictors of the observations and their mean-squared errors. These in turn can be readily calculated from the *innovations algorithm* (see [5]).

At first glance, maximization of Gaussian likelihood when the observations appear to be non-Gaussian may seem strange. However, if the noise sequence $\{Z_t\}$ in the model (12) is any independent identically distributed sequence (with finite variance), the large-sample joint distribution of the estimators

(assuming the true orders are p and q) is the same as in the Gaussian case (see [15], [5]). This large-sample distribution has a relatively simple Gaussian form that can be used to specify large-sample confidence intervals for the parameters (under the assumption that the observations are generated by the fitted model).

Maximization of L with respect to the parameter vector β is a nonlinear optimization problem, requiring the use of an efficient numerical maximization algorithm (see **Optimization Methods**). For this reason, a variety of simpler estimation methods have been developed. These generally lead to less efficient estimators, which can be used as starting points for the nonlinear optimization. Notable among these are the Hannan–Rissanen algorithm for general ARMA processes, and the Yule–Walker and Burg algorithms for purely autoregressive processes.

The previous discussion assumes that the orders p and q are *known*. However, this is rarely, if ever, the case, and they must be chosen on the basis of the observations. The choice of p and q is referred to as the problem of *order selection*. The shape of the sample autocorrelation function gives some clue as to the order of the ARMA(p, q) model that best represents the data. For example, a sample autocorrelation function that appears to be roughly of the form $\phi^{|h|}$ for some ϕ such that $|\phi| < 1$ suggests (see Example 3) that an AR(1) model might be appropriate, while a sample autocorrelation function that is small in absolute value for lags $h > q$ suggests (see Example 4) that an MA(r) model with $r \leq q$ might be appropriate.

A systematic approach to the problem was suggested by Akaike [1] when he introduced the *information criterion* known as AIC (see **Akaike's Criterion**). He proposed that p and q be chosen by minimizing $\text{AIC}(\hat{\beta}(p, q))$, where $\hat{\beta}(p, q)$ is the maximum likelihood estimator of β for fixed p and q and

$$\text{AIC}(\beta) = -2 \ln(L(\beta)) + 2(p + q + 1). \quad (18)$$

The term $2(p + q + 1)$ can be regarded as a *penalty factor* that prevents the selection of excessive values for p and q and the accumulation of additional parameter estimation errors. If the data are truly generated by an ARMA(p, q) model, it has been shown that the AIC criterion tends to overestimate p and q and is not consistent as the sample size

approaches infinity. Consistent estimation of p and q can be obtained by using information criteria with heavier penalty factors such as the (Bayesian information criterion) BIC. Since, however, there is rarely a ‘true’ model generating the data, consistency is not necessarily an essential property of order selection methods. It has been shown in [23] that although the AIC criterion does not give consistent estimation of p and q , it is optimal in a certain sense with respect to prediction of future values of the series. A refined small-sample version of AIC, known as AICC, has been developed in [18], and a comprehensive account of model selection can be found in [7].

Having arrived at a potential ARMA model for the data, the model should be checked for goodness of fit to the data. On the basis of the fitted model, the minimum mean-squared error linear predictors $\hat{Y}_t$ of each Y_t in terms of $Y_s, s < t$, and the corresponding mean-squared errors s_t can be computed (see Prediction below). In fact, they are computed in the course of evaluating the Gaussian likelihood in its innovations form. If the fitted model is valid, the properties of the rescaled one-step prediction errors $(Y_t - \hat{Y}_t)/\sqrt{s_t}$ should be similar to those of the sequence Z_t/σ in the model (12), and can, therefore, be used to check the assumed white noise properties of $\{Z_t\}$ and whether or not the assumption of independence and/or normality is justified. A number of such tests are available (see e.g., [6], chap. 5.)

Prediction

If $\{Y_t\}$ is a weakly stationary process with mean, $E(Y_t) = \mu$, and autocovariance function, $\text{Cov}(Y_{t+h}, Y_t) = \gamma(h)$, a fundamental property of conditional expectation tells us that the ‘best’ (minimum mean squared error) predictor of Y_{n+h} , $h > 0$, in terms of $Y_1, \dots, Y_n$, is the conditional expectation $E(Y_{n+h} | Y_1, \dots, Y_n)$. However, this depends in a complicated way on the joint distributions of the random variables Y_t that are virtually impossible to estimate on the basis of a single series of observations $y_1, \dots, y_n$. However, if the sequence $\{Y_t\}$ is Gaussian, the best predictor of Y_{n+h} in terms of $Y_1, \dots, Y_n$ is a linear function, and can be calculated as described below.

The best *linear* predictor of Y_{n+h} in terms of $\{1, Y_1, \dots, Y_n\}$, that is, the linear combination $P_n Y_{n+h} = a_0 + a_1 Y_1 + \dots + a_n Y_n$, which minimizes

the mean squared error, $E[(Y_{n+h} - a_0 - \dots - a_n Y_n)^2]$, is given by

$$P_n Y_{n+h} = \mu + \sum_{i=1}^n a_i (Y_{n+1-i} - \mu), \quad (19)$$

where the vector of coefficients $\mathbf{a} = (a_1, \dots, a_n)'$ satisfies the linear equation,

$$\Gamma_n \mathbf{a} = \boldsymbol{\gamma}_n(h), \quad (20)$$

with $\boldsymbol{\gamma}_n(h) = (\gamma(h), \gamma(h+1), \dots, \gamma(h+n-1))'$ and $\Gamma_n = [\gamma(i-j)]_{i,j=1}^n$. The mean-squared error of the best linear predictor is

$$E(Y_{n+h} - P_n Y_{n+h})^2 = \gamma(0) - \mathbf{a}' \boldsymbol{\gamma}_n(h). \quad (21)$$

Once a satisfactory model has been fitted to the sequence $y_1, \dots, y_n$, it is, therefore, a straightforward but possibly tedious matter to compute best linear predictors of future observations by using the mean and autocovariance function of the fitted model and solving the linear equations for the coefficients $a_0, \dots, a_n$. If n is large, then the set of linear equations for $a_0, \dots, a_n$ is large, and, so, recursive methods using the Levinson–Durbin algorithm or the Innovations algorithm have been devised to express the solution for $n = k+1$ in terms of the solution for $n = k$, and, hence, to avoid the difficulty of inverting large matrices. For details see, for example, [6], Chapter 2.

If an ARMA model is fitted to the data, the special linear structure of the ARMA process arising from the defining equations can be used to greatly simplify the calculation of the best linear h -step predictor $P_n Y_{n+h}$ and its mean-squared error, $\sigma_n^2(h)$. If the fitted model is Gaussian, we can also compute 95% prediction bounds, $P_n Y_{n+h} \pm 1.96\sigma_n(h)$. For details see, for example, [6], Chapter 5.

Example 5 In order to predict future values of the causal AR(1) process defined in Example 3, we can make use of the fact that linear prediction is a linear operation, and that $P_n Z_t = 0$ for $t > n$ to deduce that

$$\begin{aligned} P_n Y_{n+h} &= \phi P_n Y_{n+h-1} = \phi^2 P_n Y_{n+h-2} = \dots \\ &= \phi^h Y_n, \quad h \geq 1. \end{aligned}$$

In order to obtain forecasts and prediction bounds for the *original* series that were transformed to generate the residuals, we simply apply the inverse transformations to the forecasts and prediction bounds for the residuals.

The Frequency Viewpoint

The methods described so far are referred to as ‘time-domain methods’ since they focus on the evolution in time of the sequence of random variables $X_1, X_2, \dots$ representing the observed data. If, however, we regard the sequence as a random function defined on the integers, then an alternative approach is to consider the decomposition of that function into sinusoidal components, analogous to the Fourier decomposition of a deterministic function. This approach leads to the *spectral representation* of the sequence $\{X_t\}$, according to which every weakly stationary sequence has a representation

$$X_t = \int_{-\pi}^{\pi} e^{i\omega t} dZ(t), \quad (22)$$

where $\{Z(t), -\pi \leq t \leq \pi\}$ is a process with uncorrelated increments. A detailed discussion of this approach can be found, for example, in [2], [5], and [21], but is outside the scope of this article. Intuitively, however, the expression (22) can be regarded as representing the random function $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ as the limit of a linear combination of sinusoidal functions with uncorrelated random coefficients. The analysis of weakly stationary processes by means of their spectral representation is referred to as ‘frequency domain analysis’ or spectral analysis. It is equivalent to time-domain analysis, but provides an alternative way of viewing the process, which, for some applications, may be more illuminating. For example, in the design of a structure subject to a randomly fluctuating load, it is important to be aware of the presence in the loading force of a large sinusoidal component with a particular frequency in order to ensure that this is not a resonant frequency of the structure.

Multivariate Time Series

Many time series arising in practice are best analyzed as components of some vector-valued (multivariate)

time series $\{\mathbf{X}_t\} = (X_{t1}, \dots, X_{tm})'$ in which each of the component series $\{X_{ti}, t = 1, 2, 3, \dots\}$ is a univariate time series of the type already discussed. In multivariate time series modeling, the goal is to account for, and take advantage of, the dependence not only between observations of a single component at different times, but also between the different component series. For example, if X_{t1} is the daily percentage change in the closing value of the Dow-Jones Industrial Average in New York on trading day t , and if X_{t2} is the analogue for the Australian All-Ordinaries Index (see Example 2), then $\{\mathbf{X}_t\} = (X_{t1}, X_{t2})'$ is a bivariate time series in which there is very little evidence of autocorrelation in either of the two component series. However, there is strong evidence of correlation between X_{t1} and $X_{(t+1)2}$, indicating that the Dow-Jones percentage change on day t is of value in predicting the All-Ordinaries percentage change on day $t + 1$. Such dependencies in the multivariate case can be measured by the covariance matrices,

$$\Gamma(t+h, t) = [\text{cov}(X_{(t+h),i}, X_{t,j})]_{i,j=1}^m, \quad (23)$$

where the number of components, m , is two in this particular example. Weak stationarity of the multivariate series $\{\mathbf{X}_t\}$ is defined as in the univariate case to mean that all components have finite second moments and that the mean vectors $E(\mathbf{X}_t)$ and covariance matrices $\Gamma(t+h, t)$ are independent of t .

Much of the analysis of multivariate time series is analogous to that of univariate series, multivariate ARMA (or VARMA) processes being defined again by linear equations of the form (12) but with vector-valued arguments, and matrix coefficients. There are, however, some new important considerations arising in the multivariate case. One is the question of VARMA nonidentifiability. In the univariate case, an $\text{AR}(p)$ process cannot be reexpressed as a finite order moving average process. However, this is not the case for $\text{VAR}(p)$ processes. There are simple examples of $\text{VAR}(1)$ processes that are also $\text{VMA}(1)$ processes. This lack of identifiability, together with the large number of parameters in a VARMA model and the complicated shape of the likelihood surface, introduces substantial additional difficulty into maximum Gaussian likelihood estimation for VARMA processes. Restricting attention to VAR processes eliminates the identifiability problem. Moreover, the fitting of a $\text{VAR}(p)$ process by equating covariance

matrices up to lag p is asymptotically efficient and simple to implement using a multivariate version of the Levinson-Durbin algorithm (see [27], and [28]).

For nonstationary univariate time series, we discussed in the section 'Objectives' the use of differencing to transform the data to residuals suitable for modeling as zero-mean stationary series. In the multivariate case, the concept of *cointegration*, due to Granger [13], plays an important role in this connection. The m -dimensional vector time series $\{\mathbf{X}_t\}$ is said to be integrated of order d (or $I(d)$) if, when the difference operator ∇ is applied to each component $d - 1$ times, the resulting process is nonstationary, while if it is applied d times, the resulting series is stationary. The $I(d)$ process $\{\mathbf{X}_t\}$ is said to be cointegrated with cointegration vector $\boldsymbol{\alpha}$, if $\boldsymbol{\alpha}$ is an $m \times 1$ vector such that $\{\boldsymbol{\alpha}'\mathbf{X}_t\}$ is of order less than d . Cointegrated processes arise naturally in economics. [11] gives as an illustrative example the vector of tomato prices in Northern and Southern California (say X_{t1} and X_{t2} , respectively). These are linked by the fact that if one were to increase sufficiently relative to the other, the profitability of buying in one market and selling in the other would tend to drive the prices together, suggesting that although the two series separately may be nonstationary, the difference varies in a stationary manner. This corresponds to having a cointegration vector $\boldsymbol{\alpha} = (1, -1)'$. Statistical inference for multivariate models with cointegration is discussed in [10].

State-space Models

State-space models and the associated Kalman recursions have had a profound impact on time series analysis. A linear state-space model for a (possibly multivariate) time series $\{\mathbf{Y}_t, t = 1, 2, \dots\}$ consists of two equations. The first, known as the *observation equation*, expresses the w -dimensional observation vector $\mathbf{Y}_t$ as a linear function of a v -dimensional state variable $\mathbf{X}_t$ plus noise. Thus,

$$\mathbf{Y}_t = G_t \mathbf{X}_t + \mathbf{W}_t, \quad t = 1, 2, \dots, \quad (24)$$

where $\{\mathbf{W}_t\}$ is a sequence of uncorrelated random vectors with $E(\mathbf{W}_t) = \mathbf{0}$, $\text{cov}(\mathbf{W}_t) = R_t$, and $\{G_t\}$ is a sequence of $w \times v$ matrices. The second equation, called the *state equation*, determines the state $\mathbf{X}_{t+1}$ at time $t + 1$ in terms of the previous state $\mathbf{X}_t$ and a

noise term. The state equation is

$$\mathbf{X}_{t+1} = F_t \mathbf{X}_t + \mathbf{V}_t, \quad t = 1, 2, \dots, \quad (25)$$

where $\{F_t\}$ is a sequence of $v \times v$ matrices, $\{\mathbf{V}_t\}$ is a sequence of uncorrelated random vectors with $E(\mathbf{V}_t) = \mathbf{0}$, $\text{cov}(\mathbf{V}_t) = Q_t$, and $\{\mathbf{V}_t\}$ is uncorrelated with $\{\mathbf{W}_t\}$ (i.e., $E(\mathbf{W}_t \mathbf{V}_s') = 0$ for all s and t). To complete the specification, it is assumed that the initial state $\mathbf{X}_1$ is uncorrelated with all of the noise terms $\{\mathbf{V}_t\}$ and $\{\mathbf{W}_t\}$.

An extremely rich class of models for time series, including and going well beyond the ARIMA models described earlier, can be formulated within this framework (see [16]). In econometrics, the structural time series models developed in [17], in which trend and seasonal components are allowed to evolve randomly, also fall into this framework. The power of the state-space formulation of time series models depends heavily on the Kalman recursions, which allow best linear predictors and best linear estimates of various model-related variables to be computed in a routine way. Time series with missing values are also readily handled in the state-space framework.

More general state-space models, in which the linear relationships (24) and (25) are replaced by the specification of conditional distributions, are also widely used to generate an even broader class of models, including, for example, models for time series of counts such as the numbers of reported new cases of a particular disease (see e.g., [8]).

Additional Topics

Although linear models for time series data have found broad applications in many areas of the physical, biological, and behavioral sciences, there are also many areas where they have been found inadequate. Consequently, a great deal of attention has been devoted to the development of nonlinear models for such applications. These include threshold models, [25], bilinear models, [24], random-coefficient autoregressive models, [20], Markov switching models, [14], and many others. For financial time series, the ARCH (autoregressive conditionally heteroscedastic) model of Engle [9], and its generalized version, the GARCH model of Bollerslev [3], have been particularly successful in modeling financial returns data of the type illustrated in Figure 4.

Although the sample autocorrelation function of the series shown in Figure 4 is compatible with the hypothesis that the series is an independent and identically distributed white noise sequence, the sample autocorrelation functions of the absolute values and the squares of the series are significantly different from zero, contradicting the hypothesis of independence. ARCH and GARCH processes are white noise sequences that are nevertheless dependent. The dependence is introduced in such a way that these models exhibit many of the distinctive features (heavy-tailed marginal distributions and persistence of volatility) that are observed in financial time series.

The recent explosion of interest in the modeling of financial data, in particular, with a view to solving option-pricing and asset allocation problems, has led to a great deal of interest, not only in ARCH and GARCH models, but also to stochastic volatility models and to models evolving continuously in time. For a recent account of time series models specifically related to financial applications, see [26].

Apart from their importance in finance, continuous-time models provide a very useful framework for the modeling and analysis of discrete-time series with irregularly spaced data or missing observations. For such applications, see [19].

References

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, B.N. Petrov & F. Saki, eds, Akademia Kiado, Budapest, pp. 267–281.
- [2] Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*, John Wiley & Sons, New York.
- [3] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.
- [4] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis; Forecasting and Control*, 3rd Edition, Prentice-Hall, Englewood Cliffs.
- [5] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer-Verlag, New York.
- [6] Brockwell, P.J. & Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, 2nd Edition, Springer-Verlag, New York.
- [7] Burnham, K.P. & Anderson, D. (2002). *Model Selection and Multi-Model Inference*, 2nd Edition, Springer-Verlag, New York.
- [8] Chan, K.S. & Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts, *Journal American Statistical Association* **90**, 242–252.

-
- [9] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica* **50**, 987–1087.
 - [10] Engle, R.F. & Granger, C.W.J. (1987). Co-integration and error correction: representation, estimation and testing, *Econometrica* **55**, 251–276.
 - [11] Engle, R.F. & Granger, C.W.J. (1991). *Long-Run Economic Relationships*, Advanced Texts in Econometrics, Oxford University Press, Oxford.
 - [12] Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. & Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program (with discussion and reply), *The Journal of Business Economic Statistics* **16**, 127–177.
 - [13] Granger, (1981). Some properties of time series data and their use in econometric model specification, *Journal of Econometrics* **16**, 121–130.
 - [14] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, Princeton.
 - [15] Hannan, E.J. (1973). The asymptotic theory of linear time series models, *Journal of Applied Probability* **10**, 130–145.
 - [16] Hannan, E.J. & Deistler, M. (1988). *The Statistical Theory of Linear Systems*, John Wiley & Sons, New York.
 - [17] Harvey, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
 - [18] Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.
 - [19] Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics* **22**, 389–395.
 - [20] Nicholls, D.F. & Quinn, B.G. (1982). *Random Coefficient Autoregressive Models: An Introduction*, Springer Lecture Notes in Statistics, p. 11.
 - [21] Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Vols. I and II, Academic Press, New York.
 - [22] Schweppe, F.C. (1965). Evaluation of likelihood functions for Gaussian signals, *IEEE Transactions on Information Theory* **IT-11**, 61–70.
 - [23] Shibata, R. (1980). Asymptotically efficient selection of the order for estimating parameters of a linear process, *Annals of Statistics* **8**, 147–164.
 - [24] Subba-Rao, T. & Gabr, M.M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Springer Lecture Notes in Statistics, p. 24.
 - [25] Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.
 - [26] Tsay, (2001). *Analysis of Financial Time Series*, John Wiley & Sons, New York.
 - [27] Whittle, P. (1963). On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix, *Biometrika* **40**, 129–134.
 - [28] Wiggins, R.A. & Robinson, E.A. (1965). Recursive solution to the multichannel filtering problem, *Journal of Geophysics Research* **70**, 1885–1891.

(See also **Longitudinal Data Analysis; Repeated Measures Analysis of Variance**)

P.J. BROCKWELL

Tolerance and Variance Inflation Factor

JEREMY MILES

Volume 4, pp. 2055–2056

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tolerance and Variance Inflation Factor

In regression analysis (*see* **Regression Models**), one outcome is regressed onto one or more predictors in order to explain the variation in the outcome variable. Where there is more than one predictor variable, the relationships between the predictors can affect both the regression estimate and the standard error of the regression estimate (*see* **Multiple Linear Regression**).

In its original usage, multicollinearity referred to a perfect linear relationship between the independent variables, or a weighted sum of the independent variables; however, it is now used to refer to large (multiple) correlations amongst the predictor variables, known as **collinearity**.

The effect of collinearity is to increase the standard error of the regression coefficients (and hence to increase the confidence intervals and decrease the *P* values).

The standard error of a regression estimate of the variable *j* ($\hat{\beta}_j$) is given by

$$se(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{\sum x_j^2} \times \frac{1}{1 - R_j^2}} \quad (1)$$

where R_j^2 is the R^2 found when regressing all other predictors onto the predictor *j*. (Note that when there is only one variable in the regression equation, or when the correlation between the predictors is equal to zero, the value for the part of the equation $1/(1 - R_j^2)$ is equal to 1.) The term $1/(1 - R_j^2)$ is known as the *variance inflation factor* (VIF). When the correlation changes from 0 (or when additional variables are added), the value of the VIF increases, and the value of the standard error of the regression parameter increases with the square root of the VIF.

The reciprocal of the VIF is called the *tolerance*. It is equal to $1 - R_j^2$, where each predictor is regressed on all of the other predictors in the analysis.

A rule of thumb that is sometimes given for the tolerance and the VIF is that the tolerance should not be less than 0.1, and that therefore the VIF should not be greater than 10, although this is dependent on other factors, not least the sample size.

Further information on these measures can be found in [1].

Reference

- [1] Fox, J. (1991). *Regression Diagnostics*, Sage Publications, Newbury Park.

JEREMY MILES

Transformation

NEAL SCHMITT

Volume 4, pp. 2056–2058

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Transformation

A transformation is any systematic alteration in a set of scores such that some characteristics of the scores are changed while other characteristics remain unchanged. Transformations of a set of scores are done for several reasons. Statistically, transformations are done to provide a data set that possesses characteristics that make it acceptable for a variety of parametric statistical procedures. In **analysis of variance** applications, a data set that is normally distributed and that possesses homogeneity of variance across treatment conditions is necessary to draw appropriate inferences about mean differences between conditions. Normality of the distribution of a set of scores and equality in the scale units on a measure is a necessary precondition for the application of the **multivariate analyses** (e.g., **factor analyses**, **structural equation modeling**) routinely applied in the social and behavioral sciences. When data are not normally distributed, transformations that produce distributions that more closely approximate normality can be used, provided it is reasonable to assume that the underlying construct being measured is normally distributed. Transformations are also done frequently to provide data that are more easily communicated to the consuming public (e.g., percentiles, IQ scores).

The data we use in the behavioral sciences involve the assignment of numbers to objects that have some meaning in terms of the objects' physical properties. In considering the use of statistical analyses to summarize or analyze data, it is important that the transformations involved in these analyses and summaries do not alter the meaning of the basic properties to which they refer. **Stevens** [1] is credited with providing the widely accepted classification of scales into nominal, ordinal, interval, and ratio types (see **Scales of Measurement**). A nominal scale is a measurement that we use to categorize objects into discrete groups (e.g., gender, race, political party). In the case of nominal data, any one-to-one transformation that retains the categorization of individual cases into discrete groups is permissible. Typical summary statistics in this instance are the number of groups, the number of cases, and the modal category. An ordinal scale is one in which we can rank order cases such that they are greater or less than other cases on the attribute measured (e.g., class rank, pleasantness of odors). In this case, we

report the median, percentiles, and the interquartile range as summary measures and can perform any transformation that preserves the rank order of the cases being measured. An interval scale (temperature) has rank order properties but, in addition, the intervals between cases are seen as equal to each other. Appropriate summary statistics include the mean and standard deviation. Associations between variables measured on interval scales can be expressed as Pearson correlations, which are the basic unit of many multivariate analysis techniques. Any linear transformation is appropriate; the mean and standard deviation may be changed, but the rank order and relative distance between cases must be preserved. Finally, ratio scales (e.g., height, weight) include a meaningful zero point and allow the expression of the equality of ratios. Only multiplicative transformations will preserve the unique properties of a ratio scale, an absolute zero point, and the capacity to form meaningful ratios between numbers on the scale.

Examples of some commonly used transformations are provided below (see Table 1) where X might be the original scores on some scale and T₁ to T₆ represent different transformations.

In this table, the first four transformations have been accomplished by adding, subtracting, multiplying, and dividing the original numbers by 2. Each of these four transformations is a linear transformation in that the mean and standard deviation of the column of numbers is changed, but the rank order and the relative size of the intervals between units on the scale remains unchanged. Further, if one computed Pearson correlations between these four transformations and the original numbers, all would correlate 1.00. The shape of the distribution of all five sets of numbers would be identical. These transformations are said to preserve the interval nature of these numbers and are routinely part of the computation of various statistics. Note that T₁ and T₂ would be inappropriate

Table 1 Common transformations of a set of raw scores (X)

X	T ₁	T ₂	T ₃	T ₄	T ₅	T ₆
0	2	-2	0	0	0	0.00
1	3	-1	2	0.5	1	1.00
2	4	0	4	1.0	4	1.41
3	5	1	6	1.5	9	1.73
4	6	2	8	2.0	16	2.00
5	7	3	10	2.5	25	2.24

2 Transformation

with ratio data since the zero point is not preserved. The ratios between corresponding points on the transformed scale and the original scale do not have the same meaning.

The fifth transformation above was produced by taking the square of the original numbers and the sixth one was produced by taking the square root. These transformations might be used if the original distribution of a set of numbers was not normal or otherwise appropriate for statistical analyses such as analyses of variance. For example, reaction times to a stimulus might include a majority of short times and some very long ones. In this case, a square root transformation would make the data more appropriate for the use of parametric statistical computations. If a square, square root, or other nonlinear transformation is used, it is important to recognize that the units of measurement have been changed when we interpret or speak of the data. These transformations are said to be nonlinear; not only are the means and standard deviations of the transformed data different than X , but the size of the intervals between points on the scale are no longer uniform, the distribution of scores are altered, and the Pearson correlations between X and these two transformed scores would be less than 1.00. Other nonlinear transformations that are sometimes used include logarithmic and reciprocal transformations. One guide for the selection of an appropriate nonlinear transformation is to consider the ratio of the largest transformed score to the smallest transformed score and select the transformation that produces the smallest ratio. Using this 'rule of thumb', one is reducing the influence of extreme scores or outliers in an analysis, making it more likely that key assumptions of normality of the score distribution and homogeneity of variance are met.

There have been several popularly used transformations whose primary purpose was to increase the ability to communicate the meaning of data to the general public or to make data comparable across different sets of scores. One of the most widely used linear transformations is the z-score or standardized score. In this case, the mean of a set of scores is subtracted from the original or raw scores for each individual and this difference is divided by the standard deviation. Since the raw score is transformed by subtracting and dividing by a constant (i.e., the mean and standard deviation), this is a linear transformation (see examples above). Raw scores and z-scores

are correlated 1.00 and have the same distribution but different means and standard deviations. Since the means and standard deviations of all z-scores are 0 and 1 respectively, z-scores are often used to compare individuals' scores on two or more measures. It is important to recognize that the z transformation does not result in a normal distribution of scores; if the original distribution was skewed, the transformed one will also be skewed.

The standard or z-score includes decimal numbers, and half the scores (if the distribution is normal or nearly so) are negative. For this reason, it is common practice to perform a second linear transformation on the z-scores so that they have a different mean and standard deviation. Standard scores on the Graduate Record Examination, for example, are transformed by multiplying the standard score by 100 and adding 500. This provides a score whose mean and standard deviation are 500 and 100 respectively. Many other test scores are reported as 'T' scores. These are transformed z-scores that have been multiplied by 10 and to which 50 has been added. So they have means and standard deviations of 50 and 10 respectively.

One early and well-known nonlinear transformation of scores in the behavioral sciences was the computation of an intelligence quotient (IQ) by Terman [2]. The IQ was computed by taking the person's mental age as measured by the test, dividing it by the person's chronological age, and multiplying by 100. This was a nonlinear transformation of scores since persons' chronological ages were not constant. This index helped to popularize the IQ test because it produced numbers that appeared to be readily interpretable by the general public. The fact that the IQ was unusable for measured attributes that had no relationship to chronological age and the fact that mental growth tends to asymptote around age 20 doomed the original IQ transformation. Today's IQs are usually 'deviation IQs' in which a standard score for people of particular age groups is computed and then transformed as above to create a distribution of scores with the desired mean and standard deviation.

If a normal distribution of scores is desired (i.e., we hold the assumption that a variable is normally distributed and we believe that the measurement scale we are using is faulty), we can transform a set of scores to a normal distribution. This mathematical transformation is available in many statistical texts (e.g., [3]). A relatively old normalizing transformation of scores was the stanine distribution. This

transformation was done by computing the percentages of cases falling below a certain raw score and changing that raw score to its normal distribution equivalent using what is known about the normal density function and tables published at the end of most statistics texts.

Another common nonlinear transformation is the percentile scale. A percentile is the percentage of cases falling below a given raw score. Since the number of cases falling at each percentile is a constant (e.g., all percentile scores for a data set containing 1000 cases represent 10 cases), the distribution of percentiles is rectangular in shape, rather than normal or the shape of the original data. This distributional property means that it is inappropriate to use statistics that assume normally distributed data. The primary reason for computing percentiles is for public consumption since this index appears to be more easily understood and communicated to a statistically unsophisticated audience. Percentiles are the means of communicating many achievement test scores.

Transformations are useful tools both in providing a distribution of scores that is amenable to a variety of statistical analyses and in helping statisticians communicate the meaning of scores to the general public. It is, however, important to remember that we make certain assumptions about the phenomenon of interest when we transform scores and that we must return to the basic unit of measurement when we consider the practical implications of the data we observe or the manipulations of some variable(s).

References

- [1] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.
- [2] Terman, L.M. (1916). *The Measurement of Intelligence*, Houghton Mifflin, Boston.
- [3] Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, Inference, and Decision*, Holt, Rinehart, & Winston, New York.

NEAL SCHMITT

Tree Models

RICHARD DANIELS AND DAILUN SHI

Volume 4, pp. 2059–2060

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tree Models

Tree models, also known as *multinomial process tree models*, are data-analysis tools widely used in behavioral sciences to measure the contribution of different cognitive processes underlying observed data. They are developed exclusively for categorical data, with each observation belonging to exactly one of a finite set of categories. For categorical data, the most general statistical distribution is the multinomial distribution, in which observations are independent and identically distributed over categories, and each category has associated with it a parameter representing the probability that a random observation falls within that category. These probability parameters are generally expressed as functions of the statistical model's parameters, that is, they redefine the parameters of the multinomial distribution. Linear (e.g., **analysis of variance**) and nonlinear (e.g., **log-linear** and **logit**) models are routinely used for categorical data in a number of fields in the social, behavioral, and biological sciences. All that is required in these models is a suitable factorial experimental design, upon which a model can be selected without regard to the substantive nature of the paradigm being modeled.

In contrast, tree models are tailored explicitly to particular paradigms. In tree models, parameters that characterize the underlying process are often unobservable, and only the frequencies in which observed data fall into each category are known. A tree model is thus a special structure for redefining the multinomial category probabilities in terms of parameters that are designed to represent the underlying cognitive process that leads to the observed data. Tree models are formulated to permit statistical inference on the process parameters using observed data.

Tree models reflect a particular type of cognitive architecture that can be represented as a tree, that is, a graph having no cycles. In a tree that depicts the underlying cognitive process, each branch represents a different sequence of processing stages, resulting in a specific response category. From one stage to the next immediate stage in a processing sequence, one parameter is assigned to determine the link probability. The probability associated with a branch is the product of the link probabilities along that branch. Each branch must correspond to a category for which the number of observations is known; however, there can be more than one branch for a

given category. The observed response patterns can thus be considered as the final product of a number of different cognitive processes, each of which occurs with a particular probability.

A key characteristic of tree models is that category probabilities are usually nonlinear polynomial functions of the underlying process parameters (in contrast to the classical models for categorical data mentioned above, which all have linearity built in at some level). On the other hand, tree models are much less detailed than more sophisticated cognitive models like neural networks. Thus, while tree models capture some, but not all, of the important variables in a paradigm, they are necessarily approximate and incomplete, and hence are confined to particular paradigms. Despite this disadvantage, the statistical tractability of a tree model makes it an attractive alternative to standard, multipurpose statistical models.

A comprehensive review of the theory and applications of tree models is given in Batchelder and Riefer (1999) [1]. For readers interested in learning more about tree models and statistical inference, Xianguan Hu has developed an informative website at <http://irvin.psyc.memphis.edu/gpt/>.

An Example: 'Who Said What' Task

To illustrate the structure of a tree model, consider the 'Who Said What' task. Perceivers first observe a discussion that involves members of two categories (e.g., men and women). In a subsequent recognition test, subjects are shown a set of discussion statements and asked to assign each statement to its speaker. Apart from statements that occurred in the discussion (called *old statements*), new statements are also included in the assignment phase. For each statement, participants must assign Source A (male), Source B (female), or N (new statement). Figure 1 depicts a tree model for the three types of statements. Note that there are a total of 7 process parameters $\{D_1, D_2, d_1, d_2, a, b, \text{ and } g\}$, 15 branches, and 9 response categories (A, B, and N for each tree).

The model assumes that a participant first detects whether a statement is old or new with probability D_1, D_2 , or b for source A, B, or new statements, respectively. If an old statement is correctly detected as old, then d_1 and d_2 capture the capacity to correctly assign the old statement to source A and B, respectively. If the participant cannot directly

2 Tree Models

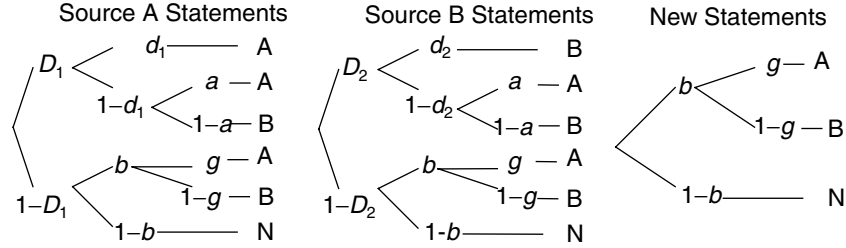


Figure 1 Tree models representing “who said what” task

attribute a statement to a source (with probability $1-d_i, i = 1, 2$), a guessing process determines the statement's source – the effectiveness of this process is measured by parameter a . If a statement is new, then another guessing process (the effectiveness of which is measured by parameter g) is used to determine the statement's source. Finally, if an old statement is not detected as old (with probability $1-D_i, i = 1, 2$), it is treated as a new statement; as such, the branches emanating from $1-D_i, i = 1, 2$, reproduce the new statement tree.

Several observations emerge from this example. First, the sequential nature of the process is based on both cognitive theory and assumptions about how statements are assigned to sources. Second, some parameters (e.g., a, g , and b) appear in more than one tree, implying, for example, that the probability of assigning a statement that is incorrectly detected as new to Source A is equal to the probability of assigning an incorrectly identified new statement to Source A. Since most of the parameters can be interpreted as conditional probabilities (i.e., conditional on the success or failure of other processes), it would perhaps be more appropriate to use different parameters to represent the same cognitive process in different trees. However, if S denotes the number of process parameters and J the number of resulting data categories, S must be no larger than $J-1$ for the model to

be statistically well defined. As a result, model realism may be traded off to gain model tractability and statistical validity.

Finally, note that the category probabilities are the sums of the products of the underlying processing parameters. For example, the probability of correctly identifying a statement from Source A is $P(A|A) = D_1d_1 + D_1(1-d_1)a + (1-D_1)bg$. Similarly, the probability that a random observation falls into each of the other eight categories can be expressed as a function of the seven process parameters ($D_1, D_2, d_1, d_2, a, b, g$). As such, the objective of tree modeling is to draw statistical inference on the process parameters using the sample frequencies of observations that fall into each data category, thus providing insight into the unknown cognitive processes.

Reference

- [1] Batchelder, W.H. & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling, *Psychonomic Bulletin & Review* 6(1), 57–86.

RICHARD DANIELS AND DAILUN SHI

Trellis Graphics

BRIAN S. EVERITT

Volume 4, pp. 2060–2063

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Trellis Graphics

Suppose in an investigation of crime in the USA we are interested in the relationship between crime rates in different states and the proportion of young males in the state, and whether this relationship differs between southern states and the others. To inspect the relationship graphically we might plot two graphs; the first a **scatterplot** of crime rate against proportion of young males for the southern states and the second the corresponding scatterplot for the rest. Such a diagram is shown in Figure 1 based on data given in [4].

Figure 1 is a simple example of a general scheme for examining high-dimensional structure in data by means of conditional one-, two- and three-dimensional graphs, introduced in [2]. The essential feature of such trellis displays (or casement displays as they sometimes called *see Scatterplot Matrices*) is the multiple conditioning that allows some type of graphic for two or more variables to be plotted for different values of a given variable (or variables). In Figure 1, for example, a simple scatterplot for two variables is shown conditional on the values of a

third, in this case categorical, variable. The aim of trellis graphics is to help in understanding both the structure of the data and how well proposed models for the data actually fit.

Some Examples of Trellis Graphics

Blood Glucose Levels

Crowder and Hand [3] report an experiment in which blood glucose levels are recorded for six volunteers before and after they had eaten a test meal. Recordings were made at times –15,0,30,60,90,120,180,240, 300 and 360 min after feeding time. The whole process was repeated six times, with the meal taken at various times of the day and night. A trellis display of the relationship between glucose level and time for each subject, conditioned on the time a meal was taken is shown in Figure 2. There are clear differences in the way glucose level changes over time between the different meal times.

Married Couples

In [4] a set of data that give the heights and ages of both couples in a sample of married couples

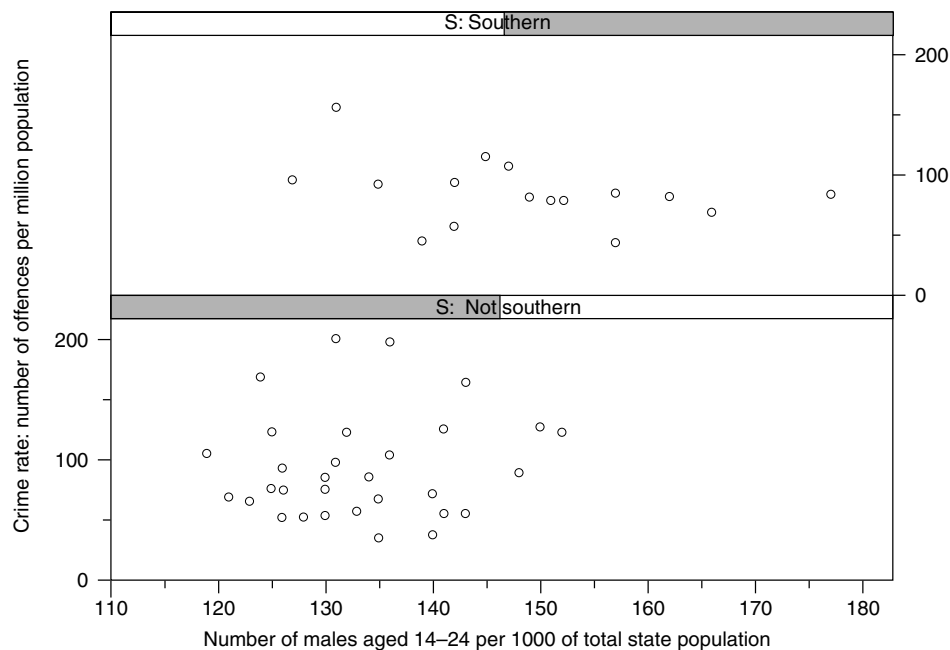


Figure 1 Scatterplots of crime rate against number of young males for southern states and other states

2 Trellis Graphics

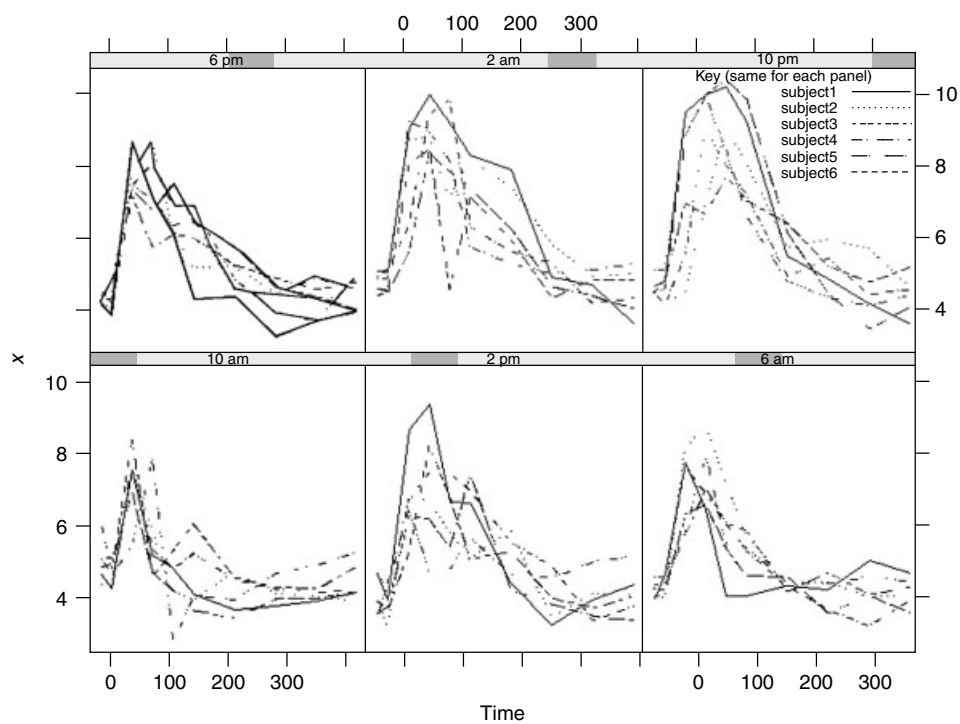


Figure 2 Trellis graphic of glucose level against time conditioned on time of eating test meal

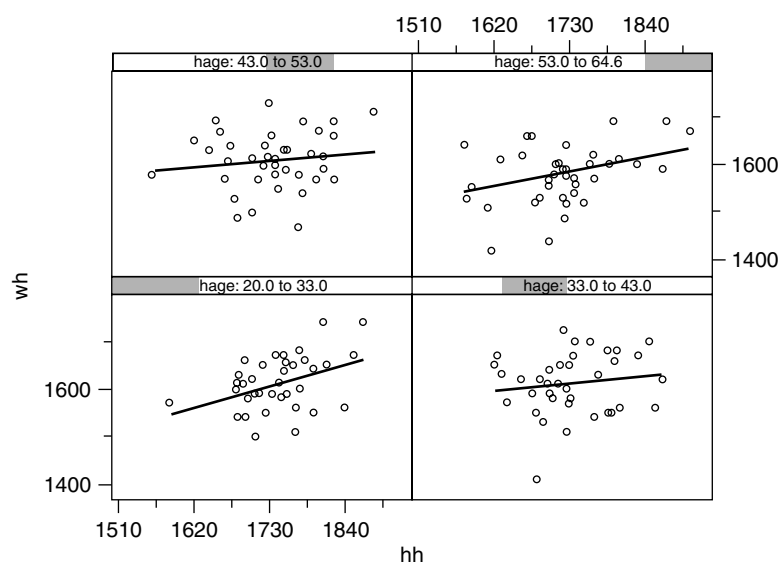


Figure 3 Trellis graphic of height of wife against height of husband conditioned on age of husband

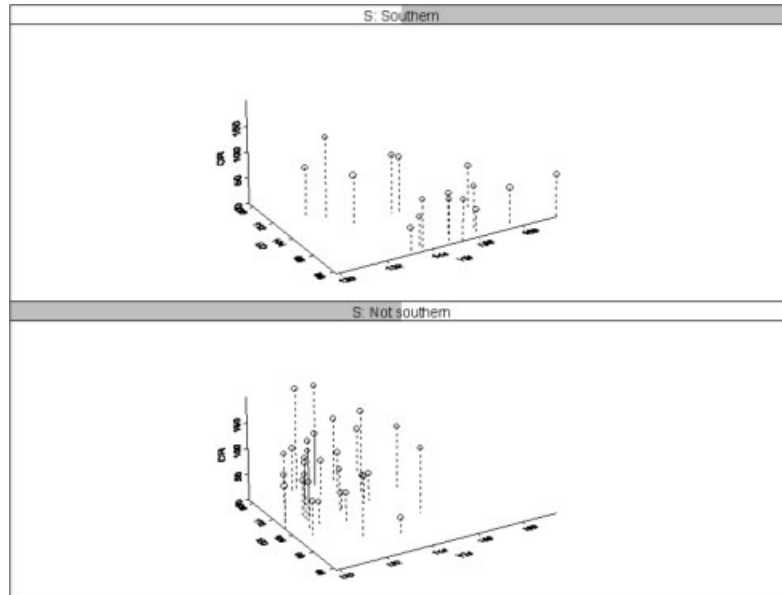


Figure 4 Three-dimensional drop-line plots for crime in southern and non-southern states

is presented. Figure 3 shows a scatterplot of the height of the wife against the height of the husband conditioned on four intervals of the age of the husband. In each of the four panels the fitted least squares regression line is shown. There is some suggestion in this diagram that amongst the youngest and the oldest husbands there is a stronger tendency for taller men to marry taller women than in the two intermediate age groups.

Crime in the USA

Finally we can return to the data set used for Figure 1 to illustrate a further trellis graphic (see Figure 4). Here a three-dimensional 'drop-line' plot is constructed for southern and non-southern states. The variables are CR-crime rate as defined in Figure 1, YM-number of young men as defined in Figure 1, and ED-educational level given by the mean number of years of schooling $\times 10$ of the population 25 years old and over.

Some other examples of trellis graphics are given in [5].

Trellis graphics are available in **S-PLUS** as described in [1].

References

- [1] Becker, R.A. & Cleveland, W.S. (1994). *S-PLUS Trellis Graphics User's Manual Version 3.3*, Insightful, Seattle.
- [2] Cleveland, W. (1993). *Visualizing Data*, Hobart Press, Summit.
- [3] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*, CRC/Chapman & Hall, London.
- [4] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1993). *Small Data Sets*, CRC/Chapman & Hall, London.
- [5] Verbyla, A.P., Cullis, B.R., Kenward, M.G. & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data using smoothing splines, *Applied Statistics* **48**, 269–312.

BRIAN S. EVERITT

Trend Tests for Counts and Proportions

LI LIU, VANCE W. BERGER AND SCOTT L. HERSHBERGER

Volume 4, pp. 2063–2066

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Trend Tests for Counts and Proportions

Cochran–Armitage Trend Test

In an R by 2 or 2 by C **contingency table**, where one variable is binary and the other is ordinal, the Cochran–Armitage test for trend can be used to test the trend in the contingency table. The binary variable can represent the response, and the ordinal variable can represent an explanatory variable with ordered levels. In an R by 2 table, the trend test can test whether the proportion increases or decreases along the row variable. In a 2 by C table, the trend test can test whether the proportion increases or decreases along the column variable.

For an R by 2 table, the Cochran–Armitage trend statistic [2, 4] is defined as

$$Z^2 = \frac{\left(\sum_{i=1}^R n_{i1}(x_i - \bar{x}) \right)^2}{p_{+1}p_{+2} \sum_{i=1}^R n_{i+}(x_i - \bar{x})^2}, \quad (1)$$

where n_{i1} is the count of response 1 (column 1) for the i th row, n_{i+} is the sum count of the i th row, p_{+1} is the sample proportion of response 1 (column 1), p_{+2} is the sample proportion of response 2 (column 2), x_i is the score assigned to the i th row, and $\bar{x} = \sum_{i=1}^R n_{i+}x_i/n$.

The statistic Z^2 has an asymptotic chi-squared distribution with 1 degree of freedom (*see Catalogue of Probability Density Functions*). The null hypothesis is that the proportion $p_{i1} = n_{i1}/n_{i+}$ is the same across all levels of the exploratory variable. The alternative hypothesis is that the proportion either increases monotonically or decreases monotonically along the exploratory variable. If we are interested in the direction of the trend, then the statistic Z can be used, which has an asymptotically standard normal distribution under the null hypothesis.

A simple score selection [6] is to use the corresponding row number as the score for that row. Other score selection can be used based on the specific problem.

The trend test is based on the linear probability model, where the response is the **binomial proportion**, and the exploratory variable is the score of each level of the ordinal variable. Let π_{i1} denote the probability of response 1 (column 1), and p_{i1} denote the sample proportion for $i = 1, \dots, R$. We have

$$\pi_{i1} = \alpha + \beta(x_i - \bar{x}). \quad (2)$$

The weighted least squares regression (*see Least Squares Estimation*) gives the estimate of α and β , which are

$$\begin{aligned} \hat{\alpha} &= p_{+1}, \\ \hat{\beta} &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - p_{+1})(x_i - \bar{x})}{\sum_{i=1}^R n_{i+}(x_i - \bar{x})^2}. \end{aligned} \quad (3)$$

The Pearson statistic for testing independence for an R by 2 table is

$$\begin{aligned} \chi^2 &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - \hat{\pi}_{i1} + \hat{\pi}_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - \hat{\pi}_{i1})^2 + \sum_{i=1}^R n_{i+}(\hat{\pi}_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - \hat{\pi}_{i1})^2}{p_{+1}p_{+2}} + \frac{\sum_{i=1}^R n_{i+}(\hat{\pi}_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^R n_{i+}(p_{i1} - \hat{\pi}_{i1})^2}{p_{+1}p_{+2}} + \frac{\sum_{i=1}^R n_{i+}\hat{\beta}^2(x_i - \bar{x})^2}{p_{+1}p_{+2}} \end{aligned}$$

2 Trend Tests for Counts and Proportions

$$= \frac{\sum_{i=1}^R n_{i+} (p_{i1} - \hat{\pi}_{i1})^2}{p_{+1} p_{+2}} + \frac{\left(\sum_{i=1}^R n_{i1} (x_i - \bar{x}) \right)^2}{p_{+1} p_{+2} \sum_{i=1}^R n_{i+} (x_i - \bar{x})^2}. \quad (4)$$

Basically, we can decompose the Pearson Statistics into two parts. The first part tests the goodness of fit of the linear model, and the second part is the Cochran–Armitage Statistic for testing a linear trend in the proportions [1].

Exact Cochran–Armitage Test for Trend

An Exact test for trend can be used if the asymptotic assumptions are not met (*see Exact Methods for Categorical Data*). For example, the sample size may not be large enough, or the data distribution may be sparse, skewed, and so on. Exact test for trend is based on the exact conditional method for contingency tables. Conditional on the row totals and column totals, to test independence against a trend, the exact P value is the sum of the probabilities for those tables having a test statistic larger than or equal to the observed test statistic Z^2 . In practice, the sufficient statistic, $T = \sum x_i n_{i1}$, can be used as a test statistic to compute the exact P value.

Jonckheere–Terpstra Trend Test

In an R by C contingency table, where the column variable represents an ordinal response, and the row variable can be nominal or ordinal, sometimes we are interested in testing whether the ordered response follows either an increasing or decreasing trend across the rows. For example, following the omnibus **Kruskal–Wallis** test for differences among doses of a sleeping medication, we might want to determine whether the proportion of subjects who fall asleep within 30 minutes increases as the dose increases. The **Jonckheere–Terpstra trend** test is designed to test the null hypothesis that the distribution of the response variable is the same across the rows [9]. The alternative hypothesis is that

$$s_1 \leq s_2 \leq \dots \leq s_R \text{ or } s_1 \geq s_2 \geq \dots \geq s_R, \quad (5)$$

with at least one of the equalities being strict, where s_i represents the i th row effect. Unlike the Cochran–Armitage test, the inequality tested by the Jonckheere–Terpstra test is not necessarily linear. The Jonckheere–Terpstra trend test was proposed independently by Jonckheere [7] and Terpstra [10], and it is a nonparametric test based on the sum of the Mann–Whitney–Wilcoxon (*see Wilcoxon–Mann–Whitney Test*) statistic M . To compare row i and row i' , we have

$$M_{ii'} = \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} I(n_{i'j'} - n_{ij}), \quad (6)$$

where $I(x)$ is equal to 0, $1/2$, 1 for $x < 0$, $x = 0$, and $x > 0$ respectively. Then the Jonckheere–Terpstra trend test statistic is

$$J = \sum_{i=1}^{R-1} \sum_{i'=i+1}^R M_{ii'}. \quad (7)$$

Under the null hypothesis that there is no difference among the rows, the standardized test

$$J^* = \frac{J - u_J}{\sigma_J} \quad (8)$$

is asymptotically distributed as a standard normal variable, where u_J is the expected mean and σ_J is the expected standard deviation under the null. Here

$$u_J = \frac{\left(n^2 - \sum_{i=1}^R n_{i+}^2 \right)}{4}, \quad (9)$$

and the variance is

$$\sigma_J^2 = \frac{1}{72} \left(n^2(2n+3) - \sum_{i=1}^R [n_{i+}^2(2n_{i+}+3)] \right). \quad (10)$$

The Jonckheere–Terpstra test is generally more powerful than the Kruskal–Wallis test, and should be used instead if there is specific interest in a trend in the data.

A modified variance adjusting for the tied values can also be used [8]. We calculate

$$\sigma_J^{*2} = \frac{J_1}{d_1} + \frac{J_2}{d_2} + \frac{J_3}{d_3}, \quad (11)$$

where

$$\begin{aligned}
 J_1 &= n(n-1)(2n+5), \\
 &\quad - \sum_{i=1}^R n_{i+}(n_{i+}-1)(2n_{i+}+5), \\
 &\quad - \sum_{j=1}^C n_{+j}(n_{+j}-1)(2n_{+j}+5), \\
 J_2 &= \left(\sum_{i=1}^R n_{i+}(n_{i+}-1)(n_{i+}-2) \right), \\
 &\quad \times \left(\sum_{j=1}^C n_{+j}(n_{+j}-1)(n_{+j}-2) \right), \\
 J_3 &= \left(\sum_{i=1}^R n_{i+}(n_{i+}-1) \right) \left(\sum_{j=1}^C n_{+j}(n_{+j}-1) \right) \\
 d_1 &= 72, \quad d_2 = 36n(n-1)(n-2), \quad d_3 = 8n(n-1).
 \end{aligned}
 \tag{12}$$

When there are no ties, all $n_{i+} = 1$ and $J_2 = J_3 = 0$, resulting in $\sigma_j^{*2} = \sigma_j^2$.

The asymptotic Jonckheere–Terpstra test is also equivalent to Kendall’s tau.

We can also compute the exact Jonckheere–Terpstra trend test, which is a **permutation test** [3], requiring computation of the test statistic for all permutations of a contingency table.

Example

This example illustrates the use of the Cochran–Armitage trend test. Table 1 is a data set from [5]. This is a retrospective study of the lung cancer and tobacco smoking among patients in hospitals in several English cities. One question of interest is whether subjects with higher numbers of cigarettes daily are more likely to have lung cancer. We use the equal interval scores {1, 2, 3, 4, 5, 6}. The trend test statistic $Z^2 = 129$ for $df = 1$, and the P value is less than 0.0001. This indicates that there is a strong linear trend along the row variable, the daily average number of cigarettes.

Alternatively, we could compute the Jonckheere–Terpstra test. For the asymptotic test, $J = 1090781$, $p < .0001$, and $z = -10.59$; whereas for the exact

Table 1 Retrospective study of lung cancer and tobacco smoking. Reproduced from Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286 [5]

Daily average Number of cigarettes	Disease group	
	Lung cancer patients	Control patient
None	7	61
<5	55	129
5–14	489	570
15–24	475	431
25–49	293	154
50+	38	12

test, evidence for the inequality of the patient and control distributions is even stronger, $p < .000001$.

Both the Cochran–Armitage and Jonckheere–Terpstra tests result in the conclusion that the proportion of lung cancer increases as the daily average number of cigarettes increases.

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.
- [2] Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics* **11**, 375–386.
- [3] Berger, V.W. (2000). Pros and Cons of Permutation Tests in Clinical Trials, *Statistics in Medicine* **19**, 1319–1328.
- [4] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.
- [5] Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286.
- [6] Ivanova, A. & Berger, V.W. (2001). Drawbacks to Integer Scoring for Ordered Categorical Data, *Biometrics* **57**, 567–570.
- [7] Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**, 133–145.
- [8] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods based on Ranks*, Holden-Day, San Francisco.
- [9] SAS Institute Inc. (1999). *SAS/STAT User’s Guide*, Version 8, SAS Institute Cary.
- [10] Terpstra, T.J. (1952). The asymptotic normality and consistency of Kendall’s test against trend, when ties are present in one ranking, *Indagationes Mathematicae* **14**, 327–333.

LI LIU, VANCE W. BERGER AND SCOTT
L. HERSHBERGER

Trimmed Means

RAND R. WILCOX

Volume 4, pp. 2066–2067

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Trimmed Means

A trimmed mean is computed by removing a proportion of the largest and smallest observations and averaging the values that remain. Included as a special case are the usual sample mean (no trimming) and the median. As a simple illustration, consider the 11 values 6, 2, 10, 14, 9, 8, 22, 15, 13, 82, and 11. To compute a 10% trimmed mean, multiply the sample size by 0.1 and round the result down to the nearest integer. In the example, this yields $g = 1$. Then, remove the g smallest values, as well as the g largest, and average the values that remain. In the illustration, this yields 12. In contrast, the sample mean is 17.45. To compute a 20% trimmed mean, proceed as before; only, now g is 0.2 times the sample sizes rounded down to the nearest integer. Some researchers have considered a more general type of trimmed mean [2], but the description just given is the one most commonly used.

Why trim observations, and if one does trim, why not use the **median**? Consider the goal of achieving a relatively low standard error. Under normality, the optimal amount of trimming is zero. That is, use the untrimmed mean. But under very small departures from normality, the mean is no longer optimal and can perform rather poorly (e.g., [1], [3], [4], [8]). As we move toward situations in which **outliers** are common, the median will have a smaller standard error than the mean, but under normality, the median's standard error is relatively high. So, the idea behind trimmed means is to use a compromise amount of trimming with the goal of achieving a relatively small standard error under both normal and nonnormal distributions. (For an alternative approach, see **M Estimators of Location**). Trimming observations with the goal of obtaining a more accurate estimator might seem counterintuitive, but this result has been known for over two centuries. For a non-technical explanation, see [6].

Another motivation for trimming arises when sampling from a skewed distribution and testing some

hypothesis. **Skewness** adversely affects control over the probability of a type I error and **power** when using methods based on means (e.g., [5], [7]). As the amount of trimming increases, these problems are reduced, but if too much trimming is used, power can be low. So, in particular, using a median to deal with a skewed distribution might make it less likely to reject when in fact the null hypothesis is false.

Testing hypotheses on the basis of trimmed means is possible, but theoretically sound methods are not immediately obvious. These issues are easily addressed, however, and easy-to-use software is available as well, some of which is free [7, 8].

References

- [1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.
- [2] Hogg, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory, *Journal of the American Statistical Association* **69**, 909–922.
- [3] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [4] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [5] Westfall, P.H. & Young, S.S. (1993). *Resampling Based Multiple Testing*, Wiley, New York.
- [6] Wilcox, R.R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy*, Springer, New York.
- [7] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.
- [8] Wilcox, R.R. (2004). (in press). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

Further Reading

- Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in *Contributions to Probability and Statistics*, I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann, eds, Stanford University Press, Stanford.

RAND R. WILCOX

T-Scores

DAVID CLARK-CARTER

Volume 4, pp. 2067–2067

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

T-Scores

T -scores are standardized scores. However, unlike z -scores, which are standardized to have a mean of 0 and an SD of 1, the T -scoring system (due to McCall [1]) produces a distribution of new scores with a mean of 50 and an SD of 10. The easiest way of calculating a T -score is to find the z -score first and then apply the following simple linear transformation:

$$T = z \times 10 + 50. \quad (1)$$

T -scoring gives the same benefits as any other standardizing system in that, for instance, it makes possible direct comparisons of a person's scores over different, similarly scored, tests. However, it has the additional advantage over z -scores of producing new

scores that are easier to interpret since they are always positive and expressed as whole numbers.

If the original scores come from a normal population with a known mean and SD, the resulting T -scores will be normally distributed with mean 50 and SD of 10. Thus, we can convert these back to z -scores and then use standard normal tables to find the percentile point for a given T -score.

T -scoring is the scoring system for several test instruments commonly used in psychology, such as the MMPI.

Reference

- [1] McCall, W.A. (1939). *Measurement*, Macmillan Publishers, New York.

DAVID CLARK-CARTER

Tukey, John Wilder

SANDY LOVIE

Volume 4, pp. 2067–2069

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tukey, John Wilder

Born: June 16, 1915, in Massachusetts, USA.

Died: July 26, 2000, in New Jersey, USA.

It is almost impossible to be a true polymath in modern science, if only because of the intense pressure to specialize as early as possible. John Tukey was one of the few exceptions to this rule in that he made fundamental contributions to many distinct areas of science including statistics, computing, and mathematics, where the latter meant his early love, topology. What seems to characterize all his work is an unwillingness to accept the accepted; thus, his most extensively articulated area of work, **Exploratory Data Analysis** (or EDA), sidesteps twentieth century statistical obsessions with inference to concentrate on extracting new meanings from data. In other words, EDA is perhaps best seen as an attempt to recreate an *inductive* approach to statistics, which may or may not have actually existed in the past. Certainly, Tukey's massive search for novel measures of location in place of the mean using large scale computer simulations seems to reflect his wish to both query the foundations of statistics and to make foundational changes in the discipline [1]. Such a work also mirrors the long running historical debates in the eighteenth and nineteenth centuries over the definition and role of the mean in statistics, where these were most decidedly not the unproblematic issues that they appear today (see [4], [5]). Tukey's invention of novel graphical methods can also be viewed in something like the same light, that is, as both an undermining throwback to the nineteenth century when plots and other displays were central tools for the statistician, and as solutions to modern problems making use of modern technology such as computers, visual displays, and plotters.

Tukey's early life was somewhat isolated and protected in that he was an only child with most of his education coming from his mother, who acted as a private tutor, since, being married, she was unable to practice her original training as a teacher. Tukey's father was also a school teacher, whose *metier* was Latin. Tukey's early degrees were from Brown University (bachelors and masters in chemistry, 1936 and 1937). He then moved to Princeton with the initial aim of obtaining a PhD in chemistry, but saw the

error of his ways and switched to mathematics (see the transcript of his early Princeton reminiscence's in [6]). After finishing his graduate work in topology and obtaining his doctorate in 1939, he was appointed to an assistant professorship in the Mathematics Department in Princeton and full professorship in 1950 at the age of 35. Meanwhile, he had discovered statistics when working during the war for Fire Control Research based in Princeton from 1941 to 1945. Other important workers attached to this unit were Tukey's statistical mentor Charles Winsor, Albert Tucker, and **Claude Shannon**. After the war, he alternated between Princeton and the Bell Laboratories at Murray Hill, New Jersey. The earliest work of note is his influential text with Blackman on the analysis of power spectra, which appeared in 1959. But while still maintaining an interest in this field, including the use of computers in the approximation of complex functions (his well regarded paper on fast Fourier transforms, for example, had been published jointly with Cooley in 1964), Tukey had, nevertheless, begun to move into EDA via the study of robustness and regression residuals (see **Robust Testing Procedures; Robustness of Standard Tests; Regression Models**). The public side of this finally emerged in 1977 in the two volumes on EDA [7], and EDA and regression [3], the latter written jointly with Frederick Mosteller.

The initial reaction to these books was strongly positive on many statisticians part, but there were equally strong negative reactions as well. The British statistician Ehrenberg, for example, is quoted as saying that if he had not known who the author was he would have dismissed EDA as a joke. Happily EDA survived and flourished, particularly in the resuscitation and redirecting of regression analysis from a rather fusty nineteenth century area of application of least squares methods into a dynamic and multifaceted technique for the robust exploration of complex data sets. Further, a great many of the displays pioneered by Tukey are now staple fodder in just about all the statistics packages that one can point to, from the hand-holding ones like SPSS and Minitab, to the write-your-own-algorithm environments of S, S-Plus and R. Tukey's defense of this apparent fuzziness is well known, arguing as he did on many occasions that an approximate answer to the correct question is always preferable to a precise one to an incorrect one. Tukey in his long and fruitful career has also inspired generations of students and

coworkers, as may be seen from the published volume of collaborative work. In addition, his collected works now run to nine volumes, although I suspect that this is not the final count; while the latest book of his that I can find is one written jointly with Kaye Basford on the graphical analysis of some classic plant breeding trials [2], which he published a year before his death at the age of 84!

References

- [1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton.
- [2] Basford, K.E. & Tukey, J.W. (1999). *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*, Chapman & Hall, London.
- [3] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading.
- [4] Porter, T.M. (1986). *The Rise of Statistical Thinking 1820–1900*, Princeton University Press, Princeton.
- [5] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Harvard.
- [6] Tucker, A. (1985). *The Princeton Mathematics Community in the 1930s: John Tukey*, Princeton University, Transcript No. 41 (PMC41).
- [7] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

SANDY LOVIE

Tukey Quick Test

SHLOMO SAWILOWSKY

Volume 4, pp. 2069–2070

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tukey Quick Test

The Tukey procedure [4] is a two-independent-samples test of differences in location. It is an alternative to the parametric t Test. Despite its simplicity, Neave and Worthington [3] noted that it is ‘an entirely valid and distribution-free test’. It has the benefit of easily memorized critical values.

Procedure

The first step is to identify the maximum score in the sample (A_{Max}) with the smaller median and also the minimum score in sample (B_{Min}) with the larger median. The second step is to count the number of scores in sample B that are lower than A_{Max} and also the number of scores in sample A that are higher than B_{Min} .

Hypotheses

The null hypothesis, H_0 , is that there is no difference in the two population medians or that the two samples were drawn from the same population. The alternative hypothesis, H_1 , is that the populations sampled have different medians or the samples originate from different populations.

Assumptions

Tukey’s test assumes there are no tied values, especially at the extreme ends. Neave and Worthington [3] suggested an iterative method of breaking ties in all possible directions, computing the test statistic T for each iteration, and making the decision based on the average value of T . Monte Carlo results by Fay and Sawilowsky [2] and Fay [1] indicated that a simpler procedure for resolving tied values is to randomly assign tied values for or against the null hypothesis.

Test Statistic

The test statistic, T , is the sum of the two counts described above. Critical values for the two-tailed test

Table 1 Sample data

Sample A	Sample B
201	334
333	418
335	419
340	442
420	469
	417

are easily remembered. As long as the ratio of the two sample sizes is less than 1.5, the critical values for $\alpha = 0.05$ and 0.01 are 7 and 10, respectively. Critical values for the one-tailed test are 6 and 9. Additional tabled critical values appear in [3].

Example

Consider the data from two samples in the table below (Table 1).

$A_{\text{Max}} = 420$ from sample A , and $B_{\text{Min}} = 334$ from sample B . Four scores in sample B are lower than A_{Max} (334, 418, 419, and 417). Three scores in Group A are higher than B_{Min} (335, 340, and 420). The test statistic is $T = 3 + 4 = 7$, which is significant at $\alpha = 0.050$. Thus, the null hypothesis of no difference in location is rejected in favor of the alternative that Sample B has the larger median.

References

- [1] Fay, B.R. (2003). A Monte Carlo computer study of the power properties of six distribution-free and/or nonparametric statistical tests under various methods of resolving tied ranks when applied to normal and nonnormal data distributions, Unpublished doctoral dissertation, Wayne State University, Detroit.
- [2] Fay, B.R. & Sawilowsky, S. (2004). The effect on type I error and power of various methods of resolving ties for six distribution-free tests of location. Manuscript under review.
- [3] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.
- [4] Tukey, J.W. (1959). A quick, compact, two-sample test to Duckworth’s specifications, *Technometrics* **1**, 31–48.

(See also **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY

Tversky, Amos

SANDY LOVIE

Volume 4, pp. 2070–2071

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Tversky, Amos

Born: March 16, 1937, in Haifa.

Died: June 2, 1996, in California.

There are two, somewhat different, Amos Tversky's: one is the high profile co-inventor (with Daniel Kahneman) of cognitive biases and heuristics, the other the considerably more retiring mathematical psychologist heavily involved with the development of polynomial conjoint measurement, the recent Great White Hope of psychological measurement [3]. Although Tversky will almost certainly be remembered for his joint attack on the statistical complacency of a generation of psychologists, and the ramifications of both the Law of Small Numbers and a raft of related phenomena, the more rigorous side of his contribution to mathematical psychology should not be overlooked.

Tversky's early education was at the Hebrew University of Jerusalem after a year of army service as a paratrooper, when he had been awarded a military decoration for rescuing a fellow soldier who had got into trouble laying an explosive charge. He also served his country in 1967 and 1973. After graduating from the University in 1961 with a BA in philosophy and psychology, Tversky had immediately moved to the University of Michigan in America to take a PhD with Ward Edwards, then the leading behavioral decision theorist. He was also taught mathematical psychology at Michigan by **Clyde Coombs** whose deep interest in scaling theory had rubbed off on Tversky. On obtaining his PhD in 1965, he had then travelled to the east coast for a year's work at the Harvard Centre for Cognitive Studies. He returned to the Hebrew University in 1966 where he was made full professor in 1972. Interspersed with his time in Israel was a stretch from 1970 as a fellow at Stanford University's Centre for the Advanced Study of the Behavioral Sciences. In 1978, he finally joined the Psychology Department at Stanford where he remained until his death.

In 1969, Tversky had been invited by Daniel Kahneman, his near contemporary at the Hebrew University, to give a series of lectures to the University on the human assessment of probability. From

this early collaboration came experimental demonstration of the Law of Small Numbers, and the rest, as they say, is history. Kahneman and Tversky presented a set of deceptively simple stories recounting statistical problems to a series of mathematically and statistically sophisticated audiences. The results showed the poor quality of the statistical intuitions of these groups, leading Kahneman and Tversky to invent an ever expanding set of biases and cognitive short cuts (heuristics), including the best known ones of *representativeness* (the small mirrors the large in all essential elements), and *availability* (the probability that people assign to events is a direct function of how easily they are generated or can be retrieved from memory). Others include the conjunction fallacy, regression to the mean, anchoring and adjustment, and the simulation heuristic, where the latter has triggered a veritable explosion of research into what is termed 'counterfactual' reasoning or thinking, that is, 'but what if...' reasoning (*see Counterfactual Reasoning*). The movement also attracted additional studies outside the immediate Tversky–Kahneman, axis, for example, the work of Fischhoff on the hindsight bias ('I always knew it would happen'), while the notion of biases and heuristics seemed to offer a framework for other related studies, such as those on illusory correlation, vividness, and human probability calibration. Indeed, Tversky's later economics-orientated Prospect theory threatened to account for most behavioral studies of risk and gambling! This approach also generated the notion of decision frames, that is, the idea that all choices are made within a context, or, put another way, that people's risky decision making can only be understood if you also understand the setting in which it takes place (classic accounts of the material can be found in [2], with extensive updates contained in [1]). Not too surprisingly, the theme of heuristics and biases has also been taken up enthusiastically by cognitive social psychologists (*see [4] for the initial reaction*) (*see Decision Making Strategies*).

Kahneman was (jointly) awarded the Nobel Prize for Economics in 2003 for his work on what is now termed *Behavioral Economics*. Unfortunately, since the Prize cannot be awarded posthumously, Tversky missed being honoured for his seminal contribution to this new and exciting area.

References

- [1] Gilovich, T., Griffin, D. & Kahneman, D., eds (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge.
- [2] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
- [3] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A., eds (1971). *Foundations of Measurement*, Vol. 1, Academic Press, New York.
- [4] Nisbett, R.E. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*, Prentice-Hall, Englewood Cliffs.

SANDY LOVIE

Twin Designs

FRANK M. SPINATH

Volume 4, pp. 2071–2074

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Twin Designs

Introduction

The classical twin study compares the phenotypic resemblances of identical or *monozygotic* (MZ) and fraternal or *dizygotic* (DZ) twins. MZ twins derive from the splitting of one fertilized zygote and therefore inherit identical genetic material. DZ twins are first-degree relatives because they develop from separately fertilized eggs and are 50% genetically identical on average. It follows that a greater within-pair similarity in MZ compared to DZ twins suggests that genetic variance influences the trait under study.

The discovery of the twin method is usually ascribed to **Galton** [9] although it is uncertain whether Galton was aware of the distinction between MZ and DZ twins [22]. It was not until almost 50 years later that explicit descriptions of the classical twin method were published [14, 25].

Terminology

To disentangle and to quantify the contributions that genes and the environment make to human complex traits, data are required either from relatives who are genetically related but who grow up in unrelated environments (*'twin adoption design'* (see **Adoption Studies**)), or from relatives who grow up in similar environments but are of differing genetic relatedness (*'twin design'*) [1]. Most twin studies that have been conducted over the past 80 years are of the latter type. Only two major studies of the former type have been conducted, one in Minnesota [2] and one in Sweden [17]. These studies have found, for example, that monozygotic twins reared apart from early in life are almost as similar in terms of general cognitive ability as are monozygotic twins reared together, a result suggesting strong genetic influence and little environmental influence caused by growing up together in the same family. These influences are typically called (see **Shared Environment**) because they refer to environmental factors contributing to the resemblance between individuals who grow up together [20]. **Nonshared environmental** influences, on the other hand, refer to environmental factors that make individuals who grow up together different from one another.

Twinning

One reason why a predominant number of twin studies have utilized the twin design instead of the twin adoption design is that twins typically grow up together, thus it is much easier to find a large number of participants for the classic twin study. In humans, about 1 in 85 live births are twins. The numbers of identical and same-sex fraternal twins are approximately equal. That is, of all twin pairs, about one third are identical twins, one third are same-sex fraternal twins, and one third are opposite-sex fraternal twins. The rate of twinning differs across countries, increases with maternal age, and may even be inherited in some families. Greater numbers of fraternal twins are the result of the increased use of fertility drugs and *in vitro* fertilization, whereas the rate of identical twinning is not affected by these factors [20].

Zygosity Determination

The best way to determine twin zygosity is by means of DNA markers (polymorphisms in DNA itself). If a pair of twins differs for any DNA marker, they must be fraternal because identical twins are identical genetically. If a reasonable number of markers are examined and no differences are found, it can be concluded that the twin pair is identical. Physical similarity on highly heritable traits such as eye color, hair color, or hair texture as well as reports about twin confusion are also often used for zygosity determination. If twins are highly similar for a number of physical traits, they are likely to be identical. Using physical similarity to determine twin zygosity typically yields accuracy of more than 90% when compared to genotyping data from DNA markers (e.g., [5]).

Deriving Heritability and Environmental Estimates from Twin Correlations

Comparing the phenotypic (see **Genotype**) resemblance of MZ and DZ twins for a trait or measure under study offers a first estimate of the extent to which genetic variance is associated with phenotypic variation of that trait. If MZ twins resemble each other to a greater extent than do DZ twins, the **heritability** (h^2) of the trait can be estimated by doubling

the difference between MZ and DZ correlations, that is, $h^2 = 2(r_{MZ} - r_{DZ})$ [7]. *Heritability* is defined as the proportion of phenotypic differences among individuals that can be attributed to genetic differences in a particular population. Whereas *broad-sense* heritability involves all additive and nonadditive sources of genetic variance, *narrow-sense* heritability is limited to additive genetic variance. The proportion of the variance that is due to the shared environment (c^2) can be derived from calculating $c^2 = r_{MZ} - h^2$ where r_{MZ} is the correlation between MZ twins because MZ similarity can be conceptualized as h^2 (similarity due to genetic influences) + c^2 (similarity due to shared environmental influences). Substituting $2(r_{MZ} - r_{DZ})$ for h^2 offers another way of calculating shared environment ($c^2 = 2r_{DZ} - r_{MZ}$). In other words, the presence of shared environmental influences on a certain trait is suggested if DZ twin similarity exceeds half the MZ twin similarity for that trait. Similarly, *nonadditive* genetic effects (dominance and/or epistasis) are implied if DZ twin similarity is less than half the MZ twin correlation. Finally, nonshared environmental influences (e^2) can be estimated from $e^2 = r_{it} - r_{MZ}$ where r_{it} is the test-retest reliability of the measure. If $1 - r_{MZ}$ is used to estimate e^2 instead, the resulting nonshared environmental influences are confounded with measurement error. For example, in studies of more than 10,000 MZ and DZ twin pairs on general cognitive ability (g), the average MZ correlation is 0.86, which is near the test-retest reliability of the measures, in contrast to the DZ correlation of 0.60 [21]. Based on this data, application of the above formulae results in a heritability estimate of $h^2 = 0.52$, shared environmental estimate of $c^2 = 0.34$ and nonshared environmental influence/measurement error estimate of $e^2 = 0.14$.

Thus, given that MZ and DZ twin correlations are available, this straightforward set of formulae can be used to derive estimates of genetic and environmental influences on any given trait under study. Instead of the **Pearson product-moment correlation** coefficient, twin similarity is typically calculated using **intra-class correlation** (ICC1.1; [24]) which is identical to the former only if there are no mean or variance differences between the twins.

Requirements and Assumptions

It should be noted that for a meaningful interpretation of twin correlations in the described manner, a

number of assumptions have to be met: The absence of assortative mating for the trait in question, the absence of **G(enotype) – E(nvironment) correlation** and interaction, and the viability of the Equal Environments Assumption. Each of these assumptions will be addressed briefly below.

Assortative mating describes nonrandom mating that results in similarity between spouses and increases correlations and the genetic similarity for first-degree relatives if the trait under study shows genetic influence. Assortative mating can be inferred from spouse correlations which are comparably low for some psychological traits (e.g., personality), yet are substantial for others (e.g., intelligence), with average spouse correlations of about .40 [10]. In twin studies, assortative mating could result in underestimates of heritability because it raises the DZ correlation but does not affect the MZ correlation. If assortative mating were not taken into account, its effects would be attributed to the shared environment.

Gene-Environment Correlation describes the phenomenon that genetic propensities can be correlated with individual differences in experiences. Three types of $G \times E$ correlations are distinguished: passive, evocative, and active [19]. Previous research indicates that genetic factors often contribute substantially to measures of the environment, especially the family environment [18]. In the classic twin study, however, $G \times E$ correlation is assumed to be zero because it is essentially an analysis of main effects.

$G \times E$ interaction (*see Gene-Environment Interaction*) is often conceptualized as the genetic control of sensitivity to the environment [11]. Heritability that is conditional on environmental exposure can indicate the presence of a $G \times E$ interaction. The classic twin study does not address $G \times E$ interaction.

The classic twin model assumes the equality of pre- and postnatal environmental influences within the two types of twins. In other words, the *Equal Environments Assumption* (EEA) assumes that environmentally caused similarity is roughly the same for both types of twins reared in the same family. Violations of the EEA because MZ twins experience more similar environments than DZ twins would inflate estimates of genetic influences. The EEA has been tested in a number of studies and even though MZ twins appear to experience more similar environments than DZ twins, it is typically concluded that these differences do not seem to be responsible for the greater MZ twin compared to DZ twin

similarity [3]. For example, empirical studies have shown that within a group of MZ twins, those pairs who were treated more individually than others do not behave more differently [13, 15]. Another way of putting the EEA to a test is studying twins who were mislabeled by their parents, that is, twins whose parents thought that they were dizygotic when they were in fact monozygotic and vice versa. Results typically show that the similarity of mislabeled twin pairs reflects their biological zygosity to a much greater extent than their assumed zygosity [12, 23].

In recent years, the *prenatal environment* among twins has received increasing attention (e.g., [4]). About two thirds of MZ twin pairs are monochorionic (MC). All DZ twins are dichorionic (DC). The type of placentation in MZ twins is a consequence of timing in zygotic division. If the division occurs at an early stage (up to day 3 after fertilization), the twins will develop separate fetal membranes (chorion and amnion), that is, they will be dichorionic diamniotic. When the division occurs later (between days 4 and 7), the twins will be monochorionic diamniotic. For anthropological measures (such as height), a 'chorion effect' has been documented: Within-pair differences are larger in MC-MZs than in DC-MZs. Findings for cognitive and personality measures are less consistent. If possible, chorionicity should be taken into account in twin studies even if the above example shows that failing to discriminate between MC and DC MZ twin pairs does not necessarily lead to an overestimate of heritability. The importance of the prenatal maternal environment on IQ has been demonstrated in a recent meta-analysis [6].

Structural Equation Modeling

The comparison of intra-class correlations between MZ versus DZ twins can be regarded as a reasonable first step in our understanding of the etiology of particular traits. This approach, however, cannot accommodate the effect of gender on variances and covariances of opposite-sex DZ twins. To model genetic and environmental effects as the contribution of unmeasured (latent) variables to phenotypic differences, **Structural Equation Modelling (SEM)** is required. Analyzing univariate data from MZ and DZ twins by means of SEM offers numerous advances over the mere use of correlations, including an overall

statistical fit of the model, tests of parsimonious sub-models, and maximum likelihood confidence intervals for each latent influence included in the model. The true strength of SEM, however, lies in its application to multivariate and multigroup data. During the last decade powerful models and programs to efficiently run these models have been developed [16]. Extended twin designs and the simultaneous analysis of correlated traits are among the most important developments that go beyond the classic twin designs, yet still use the information inherently available in twins [1].

Outlook

Results from classical twin studies have made a remarkable contribution to one of the most dramatic developments in psychology during the past few decades: The increased recognition of the important contribution of genetic factors to virtually every psychological trait [20], particularly for phenotypes such as autism [8]. Currently, worldwide registers of extensive twin data are being established and combined with data from additional family members offering completely new perspectives in a refined behavioral genetic research [1]. Large-scale longitudinal twin studies (*see Longitudinal Designs in Genetic Research*) such as the Twins Early Development Study (TEDS) [26] offer opportunities to study the etiology of traits across time and at the extremes and compare it to the etiology across the continuum of trait expression. In this way, twin data remains a valuable and vital tool in the toolbox of behavior genetics.

References

- [1] Boomsma, D., Busjahn, A. & Peltonen, L. (2002). Classical twin studies and beyond, *Nature Reviews Genetics* **3**, 872–882.
- [2] Bouchard Jr, T.J., Lykken, D.T., McGue, M., Segal, N.L. & Tellegen, A. (1990). Sources of human psychological differences: the Minnesota study of twins reared apart, *Science* **250**, 223–228.
- [3] Bouchard Jr, T.J. & Propping, P. (1993). *Twins as a Tool of Behavioral Genetics*, John Wiley & Sons, New York.
- [4] Carlier, M. & Spitz, E. (2000). The twin method, in *Neurobehavioral Genetics: Methods and Applications*, B. Jones & P. Mormède eds, CRC Press, pp. 151–159.
- [5] Chen, W.J., Chang, H.W., Lin, C.C.H., Chang, C., Chiu, Y.N. & Soong, W.T. (1999). Diagnosis of zygosity

- by questionnaire and polymarker polymerase chain reaction in young twins, *Behavior Genetics* **29**, 115–123.
- [6] Devlin, B., Daniels, M. & Roeder, K. (1997). The heritability of IQ, *Nature* **388**, 468–471.
- [7] Falconer, D.S. (1960). *Introduction to Quantitative Genetics*, Ronald Press, New York.
- [8] Folstein, S. & Rutter, M. (1977). Genetic influences and infantile autism, *Nature* **265**, 726–728.
- [9] Galton, F. (1876). The history of twins as a criterion of the relative powers of nature and nurture, *Royal Anthropological Institute of Great Britain and Ireland Journal* **6**, 391–406.
- [10] Jensen, A.R. (1998). *The g Factor: The Science of Mental Ability*, Praeger, London.
- [11] Kendler, K.S. & Eaves, L.J. (1986). Models for the joint effects of genotype and environment on liability to psychiatric illness, *American Journal of Psychiatry* **143**, 279–289.
- [12] Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness, *Behavior Genetics* **23**, 21–27.
- [13] Loehlin, J.C. & Nichols, J. (1976). *Heredity, Environment and Personality*, University of Texas, Austin.
- [14] Merriman, C. (1924). The intellectual resemblance of twins, *Psychological Monographs* **33**, 1–58.
- [15] Morris Yates, A., Andrews, G., Howie, P. & Henderson, S. (1990). Twins: a test of the equal environments assumption, *Acta Psychiatrica Scandinavica* **81**, 322–326.
- [16] Neale, M.C., Boker, S.M., Xie, G. & Maes, H. (1999). *Mx: Statistical Modeling*, 5th Edition, VCU Box 900126, Department of Psychiatry, Richmond, 23298.
- [17] Pedersen, N.L., McClearn, G.E., Plomin, R. & Nesselroade, J.R. (1992). Effects of early rearing environment on twin similarity in the last half of the life span, *British Journal of Developmental Psychology* **10**, 255–267.
- [18] Plomin, R. (1994). *Genetics and Experience: The Interplay Between Nature and Nurture*, Sage Publications, Thousand Oaks.
- [19] Plomin, R., DeFries, J.C. & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior, *Psychological Medicine* **84**, 309–322.
- [20] Plomin, R., DeFries, J.C., McClearn, G.E. & McGuffin, P. (2001). *Behavioral Genetics*, 4th Edition, Worth Publishers, New York.
- [21] Plomin, R. & Spinath, F.M. (2004). Intelligence: genetics, genes, and genomics, *Journal of Personality and Social Psychology* **86**, 112–129.
- [22] Rende, R.D., Plomin, R. & Vandenberg, S.G. (1990). Who discovered the twin method? *Behavior Genetics* **20**, 277–285.
- [23] Scarr, S. & Carter-Saltzman, L. (1979). Twin method: defense of a critical assumption, *Behavior Genetics* **9**, 527–542.
- [24] Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability, *Psychological Bulletin* **86**, 420–428.
- [25] Siemens, H.W. (1924). *Die Zwillingspathologie: Ihre Bedeutung, ihre Methodik, Ihre Bisherigen Ergebnisse. (Twin Pathology: Its Importance, its Methodology, its Previous Results)*, Springer, Berlin.
- [26] Trouton, A., Spinath, F.M. & Plomin, R. (2002). Twins early development study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood, *Twin Research* **5**, 444–448.

FRANK M. SPINATH

Twins Reared Apart Design

NANCY L. SEGAL

Volume 4, pp. 2074–2076

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Twins Reared Apart Design

Studies of identical twins raised apart from birth provide direct estimates of genetic influence on behavior. This is because when twins are brought up separately, in uncorrelated environments and with minimal or no contact until adulthood, their trait similarity is associated with their shared genes. Fraternal, or non-identical, twins reared apart offer investigators an important control group, as well as opportunities for tests of various interactions between genotypes and environments. Separated twins are rare, relative to twins reared together. However, there have been seven studies of reared apart twins, conducted in six countries. They include the United States, Great Britain, Denmark, Japan, Sweden, and Finland [9]. Identifying reared apart twins has been easier in Scandinavian nations where extensive population registries are maintained.

Heritability, the portion of population variance associated with genetic differences in measured traits can be calculated more efficiently using twins reared apart than twins reared together. For example, 400 to 500 identical and 400 to 500 fraternal twin pairs reared together allow heritability to be estimated with the same degree of confidence as 50 MZ twin pairs reared apart [5]. This is because heritability is estimated directly from reared apart twins and indirectly from reared together twins.

Findings from reared apart twin studies have been controversial. Four key objections have been raised: (a) twins reared by relatives will be more similar than twins reared by unrelated families; (b) twins separated late will be more alike than twins separated early; (c) twins meeting before they are tested will be more similar than twins meeting later because of their increased social contact; and (d) similarities in twins' separate homes will be associated with their measured similarities [12]. These objections have, however, been subjected to testing and can be ruled out [1, 2].

A large number of analyses can be conducted by using the family members of reared apart twins in ongoing studies. It is possible to compare twin-spouse similarity, spouse-spouse similarity and similarity between the two sets of offspring who are 'genetic half-siblings' [8]. Another informative addition includes the unrelated siblings with whom the

twins were raised; these comparisons offer tests of the extent to which shared environments are associated with behavioral resemblance among relatives [11].

A large number of reared apart twin studies have demonstrated genetic influence on psychological, physical, and medical characteristics [1]. One of the most provocative findings concerns personality development. A personality study combining four twin groups (identical twins raised together and apart; fraternal twins raised together and apart) found that the degree of resemblance was the same for identical twins raised together and identical twins raised apart [13]. The shared environment of the twins reared together did not make them more similar than their reared apart counterparts; thus, personality similarity in family members is explained by their similar genes, not by their similar environments. However, twins reared together are somewhat more alike in general intelligence than twins reared apart; the **intraclass correlations** are 0.86 and 0.75, respectively [6]. These values may be misleading because most twins reared together are measured as children (when the modest effects of the shared family environment on ability are operative), while twins reared apart are measured as adults. It is possible that adult twins reared together would be as similar in intelligence as adult twins reared apart.

Studies of identical reared apart twins are natural **co-twin control studies**, thus providing insights into environmental effects on behavior. For example, it is possible to see if differences in twins' rearing histories are linked to differences in their current behavior. An analysis of relationships between IQ and rearing family measures (e.g., parental socioeconomic status, facilities in the home) did not find meaningful associations [2]. Case studies of selected pairs are also illustrative in this regard. Twins in one set of identical British women were raised by parents who provided them with different educational opportunities. Despite these differences, the twins' ability levels were quite close and both twins were avid readers of the same type of books.

Twins reared apart can also be used to examine evolutionary-based questions and hypotheses [10]. A finding that girls raised in father-absent homes undergo early sexual development and poor social relationships has attracted attention [4]. It has been suggested that studies of MZ female cotwins reared separately in father-absent and father-present homes

2 Twins Reared Apart Design

could clarify factors underlying this behavioral pattern [3]. Further discussion of the potential role of reared apart twins in evolutionary psychological work is available in Mealey [7].

References

- [1] Bouchard Jr, T.J. (1997). IQ similarity in twins reared apart: findings and responses to critics, in *Intelligence: Heredity and Environment*, R.J. Sterberg & E.L. Grigorenko eds. Cambridge University Press, New York. pp. 126–160.
- [2] Bouchard Jr, T.J., Lykken, D.T., McGue, M., Segal, N.L. & Tellegen, A. (1990). Sources of human psychological differences: the minnesota study of twins reared apart. *Science* **250**, 223–228.
- [3] Crawford, C.B. & Anderson, J.L. (1989). Sociobiology: an environmentalist discipline? *American Psychologist* **44**, 1449–1459.
- [4] Ellis, B.J. & Garber, J. (2000). Psychosocial antecedents of variation in girls' pubertal timing: maternal depression, stepfather presence, and marital and family stress, *Child Development* **71**, 485–501.
- [5] Lykken, D.T., Geisser, S. & Tellegen, A. (1981). Heritability estimates from twin studies: The efficiency of the MZA design, Unpublished manuscript.
- [6] McGue, M., & Bouchard Jr, T.J. (1998). Genetic and environmental influences on human behavioral differences, *Annual Review of Neuroscience*, **21**, 1–24.
- [7] Mealey, L. (2001). Kinship: the tie that binds (disciplines), in *Conceptual Challenges in Evolutionary Psychology: Innovative Research Strategies*, Holcomb III, H.R. ed., Kluwer Academic Publisher, Dordrecht, pp. 19–38.
- [8] Segal, N.L. (2000). *Entwined Lives: Twins and What they Tell us About Human Behavior*, Plume, New York.
- [9] Segal, N.L. (2003). Spotlights: reared apart twin researchers, *Twin Research* **6**, 72–81.
- [10] Segal, N.L. & Hershberger, S.L. (1999). Cooperation and competition in adolescent twins: findings from a prisoner's dilemma game, *Evolution and Human Behavior* **20**, 29–51.
- [11] Segal, N.L., Hershberger, N.L. & Arad, S. (2003). Meeting one's twin: perceived social closeness and familiarity, *Evolutionary Psychology* **1**, 70–95.
- [12] Taylor, H.F. (1980). *The IQ Game: A Methodological Inquiry into the Heredity-Environment Controversy*, Rutgers University Press, New Brunswick.
- [13] Tellegen, A., Lykken, D.T., Bouchard Jr, T.J., Wilcox, K.J., Segal, N.L. & Rich, S. (1988). Personality similarity in twins reared apart and together, *Journal of Personality and Social Psychology*, **54**, 1031–1039.

NANCY L. SEGAL

Two by Two Contingency Tables

LI LIU AND VANCE W. BERGER

Volume 4, pp. 2076–2081

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Two by Two Contingency Tables

In a two by two **contingency table**, both the row variable and the column variable have two levels, and each cell represents the count for that specific condition. This type of table arises in many different contexts, and can be generated by different sampling schemes. We illustrate a few examples.

- Randomly sample subjects from a single group and cross classify each subject into four categories corresponding to the presence or absence of each of two conditions. This yields a multinomial distribution, and each cell count is considered to be the result of an independent Poisson process. For example, the 1982 General Social Survey [1] was used to sample and classify individuals by their opinions on gun registration and the death penalty. This sampling scheme yields a multinomial distribution with four outcomes that can be arranged as a 2×2 table, as in Table 1. The row variable represents the opinion on gun registration, and the column variable represents the opinion on the death penalty.
- Randomly sample subjects from each of two groups, and classify each of them by a single binary variable. This results in two independent binomial distributions (*see Catalogue of Probability Density Functions*). For example, a study was designed to study whether cigarette smoking is related to lung cancer [6]. Roughly equal numbers of lung cancer patients and controls (without lung cancer) were asked whether they smoked or not (see Table 2).
- Randomly assign subjects (selected with or without randomization) to one of two treatments, and then classify each subject by a binary response.

Table 1 Opinions on the death penalty cross-classified with opinions on gun registration

Gun registration	Death penalty		Total
	Favor	Oppose	
Favor	784	236	1020
Oppose	311	66	377
Total	1095	302	1397

Table 2 Lung cancer and smoking

Smoker	Lung cancer		Total
	Yes	No	
Yes	647	622	1269
No	2	27	29
Total	649	649	1298

Table 3 Treatment and depression improvement

Treatment	Depression improved?		Total
	Yes	No	
Pramipexole	8	4	12
Placebo	2	8	10
Total	10	12	22

For example, a preliminary randomized, placebo-controlled trial (*see Clinical Trials and Intervention Studies*) was conducted to determine the antidepressant efficacy of pramipexole in a group of 22 treatment-resistant bipolar depression outpatients [8]. The data are as in Table 3.

Under the null hypothesis that the two treatments produce the same response in any given subject (that is, that response is an attribute of the subject, independent of the treatments, so that there are some patients destined to respond and others destined not to) [2], the column totals are fixed and the cell counts follow the hypergeometric distribution.

In a two by two table, we are interested in studying whether there is a relationship between the row variable and column variable. If there is an association, then we also want to know how strong it is, and how the two variables are related to each other. The following topics in the paper will help us understand these questions. We use Table 4 to represent any arbitrary two by two table,

Table 4 Generic two-by-two table

Column variable	Row variable		Total
	Level 1	Level 2	
Level 1	n_{11}	n_{12}	n_{1+}
Level 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

2 Two by Two Contingency Tables

where $n_{11}, n_{12}, n_{21}, n_{22}$ represent the cell counts, $n_{1+}, n_{2+}, n_{+1}, n_{+2}$ represent the row totals and column totals, respectively, and n represents the total count in the table.

χ^2 Test of Independence

For a two by two contingency table, an initial question is whether the row variable and the column variable are independent or not, and the χ^2 test of independence is one method that can be used to answer this question. Given that all the marginal totals (row totals and column totals) are fixed under the null hypothesis that the row variable and column variable are independent, the expected value of n_{ij} , assuming independence of rows and columns, is

$$m_{ij} = \frac{n_{i+}n_{+j}}{n}, \quad (1)$$

and Pearson proposed the test statistic

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \quad (2)$$

If the cell counts are large enough, then χ^2 has approximately a chi-square distribution with one degree of freedom. Large values of χ^2 lead to the rejection of the null hypothesis of no association. For contingency Table 1, $\chi^2 = 5.15$ with one degree of freedom, and the P value is 0.0232, which suggests that there is an association between one's opinion on gun registration and one's opinion on the death penalty. Generally, the conventional wisdom is that all the expected cell counts m_{ij} should be larger than five for the test to be valid, and that if some cell counts are small, then Fisher's exact test (*see Exact Methods for Categorical Data*) can be used instead. This is not a very sensible plan, however, because it would be quite difficult to justify the use of an approximate test given the availability of the exact test it is trying to approximate [2, 4]. Moreover, it has been demonstrated that even if all expected cell counts exceed five, the approximate test can still give different results from the exact test. Just as it is a better idea to wear a seat belt in all weather rather than just in inclement weather, the safe approach is to select an exact test all the time. Hence Fisher's exact test should be used

instead of the chi-square test, for any expected cell counts.

Difference in Proportions

If there is an association based on the chi-square test of independence, or preferably Fisher's exact test, then we may be interested in knowing how the two variables are related. One way is to study the proportions for the two groups, and see how they differ. The difference in proportions is used to compare the conditional (on the row) distributions of a column response variable across the two rows. For these measures, the rows are treated as independent binomial samples. Consider Table 1, as an example. Let π_1 and π_2 represent the probabilities of favoring death penalty for those favoring and opposing gun registration from the population, respectively. Then we are interested in estimating the difference between π_1 and π_2 . The sample proportion (using the notation of Table 4) of those favoring the death penalty, among those favoring gun registration, is $p_1 = n_{11}/n_{1+}$, and has expectation π_1 and variance $\pi_1(1 - \pi_1)/n_{1+}$, and the sample proportion of those favoring the death penalty among those opposing gun registration can be computed accordingly. Thus, the difference in proportions has expectation of

$$E(p_1 - p_2) = \pi_1 - \pi_2, \quad (3)$$

and variance (using the notation of Table 4)

$$\sigma^2(p_1 - p_2) = \frac{\pi_1(1 - \pi_1)}{n_{1+}} + \frac{\pi_2(1 - \pi_2)}{n_{2+}}, \quad (4)$$

and the estimated variance (using the notation of Table 4) is

$$\hat{\sigma}^2(p_1 - p_2) = \frac{p_1(1 - p_1)}{n_{1+}} + \frac{p_2(1 - p_2)}{n_{2+}}. \quad (5)$$

Then a $100(1 - \alpha)\%$ **confidence interval** for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z_{\alpha/2} \hat{\sigma}(p_1 - p_2). \quad (6)$$

For Table 1, we have

$$\begin{aligned} p_1 - p_2 &= \frac{784}{1020} - \frac{311}{377} = 0.7686 - 0.8249 \\ &= -0.0563, \end{aligned} \quad (7)$$

and

$$\hat{\sigma}^2(p_1 - p_2) = \frac{0.7686(1 - 0.7686)}{1020} + \frac{0.8249(1 - 0.8249)}{377} = 0.000557. \quad (8)$$

So the 95% confidence interval for the true difference is $-0.0563 \pm 1.96(0.0236)$, or $(-0.010, -0.103)$. Since this interval contains only negative values, we can conclude that $\pi_1 - \pi_2 < 0$, so people who favor gun registration are less likely to favor the death penalty.

Relative Risk and Odds Ratio

To compare how the two groups differ, we can use the difference of the two proportions. Naturally, we can also use the ratio of the two proportions, and this is called the **relative risk**. The estimated relative risk is

$$RR = \frac{p_1}{p_2} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}, \quad (9)$$

and the estimated variance of $\log(RR)$ is

$$\hat{\sigma}^2(\log(RR)) = \frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}. \quad (10)$$

Thus a $100(1 - \alpha)\%$ confidence interval for the relative risk π_1/π_2 is

$$\exp(RR \pm z_{\alpha/2} \hat{\sigma}(\log(RR))). \quad (11)$$

Instead of computing the ratio of the proportion of yes for group 1 versus group 2, we can also compute the ratio of the odds of yes for group 1 versus for group 2. This is called the **odds ratio**. The estimated odds ratio is

$$OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}, \quad (12)$$

and the estimated variance of $\log(OR)$ is

$$\hat{\sigma}^2(\log(OR)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}. \quad (13)$$

Then a $100(1 - \alpha)\%$ confidence interval for odds ratio $\pi_1/(1 - \pi_1)/\pi_2/(1 - \pi_2)$ is

$$\exp(OR \pm z_{\alpha/2} \hat{\sigma}(\log(OR))). \quad (14)$$

The sample odds ratio is either to 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries are 0 for a row or column. To solve this problem, a modified estimate can be used by adding 1/2 to each cell count, and its variance can be estimated accordingly. The relationship between the odds ratio and the relative risk is shown in the following formula

$$RR = OR \times \frac{1 - p_1}{1 - p_2}. \quad (15)$$

Unlike the relative risk, the odds ratio can be used to measure an association no matter how the data were collected. This is very useful for rare disease retrospective studies such as the study in Table 2. In such a study, we cannot obtain the relative risk directly, however, we can still compute the odds ratio. Since $1 - p_1 \approx 1$ and $1 - p_2 \approx 1$ for rare diseases, the relative risk and odds ratio are numerically very close in such studies, and so the odds ratio can be used to estimate the relative risk. For example, if $p_1 = 0.01$ and $p_2 = 0.001$, then we have

$$\begin{aligned} RR &= \frac{p_1}{p_2} = 10, \quad \text{and} \\ OR &= \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{0.01/0.99}{0.001/0.999} = 10.09, \end{aligned} \quad (16)$$

which are very close. For the data in Table 2, we have

$$\begin{aligned} OR &= \frac{P(E/D)/P(\bar{E}/D)}{P(E/\bar{D})/P(\bar{E}/\bar{D})} = \frac{P(E \cap D)/P(\bar{E} \cap D)}{P(E \cap \bar{D})/P(\bar{E} \cap \bar{D})} \\ &= \frac{P(D/E)/P(D/\bar{E})}{P(\bar{D}/E)/P(\bar{D}/\bar{E})} \approx \frac{P(D/E)}{P(D/\bar{E})} = RR, \end{aligned} \quad (17)$$

since $P(\bar{D}/E)$ and $P(\bar{D}/\bar{E})$ are almost 1 for rare disease. So

$$RR \approx OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{647 \times 27}{622 \times 2} = 14.04. \quad (18)$$

This statistic indicates that the risk of getting lung cancer is much higher for smokers than it is for nonsmokers.

Sensitivity, Specificity, False Positive Rate, and False Negative Rate

These measures are commonly used when evaluating the efficacy of a screening test for a disease outcome.

4 Two by Two Contingency Tables

Table 5 VAP diagnosis results

Disease status	Chest radiograph		Total
	Diagnosis +	Diagnosis −	
VAP	12	1	13
No VAP	8	4	12
Total	20	5	25

Table 5 contains a study assessing the accuracy of chest radiograph (used to detect radiographic infiltrates) for the diagnosis of ventilator associated pneumonia (VAP) [7]. The sensitivity is the true proportion of positive diagnosis results among the VAP patients, and the specificity is the true proportion of negative diagnosis results among those without VAP. Both the sensitivity and the specificity can be estimated by

$$\begin{aligned}\text{Sensitivity} &= \frac{n_{11}}{n_{1+}} = P(\text{Diagnosis} + | \text{VAP}) \\ \text{Specificity} &= \frac{n_{22}}{n_{2+}} = P(\text{Diagnosis} - | \text{No VAP})\end{aligned}\quad (19)$$

For the data in Table 5, sensitivity = 12/13 = 0.92 and specificity = 4/12 = 0.33.

Sensitivity is also called the *true positive rate*, and specificity is also called the *true negative rate*. They are closely related to two other two rates, specifically the false positive rate and the false negative rate. The false negative rate is the true proportion of negative diagnosis results among the VAP patients, and the false positive rate is the true proportion of positive diagnosis results among those without VAP. The false positive rate and false negative can be estimated by

$$\begin{aligned}\text{False positive rate} &= \frac{n_{21}}{n_{2+}} \\ &= P(\text{Diagnosis} + | \text{No VAP}) \\ &= 1 - \text{specificity} \\ \text{False negative rate} &= \frac{n_{12}}{n_{1+}} \\ &= P(\text{Diagnosis} - | \text{VAP}) \\ &= 1 - \text{sensitivity},\end{aligned}\quad (20)$$

Another useful term is the false discovery rate, which is the proportion of subjects without VAP among all the positive diagnosis results.

Fisher's Exact Test

When the sample size is small, the χ^2 test based on large samples may not be valid, and Fisher's exact test can be used to test the independence of a contingency table. For given row and column totals, the value n_{11} determines the other three cell counts (there is but one degree of freedom). Under the null hypothesis of independence, the probability of a particular value n_{11} given the marginal totals is

$$P(n_{11}) = \frac{(n_{1+}!n_{2+}!)(n_{+1}!n_{+2}!)}{n!(n_{11}!n_{12}!n_{21}!n_{22}!)}, \quad (21)$$

which is the hypergeometric probability function (*see Catalogue of Probability Density Functions*). One would enumerate all possible tables of counts consistent with the row and column totals n_{i+} and n_{+j} . For each one, the associated conditional probability can be calculated using the above formula, and the sum of these probabilities must be one. To test independence, the P value is the sum of the hypergeometric probabilities for those tables at least as favorable to the alternative hypothesis as the observed one. That is, for a given two by two table, the P value of the Fisher exact test is the sum of all the conditional probabilities that correspond to tables that are as extreme as or more extreme than the observed table. Consider Table 3, for example. The null distribution of n_{11} is the hypergeometric distribution defined for all the two by two tables having row totals and column totals (12,10) and (10,12). The potential values for n_{11} are (0, 1, 2, 3, ..., 10). First, we can compute the probability of the observed table as

$$P(4) = \frac{(12!10!)(10!12!)}{22!(8!4!2!8!)} = 0.0344. \quad (22)$$

Other possible two by two tables and their probabilities are

$$\begin{aligned}\begin{bmatrix} 10 & 2 \\ 0 & 10 \end{bmatrix} \text{ Prob} &= 0.0001 & \begin{bmatrix} 9 & 3 \\ 1 & 9 \end{bmatrix} \text{ Prob} &= 0.0034 \\ \begin{bmatrix} 7 & 5 \\ 3 & 7 \end{bmatrix} \text{ Prob} &= 0.1470 & \begin{bmatrix} 6 & 6 \\ 4 & 6 \end{bmatrix} \text{ Prob} &= 0.3001 \\ \begin{bmatrix} 5 & 7 \\ 5 & 5 \end{bmatrix} \text{ Prob} &= 0.3086 & \begin{bmatrix} 4 & 8 \\ 6 & 4 \end{bmatrix} \text{ Prob} &= 0.1608 \\ \begin{bmatrix} 3 & 9 \\ 7 & 3 \end{bmatrix} \text{ Prob} &= 0.0408 & \begin{bmatrix} 2 & 10 \\ 8 & 2 \end{bmatrix} \text{ Prob} &= 0.0046 \\ \begin{bmatrix} 1 & 11 \\ 9 & 1 \end{bmatrix} \text{ Prob} &= 0.0002 & \begin{bmatrix} 0 & 12 \\ 10 & 0 \end{bmatrix} \text{ Prob} &= 0.0000015\end{aligned}\quad (23)$$

Together with the observed probability, these probabilities sum up to one. The sum of the probabilities of the tables in the two-sided rejection region is 0.0427. This rejects the null hypothesis of independence (at the customary 0.05 level), and suggests that treatment and improvement are correlated. The one-sided rejection region would consist of the tables with upper-left cell counts of 10, 9, and 8, and the one-sided P value is $0.0001 + 0.0034 + 0.0344 = 0.0379$. Fisher's exact test can be conservative, and so one can use the **mid- P value** or the P value interval [3].

Simpson's Paradox

For a $2 \times 2 \times 2$ contingency table, **Simpson's Paradox** refers to the situation in which a marginal association has a different direction from the conditional associations (see **Measures of Association**). In a $2 \times 2 \times 2$ contingency table, if we cross classify two variables X and Y at a fixed level of binary variable Z , we can obtain two 2×2 tables with variables X and Y and they are called *partial tables*. If we combine the two partial tables, we can obtain one 2×2 table and this is called the *marginal table*. The associations in the partial tables are called *partial associations*, and the association in the marginal table is called *marginal association*. Table 6 is a $2 \times 2 \times 2$ contingency table that studied the relationship between urinary tract infections (UTI) and antibiotic prophylaxis (ABP) [9].

The last section of Table 6 displays the marginal association. As we can see, 3.28% of the patients who used antibiotic prophylaxis got UTI, and 4.64% of patients who did not use antibiotic prophylaxis got UTI. Clearly, the probability of UTI is lower for those who used antibiotic prophylaxis than it is for those

who did not. This is consistent with previous findings that antibiotic prophylaxis is effective in preventing UTI. But now consider the first two partial tables in Table 6, which display the conditional associations between antibiotic prophylaxis and UTI. When the patients were from four hospitals with low incidence of UTI, the probability of UTI is higher for those who used antibiotic prophylaxis $1.80\% - 0.70\% = 1.10\%$. It might seem that to compensate for this reversed effect, the patients from the four hospitals with high incidence of UTI would have shown a very strong trend in the direction of higher UTI incidence for those without the antibiotic prophylaxis. But alas this was not the case. In fact, the probability of UTI is higher for those who used antibiotic prophylaxis $13.25\% - 6.51\% = 6.74\%$.

Thus, controlling for hospital type, the probability of UTI is higher for those who used antibiotic prophylaxis. The partial association gives a different direction of association compared to the marginal associations. This is called *Simpson's Paradox* [9]. The reason that the marginal association and partial associations have different directions is related to the association between the control variable, hospital type, and the other two variables. First, consider the association between hospital type and the usage of antibiotic prophylaxis based on the marginal table with these two variables. The odds ratio equals

$$\frac{1113 \times 1520}{(720 \times 166)} = 14.15, \quad (24)$$

which indicates a strong association between hospital type and the usage of antibiotic prophylaxis. Patients from the four hospitals with low incidence of UTI were more likely to have used antibiotic prophylaxis. Second, the probability of UTI is tautologically higher for the four high incidence hospitals. The odds

Table 6 Urinary tract infections (UTI) and antibiotic prophylaxis (ABP)

Hospital	Antibiotic prophylaxis	UTI?		Percentage
		Yes	No	
Patients from four hospitals with Low incidence of UTI ($\leq 2.5\%$)	Yes	20	1093	1.80%
	No	5	715	0.70%
Patients from four hospitals with High incidence of UTI ($> 2.5\%$)	Yes	22	144	13.25%
	No	99	1421	6.51%
Total	Yes	42	1237	3.28%
	No	104	2136	4.64%

6 Two by Two Contingency Tables

ratio based on the marginal table of hospital and UTI is

$$\frac{25 \times 1565}{(1808 \times 121)} = 0.18. \quad (25)$$

An explanation of the contrary results in Simpson's Paradox is that there are other confounding variables that may have been unrecognized.

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.
- [2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [3] Berger, V.W. (2001). The p-Value Interval as an Inferential Tool, *Journal of the Royal Statistical Society D (The Statistician)* **50**, **1**, 79–85.
- [4] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**, 74–82.
- [5] Berger, V.W. (2004). Valid Adjustment of Randomized Comparisons for Binary Covariates, *Biometrical Journal* **46**(5), 589–594.
- [6] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung; preliminary report, *British Medical Journal* **2**(4682), 739–748.
- [7] Fabregas, N., Ewig, S., Torres, A., El-Ebiary, M., Ramirez, J., de La Bellacasa, J.P., Bauer, T., Cabello, H., (1999). Clinical diagnosis of ventilator associated pneumonia revisited: comparative validation using immediate post-mortem lung biopsies, *Thorax* **54**, 867–873.
- [8] Goldberg, J.F., Burdick, K.E. & Endick, C.J. (2004). Preliminary randomized, double-blind, placebo-controlled trial of pramipexole added to mood stabilizers for treatment-resistant bipolar depression, *The American Journal of Psychiatry* **161**(3), 564–566.
- [9] Reintjes, R., de Boer, A., van Pelt, W. & Mintjes-de Groot, J. (2000). Simpson's paradox: an example from hospital epidemiology, *Epidemiology* **11**(1), 81–83.

LI LIU AND VANCE W. BERGER

Two-mode Clustering

IVEN VAN MECHELEN, HANS-HERMANN BOCK AND PAUL DE BOECK

Volume 4, pp. 2081–2086

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Two-mode Clustering

Data in the behavioral sciences often can be written in the form of a rectangular matrix. Common examples include observed performances denoted in a person by task matrix, questionnaire data written in a person by item matrix, and semantic information written in a concept by feature matrix. Rectangular matrix data imply two sets of entities or *modes*, the row mode and the column mode.

Within the family of two-mode data, several subtypes may be distinguished. Common subtypes include case by variable data that denote the value (data entry) that each of the variables (columns) under study takes for each case (row) under study, contingency table data (*see* **Contingency Tables**) that denote the frequency (data entry) with which each combination of a row entity and a column entity is being observed, and categorical predictor/criterion data that denote the value of some criterion variable (data entry) observed for each pair of values of two categorical predictor variables (corresponding to the elements of the row and column modes).

Two-mode data may be subjected to various kinds of clustering methods. Two major subtypes are *indirect* and *direct approaches*. Indirect clustering methods presuppose the conversion of the two-mode data into a set of (one-mode) proximities (similarities or dissimilarities *see* **Proximity Measures**) among the elements of either the row or the column mode; this conversion is usually done as a separate step prior to the actual **cluster analysis**. Direct clustering methods, on the other hand, directly operate on the two-mode data without a preceding proximity transformation. Several direct clustering methods yield a clustering of the elements of *only one of the modes* of the data. Optionally, such methods can be applied twice to yield successively a clustering of the row entities and a clustering of the column entities.

As an alternative, one may wish to rely on direct *two-mode clustering methods*. Such methods yield clusterings of rows and columns *simultaneously* rather than successively. The most important advantage of simultaneous approaches is that they may reveal information on the linkage between the two modes of the data.

As most methods of data analysis, two-mode clustering methods imply a reduction of the data, and, hence, a loss of information. The goal of the

clustering methods, however, is that the loss is as small as possible with regard to the particular subtype of information that constitutes the target of the clustering method under study, and into which the method aims at providing more insight. For case by variable data, the target information typically consists of the actual values the variables take for each of the cases; two-mode clustering methods for this type of data aim at reconstructing these values as well as possible. Furthermore, for contingency table data, the target information typically pertains to the amount of dependence between the row and column modes, whereas for categorical predictor/criterion data it consists of the amount of interaction as implied by the prediction of the data entries on the basis of the categorical row and column variables.

Basic Two-mode Clustering Concepts

Since the pioneering conceptual and algorithmic work by Hartigan [3] and Bock [1], a large number of quite diverse simultaneous clustering methods has been developed. Those range from heuristic *ad hoc* procedures, over deterministic structures estimated in terms of some objective or loss function, to fully stochastic model-based approaches. A structured overview of the area can be found in [4].

To grasp the multitude of methods, two conceptual distinctions may be useful:

1. The *nature of the clusters*: a cluster may be a set of row elements (*row cluster*), a set of column elements (*column cluster*), or a Cartesian product of a set of row elements and a set of column elements (*data cluster*). One may note that each data cluster as obtained from a two-mode clustering procedure always implies a row and a column cluster; the reverse, however, does not necessarily hold.
2. The *set-theoretical structure* of a particular set of clusters or clustering (see also Figure 1, to be discussed below): this may be (a) a partitioning, (b) a nested clustering (i.e., a clustering that includes intersecting clusters, albeit such that intersecting clusters are always in a subset-superset relation), and (c) an overlapping clustering (i.e., a clustering that includes intersecting, nonnested clusters).

2 Two-mode Clustering

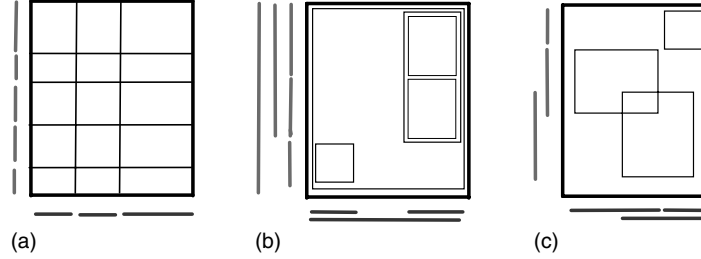


Figure 1 Schematic representation of hypothetical examples of three types of two-mode clustering: (a) partitioning, (b) nested clustering, (c) overlapping clustering

The data clustering constitutes the cornerstone of any two-mode cluster analysis. As such it is the key element in the representation of the value, dependence, or interaction information in the data. For case by variable data, this representation (and, more in particular, the reconstruction of the data values) further relies on parameters (scalar or vector) associated with the data clusters. Otherwise, we will limit the remainder of this exposition to case by variable data.

In general, two-mode clustering methods may be considered that yield row, column and data clusterings with different set-theoretical structures as distinguished above. However, here we will consider further only methods that yield row, column, and data clusterings of the same type. Taking into account the three possible set-theoretical structures, we will therefore focus on the three types of clustering as schematically presented in Figure 1.

In what follows, we will briefly present one instance of each type of clustering method. Each instance will further be illustrated making use of the same 14×11 response by situation data matrix obtained from a single participant who was asked to rate the applicability of each of 14 anxiety responses to each of 11 stressful situations, on a 5-point scale ranging from 1 (=not applicable at all) to 5 (=applicable to a strong extent).

Partitioning

Two-mode partitioning methods of a data matrix $\mathbf{X}$ imply a partitioning of the Cartesian product of the row and column modes that is obtained by fully crossing a row and a column partitioning (see leftmost panel in Figure 1). In one common instance of this class of methods, each data cluster

$A \times B$ is associated with a (scalar) parameter $\mu_{A,B}$. The clustering and parameters are further such that all entries x_{ab} (with $a \in A, b \in B$) are as close as possible to the corresponding value $\mu_{A,B}$ (which acts as the reconstructed data value). This implies that row entries of the same row cluster behave similarly across columns, that column entries of the same column cluster behave similarly across rows, and that the data values are as homogeneous as possible within each data cluster.

The result of a two-mode partitioning of the anxiety data is graphically represented in Figure 2. The representation is one in terms of a so-called *heat map*, with the estimated $\mu_{A,B}$ -parameters being represented in terms of grey values. The analysis reveals, for instance, an avoidance behavior class (third row cluster) that is associated fairly strongly with a class of situations that imply some form of psychological assessment (third column cluster).

Nested Clustering

One instance in this class of methods is *two-mode ultrametric tree modeling*. In this method, too, each data cluster $A \times B$ is associated with a (scalar) parameter $\mu_{A,B}$. As for all two-mode clustering methods for case by variable data that are not partitions, ultrametric tree models include a rule for the reconstruction of data entries in intersections of different data clusters. In the ultrametric tree model, the data reconstruction rule makes use of the Maximum (or Minimum) operator. In particular, the clustering and $\mu_{A,B}$ -parameters are to be such that all entries x_{ab} are as close as possible to the maximum (or minimum) of the $\mu_{A,B}$ -values of all data clusters $A \times B$ to which (a, b) belongs.

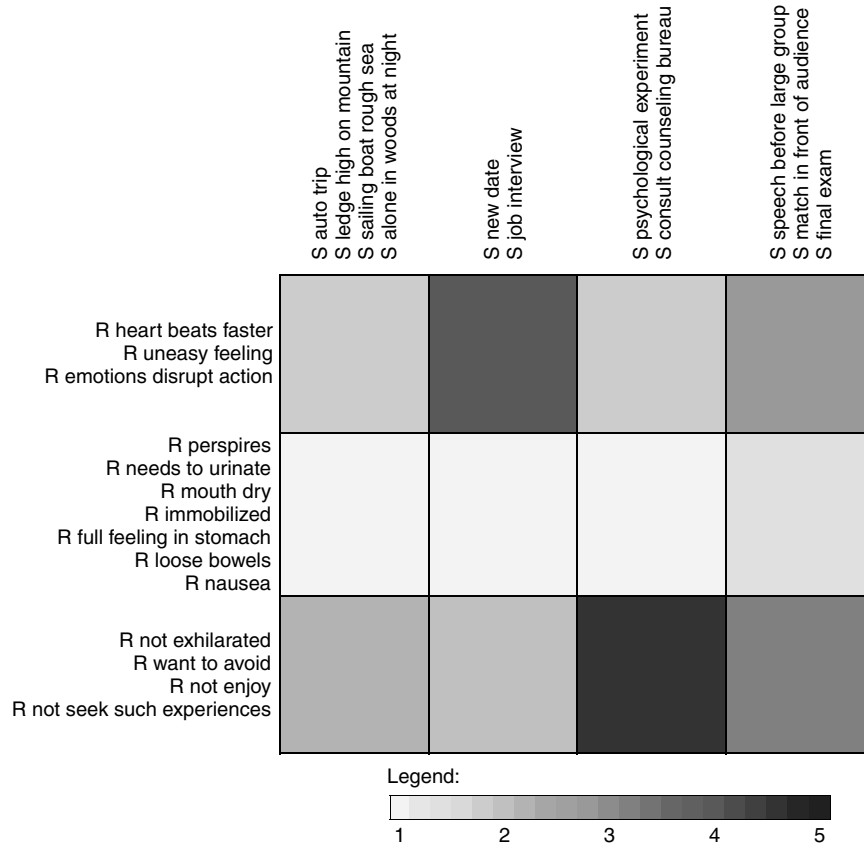


Figure 2 Two-mode partitioning of anxiety data

Optionally, a data cluster can be interpreted as a feature that applies to the row and column entries involved in it.

Part of the result of a two-mode ultrametric tree analysis of the anxiety data (making use of a constrained optimization algorithm) is graphically represented as a tree diagram in Figure 3. (The representation is limited to a subset of situations and responses to improve readability.) In Figure 3, the data clusters correspond to the vertical lines and comprise all leaves at the right linked to them. The $\mu_{A,B}$ -values can further be read from the hierarchical scale at the bottom of the figure. As such, one can read from the lower part of Figure 3 that the two psychological assessment-related situations ‘consult counseling bureau’ and ‘psychological experiment’ elicit ‘not enjoy’ and ‘not feel exhilarated’ with strength 5, and ‘not seek experiences like this’ with strength 4.5.

Overlapping Clustering

Hierarchical classes models [2] are overlapping two-mode clustering methods for case by variable data that make use of a data reconstruction rule with a Maximum (or Minimum) operator. In case of positive real-valued data, the hierarchical classes model can be considered a generalization of the two-mode ultrametric tree, with each data cluster again being associated with a $\mu_{A,B}$ -parameter, and with the clustering and parameters being such that all entries x_{ab} are as close as possible to the maximum of the $\mu_{A,B}$ -values of all data clusters $A \times B$ to which (a, b) belongs. The generalization implies that row, column, and data clusters are allowed to overlap. The row and column clusters of the model then can be considered to be (possibly overlapping) types that are associated with a value of association strength as specified in the model. A distinctive feature of

4 Two-mode Clustering

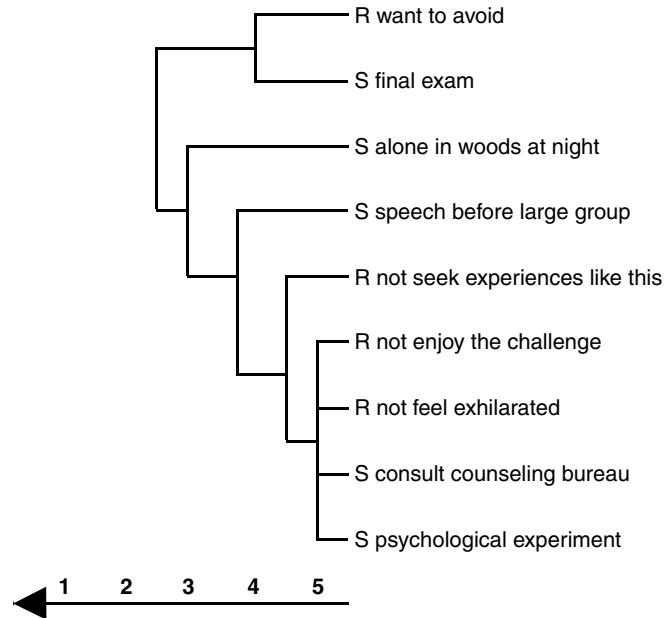


Figure 3 Part of a two-mode ultrametric tree for anxiety data

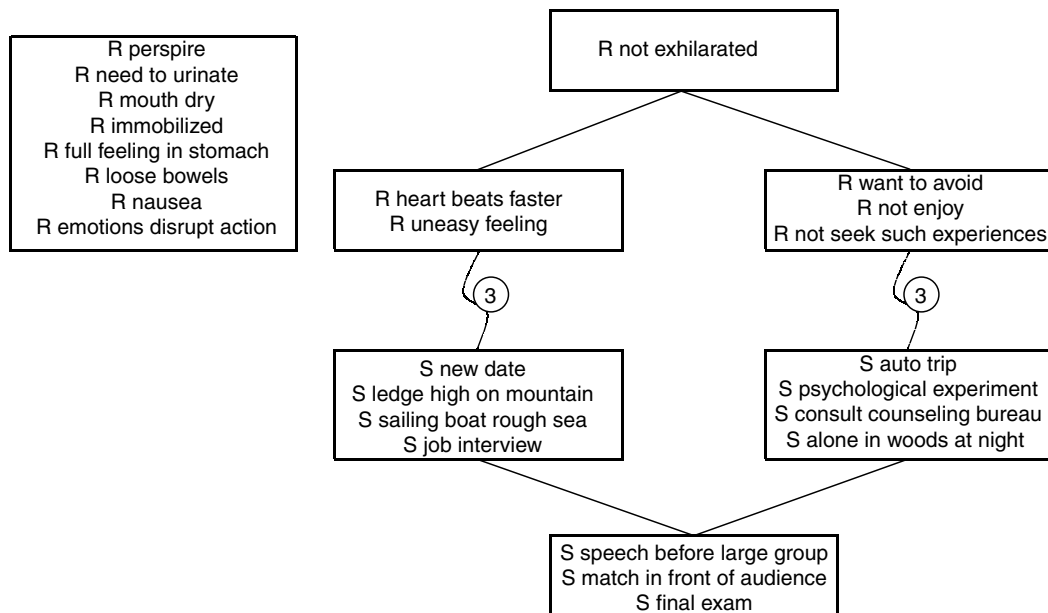


Figure 4 Overlapping clustering (HICLAS-R model) of anxiety data

hierarchical classes models further reads that they represent implicational if-then type relations among row and among column elements.

The result of a hierarchical classes analysis of the anxiety data is presented in Figure 4. A data cluster can be read from this figure as the set of

all situations and responses linked to a particular zigzag; the $\mu_{A,B}$ -value associated with this cluster is further attached as a label to the zigzag. The represented model, for instance, implies that the two psychological assessment-related situations elicit ‘not exhilarated’ and ‘not enjoy’ with value 3 (encircled number). From the representation, one may further read that, for instance, if a situation elicits ‘not enjoy’ (with some association value) then it also elicits ‘not exhilarated’ (with at least the same value of association strength) (see **Cluster Analysis: Overview; Overlapping Clusters**).

Software

Two-mode clustering methods mostly have not been incorporated in general purpose statistical packages. Exceptions include variants of a nested method (two-way joining) due to Hartigan as incorporated within *Statistica*TM and *Systat*TM, and an overlapping method (Boolean factor analysis) as incorporated within *BMDP*TM. A number of two-mode clustering methods are available from their developers as stand-alone

programs. Finally, more specific packages are being developed within specialized areas such as bioinformatics, in particular, for applications in microarray data analysis (see **Microarrays**).

References

- [1] Bock, H.-H. (1968). *Stochastische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen*, Dissertation, University of Freiburg, Germany.
- [2] De Boeck, P. & Rosenberg, S. (1988). Hierarchical classes: model and data analysis, *Psychometrika* **53**, 361–381.
- [3] Hartigan, J. (1972). Direct clustering of a data matrix, *Journal of the American Statistical Association* **67**, 123–129.
- [4] Van Mechelen, I., Bock, H.-H. & De Boeck, P. (2004). Two-mode clustering methods: a structured overview, *Statistical Methods in Medical Research* **13**, 363–394.

IVEN VAN MECHELEN, HANS-HERMANN BOCK
AND PAUL DE BOECK

Two-way Factorial: Distribution-Free Methods

LARRY E. TOOTHAKER

Volume 4, pp. 2086–2089

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Two-way Factorial: Distribution-Free Methods

The classical tests for main effects and interaction in a $J \times K$ two-way **factorial design** are the F tests in the two-way fixed-effects **analysis of variance** (ANOVA) (see **Fixed and Random Effects**), with the assumptions of normality, independence, and equal variances of errors in the $J \times K$ cells. When all assumptions are met and the null hypotheses are true, the **sampling distribution** of each F is distributed as a theoretical F distribution with appropriate degrees of freedom, in which the F -critical values cut off exactly the set α in the upper tail.

When any assumption is violated, the sampling distributions might not be well-fit by the F distributions, and the F -critical values might not cut off exactly α . Also, violation of assumptions can lead to poor **power** properties, such as when the power of F decreases as differences in means increase. The extent to which the F -statistics are resistant to violations of the assumptions is called *robustness*, and is often measured by how close the true α in the sampling distribution is to the set α , and by having power functions where power increases as mean differences increase. Also, power should be compared for different statistical methods in the same circumstances, with preference being given to those methods that maintain control of α and have the best power.

Alternatives to relying on the robustness of the F tests might be available in the area of nonparametric methods. These methods are so named because their early ancestors were originally designed to test hypotheses that had no parameters, but were tests of equality of entire distributions. Many of the modern nonparametric methods test hypotheses about parameters, although often not the means nor treatment effects of the model. They also free the researcher from the normality assumption of the ANOVA (see **Distribution-free Inference, an Overview**).

Such alternative methods include **permutation tests** [5], **bootstrap** tests [6], the rank transform [7], aligned ranks tests [9], tests on **trimmed means** [21], and a rank-based ANOVA method that allows heteroscedastic variances, called the *BDM method* (after the surnames of the authors of [4]). Some of these are more general procedures that can be useful

when combined with any statistic, for example, the bootstrap method might be combined with tests on trimmed means, using the bootstrap to make the decision on the hypothesis rather than using a standard critical value approach.

Permutation Tests

Permutation tests (see **Linear Models: Permutation Methods**) rely on permuting the observations among treatments in all ways possible, given the design of the study and the treatments to be compared. For each permutation, a treatment comparison statistic is computed, forming a null reference distribution. The proportion of reference statistics equal to or more extreme than that computed from the original data is the P value used to test the null hypothesis. If the number of permissible permutations is prohibitively large, the P value is computed from a large random sample of the permutations.

Using a simple 2×2 example with $n = 2$ per cell, if the original data were as in Table 1, then one permutation of the scores among the four treatment combinations would be as in Table 2.

Permutation tests rely on the original sampling, so the permutation illustrated above would be appropriate for a completely randomized two-way design. As the random sample of subjects was randomly assigned two to each cell, any permutation of the results, two scores to each cell would be permissible. Observations may be exchanged between any pair of cells. For randomized block designs, however, the only permissible permutations would be those in

Table 1 Original data

1	3
2	4
5	7
6	8

Table 2 Permutation of data

3	1
2	7
8	4
6	5

Table 3 Bootstrap sample of data

3	1
1	7
8	6
8	5

which scores are exchanged between treatments but within the same block. Software for permutation tests includes [5, 12, 14, 16, and 17].

Bootstrap Tests

Bootstrap tests follow a pattern similar to permutation tests except that the empirical distribution is based on B random samples from the original data rather than all possible rearrangements of the data. Put another way, the bootstrap uses sampling with replacement. So from the original data above, a possible sample with replacement would be as in Table 3.

A generally useful taxonomy is given by [13] for some of these methods that use an empirically generated sampling distribution. Consider these methods as resampling methods, and define the sampling frame from which the sample is to be selected as the set of scores that was actually observed. Then, resampling can be done without replacement (the permutation test) or with replacement (the bootstrap). Both of these use the original sample size, n , as the sample size for the resample (using a subsample of size $m < n$ leads to the **jackknife**).

The percentile bootstrap method for the two-way **factorial design** is succinctly covered in [21, p. 350]. For broader coverage, [6] and [10] give book-length coverage of the topic, including the percentile bootstrap method. Software for bootstrap tests includes [12], and macros from www.sas.com [15].

Rank Transform

The concept behind the rank transform is to substitute ranks for the original observations and compute the usual statistic (*see Rank Based Inference*). While this concept is simple and intuitively appealing, and while it works in some simpler settings (including the correlation, the two-independent-sample case, and the one-way design), it has some problems when

applied to the two-way factorial design. Notably, the ranks are not independent, resulting in F tests that do not maintain α -control. Put another way, the linear model and the F tests are not invariant under the rank transformation. Ample evidence exists that the rank transform should not be considered for the two-way factorial design [2, 18, 19, and 20] but it persists in the literature because of suggestions like that in [15].

Indeed, documentation for the SAS programming language [15] suggest that a wide range of **linear model** hypotheses can be tested nonparametrically by taking ranks of the data (using the RANK procedure) and using a regular parametric procedure (such as GLM or ANOVA) to perform the analysis. It is likely that these tests are as suspect in the wider context of linear models as for the $J \times K$ factorial design.

Other authors [11] have proposed analogous rank-based tests that rely on chi-squared distributions, rather than those of the t and F random variables. However, [20] establishes that these methods have problems similar to those that employ F tests.

Aligned Ranks Tests

When one or more of the estimated sources in the linear model are first subtracted from the scores, and the subsequent residuals are then ranked, the scores are said to have been aligned, and the tests are called *aligned ranks tests* [9]. Results for the two-way factorial design show problems with liberal α for some F tests in selected factorial designs with cell sizes of 10 or fewer [20]. However, results from other designs, such as the factorial **Analysis of Covariance (ANCOVA)**, are promising for larger cell sizes, that is, 20 or more [8].

Tests on Trimmed Means

For a method that is based on robust estimators, tests on 20% **trimmed means** are an option. For the ordered scores in each cell, remove 20% of both the largest and smallest scores, and average the remaining 60% of the scores, yielding a 20% trimmed mean. For example, if $n = 10$, then $0.2(n) = 0.2(10) = 2$. If the data for one cell are

23 24 26 27 34 35 38 45 46 56,

then the process to get the 20% trimmed mean would be to remove the 23 and 24 and the 46 and 56. Then

the 20% trimmed mean would sum 26 through 45 to get 205, then divide by $0.6n$ to get $205/6 = 34.17$.

The numerators of the test statistics for the main effects and interactions are functions of these trimmed means. The denominators of the test statistics are functions of 20% Winsorized variances (*see Winsorized Robust Measures*).

To obtain Winsorized scores for the data in each cell, instead of removing the 20% largest and smallest scores, replace the 20% smallest scores with the next score up in order, and replace the 20% largest scores with the next score down in order. Then compute the unbiased sample variance on these Winsorized scores.

For the data above, the 20% Winsorized scores would be

26 26 26 27 34 35 38 45 45 45

and the 20% Winsorized variance would be obtained by computing the unbiased sample variance of these $n = 10$ Winsorized scores, yielding 68.46.

For complete details on heteroscedastic methods using 20% trimmed means and 20% Winsorized variances, see [21]. While t Tests or F tests based on trimmed means and Winsorized variances are not nonparametric methods, they are robust to nonnormality and unequal variances, and provide an alternative to the two-way ANOVA. Of course, when combined with, say, bootstrap sampling distributions, tests based on trimmed means take on the additional property of being nonparametric.

BDM

Heteroscedastic nonparametric F tests for factorial designs that allow for tied values and test hypotheses of equality of distributions were developed by [4]. This method, called the BDM method (after the authors of [4]), is based on F distributions with estimated degrees of freedom generalized from [3]. All of the N observations are pooled and ranked, with tied ranks resolved by the mid-rank solution where the tied observations are given the average of the ranks among the tied values. Computation of the subsequent ANOVA-type rank statistics (BDM tests) is shown in [21, p. 572]. Simulations by [4] showed these BDM tests to have adequate control of α and competitive power.

Hypotheses

The hypotheses tested by the factorial ANOVA for the A main effect, B main effect, and interaction, are, respectively,

$$H_0: \alpha_j = \mu_j - \mu = 0 \text{ for all } j = 1 \text{ to } J$$

$$H_0: \alpha_k = \mu_k - \mu = 0 \text{ for all } k = 1 \text{ to } K$$

$$H_0: \alpha\beta_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu = 0 \text{ for all } j = 1 \text{ to } J, \text{ for all } k = 1 \text{ to } K. \quad (1)$$

Some of the rank procedures might claim to test hypotheses where rank mean counterparts are substituted for the appropriate μ 's in the above hypotheses, but in reality they test truly nonparametric hypotheses. Such hypotheses are given as a function of the cumulative distribution for each cell, $F_{jk}(x)$, see [1]. $F_{j.}$ is the average of the $F_{jk}(x)$ across the K levels of B, $F_{.k}$ is the average of the $F_{jk}(x)$ across the J levels of A, and $F_{..}$ is the average of the $F_{jk}(x)$ across the JK cells. Then the hypotheses tested by these nonparametric methods for the A main effect, B main effect, and interaction, are, respectively,

$$H_0: A_j = F_{j.} - F_{..} = 0 \text{ for all } j = 1 \text{ to } J$$

$$H_0: B_k = F_{.k} - F_{..} = 0 \text{ for all } k = 1 \text{ to } K$$

$$H_0: AB_{jk} = F_{jk}(x) - F_{j.} - F_{.k} + F_{..} = 0 \text{ for all } j = 1 \text{ to } J, \text{ for all } k = 1 \text{ to } K. \quad (2)$$

So the permutation tests, bootstrap tests, aligned ranks tests, and BDM all test this last set of hypotheses. The tests based on trimmed means test hypotheses about population trimmed means, unless, of course, they are used in the bootstrap. Practically, the nonparametric tests allow a researcher to say that the distributions are different, without specifying a particular parameter for the difference.

References

- [1] Akritas, M.G., Arnold, S.F. & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of the American Statistical Association* **92**, 258–265.
- [2] Blair, R.C., Sawilowsky, S.S. & Higgins, J.J. (1987). Limitations of the rank transform statistic in tests for interaction, *Communications in Statistics-Simulation and Computation* **16**, 1133–1145.

4 Two-way Factorial: Distribution-Free Methods

- [3] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: effect of inequality of variance in the one-way classification, *The Annals of Mathematical Statistics* **25**, 290–302.
- [4] Brunner, E., Dette, H. & Munk, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association* **92**, 1494–1502.
- [5] Edgington, E.S. (1987). *Randomization Tests*, Marcel Dekker, New York.
- [6] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.
- [7] Headrick, T.C. & Sawilowsky, S.S. (2000). Properties of the rank transformation in factorial analysis of covariance, *Communications in Statistics: Simulation and Computation* **29**, 1059–1088.
- [8] Headrick, T.C. & Vineyard, G. (2001). An empirical investigation of four tests for interaction in the context of factorial analysis of covariance, *Multiple Linear Regression Viewpoints* **27**, 3–15.
- [9] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*, Wiley, New York.
- [10] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Grove.
- [11] Puri, M.L. & Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*, Wiley, New York.
- [12] Resampling Stats [Computer Software]. (1999). Resampling Stats, Arlington.
- [13] Rodgers, J. (1999). The bootstrap, the jackknife, and the randomization test: a sampling taxonomy, *Multivariate Behavioral Research* **34**(4), 441–456.
- [14] RT [Computer Software]. (1999). West, Cheyenne.
- [15] SAS [Computer Software]. (1999). SAS Institute, Cary.
- [16] StatXact [Computer Software]. (1999). Cytel Software, Cambridge.
- [17] Testimate [Computer Software]. (1999). SciTech, Chicago.
- [18] Thompson, G.L. (1991). A note on the rank transform for interactions, *Biometrika* **78**, 697–701.
- [19] Thompson, G.L. (1993). A correction note on the rank transform for interactions, *Biometrika* **80**, 711.
- [20] Toothaker, L.E. & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA, *Journal of Educational and Behavioral Statistics* **19**, 237–273.
- [21] Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*, Academic Press, San Diego.

LARRY E. TOOTHAKER

Type I, Type II and Type III Sums of Squares

SCOTT L. HERSHBERGER

Volume 4, pp. 2089–2092

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Type I, Type II and Type III Sums of Squares

Introduction

In structured experiments to compare a number of treatments, two factors are said to be *balanced* if each level of the first factor occurs as frequently at each level of the second. One example of this is a **randomized block design**, where each level of J treatments occurs only once in each of the K blocks; thus, every possible block-treatment combination occurs once. In balanced designs, blocks and treatments are *orthogonal* factors.

Orthogonal factors are linearly independent: When graphed in n -dimensional space, orthogonal factors lie at a right angle (90°) to one another. As in other linearly independent relationships, orthogonal factors are often, but not always, uncorrelated with one another [4]. Orthogonal factors become correlated if, upon centering the factors, the angle between the factors deviates from 90° .

In balanced designs, (1) least-squares estimates (see **Least Squares Estimation**) of treatment parameters (each of the factor's levels) are simply the contrast of the levels' means, and (2) the sum of squares for testing the main effect of A depends only on the means of the factor's levels and does not involve elimination of blocks [3]. Property (2) implies that each level of the blocking variable contributes equally to the estimation of the main effect. Whenever properties (1) and (2) are true, the factor and blocking variable are orthogonal, and as a result, the factor's main effect is estimated independently of the blocking variable. Similarly, in a two-way ANOVA design in which cells have equal numbers of observations and every level of factor A is crossed once with every level of factor B , the A and B main effects are independently estimated, neither one affecting the other (see **Factorial Designs**).

Properties (1) and (2) also hold in another type of orthogonal design, the **Latin square**. In an $n \times n$ Latin square, each of the n treatments occurs only once in each row and once in each column. Here, treatments are orthogonal to both rows and columns. Indeed, rows and columns are themselves orthogonal. Thus, when we say that a Latin square design is an

orthogonal design, we mean that it is orthogonal for the estimation of row, column, and main effects.

In *nonorthogonal* designs, the estimation of the main effect of one factor is determined in part by the estimation of the main effects of other factors. Nonorthogonal factors occur when the combination of their levels is *unbalanced*. Generally speaking, unbalanced designs arise under one of two circumstances: either one or more factor level combinations are missing because of the complete absence of observations for one or more cells or factor level combinations vary in number of observations.

Whenever nonorthogonality is present, the effects in an experiment are **confounded** (yoked). Consider the two 3×3 designs below in which each cell's sample size is given.

Design I					
		B			
		<u>1</u>	<u>2</u>		<u>3</u>
A	<u>1</u>	5	5	5	
	<u>2</u>	5	5		5
	<u>3</u>	5	5		0
Design II					
		B			
		<u>1</u>	<u>2</u>		<u>3</u>
A	<u>1</u>	5	5	5	
	<u>2</u>	5	5		5
	<u>3</u>	5	5	2	

In both designs, factors A and B are confounded, and thus, nonorthogonal. In Design I, nonorthogonality is due to the zero frequency of cell a_3b_3 : (see **Structural Zeros**) main effect B (i.e., comparing levels B_1 , B_2 , and B_3) is evaluated by collapsing within each level of B rows A_1 , A_2 , and A_3 . However, while A_1 , A_2 , and A_3 are available to be collapsed within B_1 and B_2 , only A_1 and A_2 are available within B_3 . As a result, the main effect hypothesis for factor B cannot be constructed so that its marginal means are based on cell means that have all of the same levels of factor A . Consequently, the test of B 's marginal means is confounded and dependent on factor A 's marginal means. Similarly, the test of A 's marginal means is confounded and dependent on factor B 's marginal means. In Design II, factors A and B are confounded because of the smaller sample size of cell a_3b_3 : the main effect hypothesis for factor B

2 Type I, Type II and Type III Sums of Squares

(and A) cannot be constructed so that each of factor A 's (and B 's) levels are weighted equally in testing the B (and A) main effect.

Sum of Squares

In an orthogonal or balanced ANOVA, there is no need to worry about the decomposition of sums of squares. Here, one ANOVA factor is independent of another ANOVA factor, so a test for, say, a sex effect is independent of a test for, say, an age effect. When the design is unbalanced or nonorthogonal, there is not a unique decomposition of the sums of squares. Hence, decisions must be made to account for the dependence between the ANOVA factors in quantifying the effects of any single factor. The situation is mathematically equivalent to a multiple regression model where there are correlations among the predictor variables. Each variable has direct and indirect effects on the dependent variable. In an ANOVA, each factor will have direct and indirect effects on the dependent variable. Four different types of sums of squares are available for the estimation of factor effects. In an orthogonal design, all four will be equal. In a nonorthogonal design, the correct sums of squares will depend upon the logic of the design.

Type I SS

Type I SS are order-dependent (hierarchical). Each effect is adjusted for all other effects that appear earlier in the model, but not for any effects that appear later in the model. For example, if a three-way ANOVA model was specified to have the following order of effects,

$$A, B, A \times B, C, A \times C, B \times C, A \times B \times C,$$

the sums of squares would be calculated with the following adjustments:

Effect	Adjusted for
A	—
B	A
$A \times B$	A, B
C	$A, B, A \times B$
$A \times C$	$A, B, C, A \times B$
$B \times C$	$A, B, C, A \times B, A \times C$
$A \times B \times C$	$A, B, C, A \times B, A \times C, A \times B \times C$

Type I *SS* are computed as the decrease in the error *SS* (*SSE*) when the effect is added to a model. For example, if *SSE* for $Y = A$ is 15 and *SSE* for $Y = A \times B$ is 5, then the Type I *SS* for B is 10. The sum of all of the effects' *SS* will equal the total model *SS* for Type I *SS* – this is not generally true for the other types of *SS* (which exclude some or all of the variance that cannot be unambiguously allocated to one and only one effect). In fact, specifying effects hierarchically is the only method of determining the unique amount of variance in a dependent variable explained by an effect. Type I *SS* are appropriate for balanced (orthogonal, equal n) analyses of variance in which the effects are specified in proper order (main effects, then two-way interactions, then three-way interactions, etc.), for trend analysis where the powers for the quantitative factor are ordered from lowest to highest in the model statement, and the **analysis of covariance** (ANCOVA) in which covariates are specified first. Type I *SS* are also used for hierarchical step-down nonorthogonal analyses of variance [1] and hierarchical regression [2]. With such procedures, one obtains the particular *SS* needed (adjusted for some effects but not for others) by carefully ordering the effects. The order of effects is usually based on temporal priority, on the causal relations between effects (an outcome should not be added to the model before its cause), or on theoretical importance.

Type II SS

Type II SS are the reduction in the *SSE* as a result of adding the effect to a model that contains all other effects except effects that contain the effect being tested. An effect is contained in another effect if it can be derived by deleting terms in that effect – for example, $A, B, C, A \times B, A \times C$, and $B \times C$ are all contained in $A \times B \times C$. The Type II *SS* for our example involve the following adjustments:

Effect	Adjusted for
A	$B, C, B \times C$
B	$A, C, A \times C$
$A \times B$	$A, B, C, A \times C, B \times C$
C	$A, B, A \times B$
$A \times C$	$A, B, C, A \times B, B \times C$
$B \times C$	$A, B, C, A \times B, A \times C$
$A \times B \times C$	$A, B, C, A \times B, A \times C, B \times C$

When the design is balanced, Type I and Type II *SS* are identical.

Type III *SS*

Type III *SS* are identical to those of Type II *SS* when the design is balanced. For example, the sum of squares for *A* is adjusted for the effects of *B* and for $A \times B$. When the design is unbalanced, these are the *SS* that are approximated by the traditional unweighted means ANOVA that uses **harmonic mean** sample sizes to adjust cell totals: Type III *SS* adjusts the sums of squares to estimate what they might be if the design were truly balanced. To illustrate the difference between Type II and Type III *SS*, consider factor *A* to be a dichotomous variable such as gender. If the data contained 60% females and 40% males, the Type II sums of squares are based on those percentages. In contrast, the Type III *SS* assume that the sex difference came about because of sampling and tries to generalize to a population in which the number of males and females is equal.

Type IV *SS*

Type IV *SS* differ from Types I, II, and III *SS* in that it was developed for designs that have one or more

empty cells, that is, cells that contain no observations. (see **Structural Zeros**) Type IV *SS* evaluate marginal means that are based on equally weighted cell means. They yield the same results as a Type III *SS* if all cells in the design have at least one observation. As a result, with Type IV *SS*, marginal means of one factor are based on cell means that have all of the same levels of the other factor, avoiding the confounding of factors that would occur if cells were empty.

References

- [1] Applebaum, M.I. & Cramer, E.M. (1974). Some problems in the nonorthogonal analysis of variance, *Psychological Bulletin* **81**, 335–343.
- [2] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*, 2nd Edition. Erlbaum, Hillsdale.
- [3] Milliken, G.A. & Johnson, D.E. (1984). *Analysis of Messy Data*, Vol. 1, Wadsworth, Belmont.
- [4] Rodgers, J.L., Nicewander, W.A. & Toothaker, L. (1984). Linearly independent, orthogonal, and uncorrelated variables, *American Statistician* **38**, 133–134.

SCOTT L. HERSHBERGER

Ultrametric Inequality

FIONN MURTAGH

Volume 4, pp. 2093–2094

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ultrametric Inequality

The triangular inequality holds for a metric space: $d(x, z) \leq d(x, y) + d(y, z)$ for any triplet of points x, y, z . In addition, the properties of symmetry ($d(x, y) = d(y, x)$) and positive definiteness ($d(x, y) \geq 0$ with $x = y$ if $d(x, y) = 0$) are respected.

The ‘strong triangular inequality’ or ultrametric inequality is $d(x, z) \leq \max \{d(x, y), d(y, z)\}$ for any triplet x, y, z .

An ultrametric space implies respect for a range of stringent properties. For example, the triangle formed by any triplet is necessarily isosceles, with the two large sides equal, or is equilateral.

Consider the dissimilarity data shown on the left side of Table 1. (For instance, this could be the similarity of performance between five test subjects on a scale of 1 = very similar, to 9 = very dissimilar.) The single link criterion was used to construct the dendrogram shown (Figure 1). On the right side of Table 1, the ultrametric distances defined from the sequence of agglomeration values are given.

Among the ultrametric distances, consider for example, $d(2, 3)$, $d(3, 4)$, and $d(4, 1)$. We see that $d(2, 3) \leq \max \{d(3, 4), d(4, 1)\}$ since here we have $5 \leq \max \{5, 4\}$. We can turn around any way we like what we take as x, y, z but with ultrametric distances, we will always find that $d(x, z) \leq \max \{d(x, y), d(y, z)\}$.

Table 1 Left: original pairwise dissimilarities. Right: derived ultrametric distances

	1	2	3	4	5	1	2	3	4	5
1	0	4	9	5	8	0	4	5	4	5
2	4	0	6	3	6	4	0	5	3	5
3	9	6	0	6	3	5	5	0	5	3
4	5	3	6	0	5	4	3	5	0	5
5	8	6	3	5	0	5	5	3	5	0

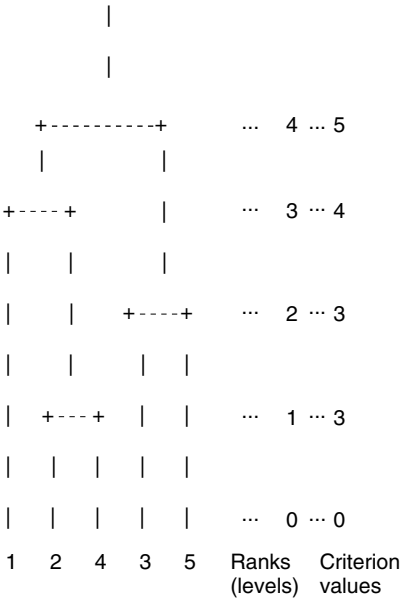


Figure 1 Resulting dendrogram

In data analysis, the ultrametric inequality is important because a **hierarchical clustering** is tantamount to defining ultrametric distances on the objects under investigation. More formally, we say that in clustering, a bijection is defined between a rooted, binary, ranked, indexed tree, called a *dendrogram* (see **Hierarchical Clustering**), and a set of ultrametric distances ([1], representing work going back to the early 1960s; [2]).

References

[1] Benzécri, J.P. (1979). *La Taxinomie*, 2nd Edition, Dunod, Paris.
 [2] Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika* **32**, 241–254.

FIONN MURTAGH

Ultrametric Trees

FIONN MURTAGH

Volume 4, pp. 2094–2095

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Ultrametric Trees

Let us call the row of a data table an observation vector. In the following, we will consider the case of n rows. The set of rows will be denoted I .

Hierarchical agglomeration on n observation vectors, $i \in I$, involves a series of pairwise agglomerations of observations or clusters (*see Hierarchical Clustering*). In Figure 1 we see that observation vectors x_1 and x_2 are first agglomerated, followed by the incorporation of x_3 , and so on. Since an agglomeration is always pairwise, we see that there are precisely $n - 1$ agglomerations for n observation vectors.

We will briefly look at three properties of a hierarchical clustering tree (referred to as conditions (i), (ii) and (iii)). A hierarchy, H , is a set of sets, q . Each q is a subset of I . By convention, we include I itself as a potential subset, but we do not include the empty set. Our input data (observation vectors), denoted $x_1, x_2, \dots$ are also subsets of I . All subsets of I are collectively termed the power set of I . The notation sometimes used for the power set of I is 2^I . Condition (i) is that the hierarchy includes the set I , and this corresponds to the top (uppermost) level in Figure 1. Condition (ii) requires that $x_1, x_2, \dots$ are in H , and this corresponds to the bottom (lowermost) level in Figure 1. Finally, condition (iii) is that different subsets do not intersect unless one is a member of the other that is, they do not overlap other than through 100% inclusion of one in the other.

In Figure 1, the observation set $I = \{x_1, x_2, \dots, x_8\}$. The $x_1, x_2, \dots$ are called *singleton sets* (as

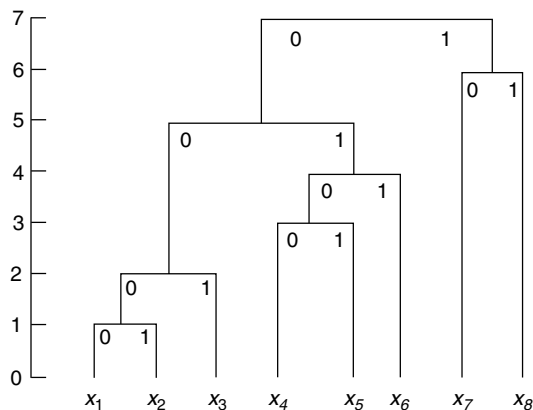


Figure 1 Labeled, ranked dendrogram on 8 terminal nodes. Branches labeled 0 and 1

opposed to clusters), and are associated with terminal nodes in the dendrogram tree.

A dendrogram, displayed in Figure 1, is the name given to a binary, ranked tree, which is a convenient representation for the set of agglomerations (*see Hierarchical Clustering*).

Now we will show the link with the **ultrametric inequality**. An indexed hierarchy is the pair (H, ν) where the positive function defined on H , that is, $\nu : H \rightarrow \mathbb{R}^+$, ($\mathbb{R}^+$ denotes the set of positive real numbers) satisfies $\nu(i) = 0$ if $i \in H$ is a singleton; and $q \subset q' \implies \nu(q) < \nu(q')$. Function ν is the agglomeration level. Typically the index, or agglomeration level, is defined from the succession of agglomerative values that are used in the creation of the dendrogram. In practice, this means that the hierarchic clustering algorithm used to produce the dendrogram representation yields the values of ν .

The distance between any two clusters is based on the 'common ancestor', that is, how high in the dendrogram we have to go to find a cluster that contains both. Take $q \subset q'$, let $q \subset q''$ and $q' \subset q''$, and let q'' be the lowest level cluster for which this is true. Then if we define $D(q, q') = \nu(q'')$, it can be shown that D is an ultrametric.

In practice, we start with our given data, and a Euclidean or other dissimilarity, we use some compactness agglomeration criterion such as minimizing the change in variance resulting from the agglomerations, and then define $\nu(q)$ as the dissimilarity associated with the agglomeration carried out.

The standard agglomerative hierarchical clustering algorithm developed in the early 1980s is based on the construction of nearest neighbor chains, followed by agglomeration of reciprocal nearest neighbors. For a survey, see Murtagh [1, 2].

References

- [1] Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms, *The Computer Journal* **26**, 354–359.
- [2] Murtagh, F. (1985). *Multidimensional Clustering Algorithms*, COMPSTAT Lectures Volume 4, Physica-Verlag, Vienna.

Further Reading

Benzécri, J.P. (1976). *L'Analyse des Données. Tome 1. La Taxinomie*, 2nd Edition, Dunod, Paris.

FIONN MURTAGH

Unidimensional Scaling

JAN DE LEEUW

Volume 4, pp. 2095–2097

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Unidimensional Scaling

Unidimensional scaling is the special one-dimensional case of **multidimensional scaling**. It is often discussed separately, because the unidimensional case is quite different from the general multidimensional case. It is applied in situations in which we have a strong reason to believe there is only one interesting underlying dimension, such as time, ability, or preference. In the unidimensional case, we do not have to choose between different metrics, such as the Euclidean metric, the City Block metric, or the Dominance metric (*see Proximity Measures*). Unidimensional scaling techniques are very different from multidimensional scaling techniques, because they use very different algorithms to minimize their loss functions.

The classical form of unidimensional scaling starts with a symmetric and nonnegative matrix $\Delta = \{\delta_{ij}\}$ of *dissimilarities* and another symmetric and nonnegative matrix $W = \{w_{ij}\}$ of *weights*. Both W and Δ have a zero diagonal. Unidimensional scaling finds *coordinates* x_i for n points on the line such that

$$\sigma(x) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (\delta_{ij} - |x_i - x_j|)^2 \quad (1)$$

is minimized. Those n coordinates in x define the scale we were looking for.

To analyze this unidimensional scaling problem in more detail, let us start with the situation in which we know the order of the x_i , and we are just looking for their scale values. Now $|x_i - x_j| = s_{ij}(x)(x_i - x_j)$, where $s_{ij}(x) = \text{sign}(x_i - x_j)$. If the order of the x_i is known, then the $s_{ij}(x)$ are known numbers, equal to either -1 or $+1$ or 0 , and thus our problem becomes minimization of

$$\sigma(x) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (\delta_{ij} - s_{ij}(x_i - x_j))^2 \quad (2)$$

over all x such that $s_{ij}(x_i - x_j) \geq 0$. Assume, without loss of generality, that the weighted sum of squares of the dissimilarities is one. By expanding the sum of squares we see that

$$\sigma(x) = 1 - t'Vt + (x - t)'V(x - t). \quad (3)$$

Here V is the matrix with off-diagonal elements $v_{ij} = -w_{ij}$ and diagonal elements $v_{ii} = \sum_{j=1}^n w_{ij}$.

Also, $t = V^+r$, where r is the vector with elements $r_i = \sum_{j=1}^n w_{ij}\delta_{ij}s_{ij}$, and V^+ is a generalized inverse of V . If all the off-diagonal weights are equal, we simply have $t = r/n$.

Thus, the unidimensional scaling problem, with a known scale order, requires us to minimize $(x - t)'V(x - t)$ over all x , satisfying the order restrictions. This is a **monotone regression** problem, possibly with a nondiagonal weight matrix, which can be solved quickly and uniquely by simple quadratic programming methods.

Now for some geometry. The vectors x satisfying the same set of order constraints form a polyhedral convex cone $\mathcal{K}$ in $\mathbb{R}^n$. Think of $\mathcal{K}$ as an ice-cream cone with its apex at the origin, except for the fact that the ice-cream cone is not round, but instead bounded by a finite number of hyperplanes. Since there are $n!$ different possible orderings of x , there are $n!$ cones, all with their apex at the origin. The interior of the cone consists of the vectors without ties, intersections of different cones are vectors with at least one tie. Obviously, the union of the $n!$ cones is all of $\mathbb{R}^n$.

Thus, the unidimensional scaling problem can be solved by solving $n!$ monotone regression problems, one for each of the $n!$ cones [2]. The problem has a solution, which is at the same time very simple and prohibitively complicated. The simplicity comes from the $n!$ subproblems, which are easy to solve, and the complications come from the fact that there are simply too many different subproblems. Enumeration of all possible orders is impractical for $n > 10$, although using combinatorial programming techniques makes it possible to find solutions for n as large as 30 [6].

Actually, the subproblems are even simpler than we suggested above. The geometry tells us that we solve the subproblem for cone $\mathcal{K}$ by finding the closest vector to t in the cone, or, in other words, by projecting t on the cone. There are three possibilities. Either t is in the interior of its cone, or on the boundary of its cone, or outside its cone. In the first two cases, t is equal to its projection, in the third case, the projection is on the boundary. The general result in [1] tells us that the loss function σ cannot have a local minimum at a point in which there are ties, and, thus, local minima can only occur in the interior of the cones. This means that we can only have a local minimum if t is in the interior of its cone [4], and it also means that we actually never have to compute the monotone regression. We just have to verify if t

2 Unidimensional Scaling

is in the interior, if it is not, then σ does not have a local minimum in this cone.

There have been many proposals to solve the combinatorial optimization problem of moving through the $n!$ cones until the global optimum of σ has been found. A recent review is [7].

We illustrate the method with a simple example, using the vegetable paired-comparison data from [5, p. 160]. Paired comparison data are usually given in a matrix P of proportions, indicating how many times stimulus i is preferred over stimulus j . P has 0.5 on the diagonal, while corresponding elements p_{ij} and p_{ji} on both sides of the diagonal add up to 1.0. We transform the proportions to dissimilarities by using the probit transformation $z_{ij} = \Phi^{-1}(p_{ij})$ and then defining $\delta_{ij} = |z_{ij}|$. There are nine vegetables in the experiment, and we evaluate all $9! = 362\,880$ permutations. Of these cones, 14 354 or 4% have a local minimum in their interior. This may be a small percentage, but the fact that σ has 14 354 isolated local minima indicates how complicated the unidimensional scaling problem is. The global minimum is obtained for the order given in Guilford's book, which is Turnips < Cabbage < Beets < Asparagus < Carrots < Spinach < String Beans < Peas < Corn. Since there are no weights in this example, the optimal unidimensional scaling values are the row averages of the matrix with elements $s_{ij}(x)\delta_{ij}$. Except for a single sign change of the smallest element (the Carrots and Spinach comparison), this matrix is identical to the probit matrix Z . And because the Thurstone Case V scale values are the row averages of Z , they are virtually identical to the unidimensional scaling solution in this case.

The second example is quite different. It has weights and incomplete information. We take it from an early paper by Fisher [3], in which he studies crossover percentages of eight genes on the sex chromosome of *Drosophila willistoni*. He takes the crossover percentage as a measure of distance, and supposes the number n_{ij} of crossovers in N_{ij} observations is binomial. Although there are eight genes, and thus $\binom{8}{2} = 28$ possible dissimilarities, there are only 15 pairs that are actually observed. Thus, 13 of the off-diagonal weights are zero, and the other weights are set to the inverses of the standard errors of the proportions. We investigate all $8! = 40\,320$ permutations, and we find 78 local minima. The solution given by **Fisher**, computed by solving linearized likelihood equations, has Reduced

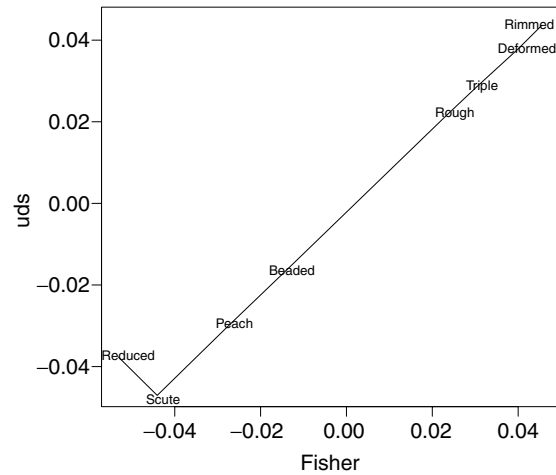


Figure 1 Genes on Chromosome

< Scute < Peach < Beaded < Rough < Triple < Deformed < Rimmed. This order corresponds with a local minimum of σ equal to 40.16. The global minimum is obtained for the permutation that interchanges Reduced and Scute, with value 35.88. In Figure 1, we see the scales for the two local minima, one corresponding with Fisher's order and the other one with the optimal order.

In this entry, we have only discussed least squares metric unidimensional scaling. The first obvious generalization is to replace the least squares loss function, for example, by the least absolute value or $\mathcal{L}_1$ loss function. The second generalization is to look at nonmetric unidimensional scaling. These generalizations have not been studied in much detail, but in both, we can continue to use the basic geometry we have discussed. The combinatorial nature of the problem remains intact.

References

- [1] De Leeuw, J. (1984). Differentiability of Kruskal's stress at a local minimum, *Psychometrika* **49**, 111–113.
- [2] De Leeuw, J. & Heiser, W.J. (1977). Convergence of correction matrix algorithms for multidimensional scaling, in *Geometric Representations of Relational Data*, J.C., Lingoes, ed., Mathesis Press, Ann Arbor, pp. 735–752.
- [3] Fisher, R.A. (1922). The systematic location of genes by means of crossover observations, *American Naturalist* **56**, 406–411.
- [4] Groenen, P.J.F. (1993). The majorization approach to multidimensional scaling: some problems and extensions, Ph.D. thesis, University of Leiden.

- [5] Guilford, J.P. (1954). *Psychometric Methods*, 2nd Edition, McGraw-Hill.
- [6] Hubert, L.J. & Arabie, P. (1986). Unidimensional scaling and combinatorial optimization, in *Multidimensional Data Analysis*, J. De Leeuw, W. Heiser, J. Meulman & F. Critchley, eds, DSWO-Press, Leiden.
- [7] Hubert, L.J., Arabie, P. & Meulman, J.J. (2002a). Linear unidimensional scaling in the L_2 -Norm: basic optimization methods using MATLAB, *Journal of Classification* **19**, 303–328.

JAN DE LEEUW

Urban, F M

HELEN ROSS

Volume 4, pp. 2097–2098

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Urban, F M

Born: December 28, 1878, Brünn, Moravia (now Austria).

Died: May 4, 1964, Paris, France.

F. M. Urban's main claim to fame was the probability weightings that he added to the calculation of the psychometric function. In 1968, he was ranked the 14th most famous psychologist, but is now almost forgotten – so much so that a professor who was equipping a new psychological laboratory reputedly attempted to order a set of Urban weights. There are few biographical sources about him, and even his usual first name is in doubt (see [1]).

Friedrich Johann Victor Urban was born into a German-speaking family in Austro-Hungary. He is known by the initials F. M. because he adopted Friedrich Maria as a pen name, which became Frederick Mary, or perhaps Francis M., in the United States. He studied philosophy at the University of Vienna from 1897, obtaining a Ph.D. in 1902 in aesthetics. He also studied probability under Wilhelm Wirth at the University of Leipzig. He taught psychological acoustics at Harvard University in 1904–1905. His main work on statistics was conducted from 1905 at the University of Pennsylvania, where he rose to assistant professor in 1910. He published in German and English in 1908–1909 [3, 4]. He was elected a member of the American Psychological Association in 1906, the American Association for the Advancement of Science in 1907, and a Fellow of the latter in 1911. Urban's life in America ended in 1914, when he returned home and married Adele Königsgarten. Lacking US citizenship, he was not allowed to return to the United States on the outbreak of war. He moved to Sweden, and spent 1914 to 1917 in Göteborg and Stockholm, where he did research at the Kungliga Vetenskapliga Akademien. He returned to Brünn in 1917, and served in the Austrian army. Moravia became part of the Czechoslovak Republic in 1918. Urban then worked as an insurance statistician, his Czech language ability being inadequate for a university post. He corresponded with his colleagues

abroad, and continued to publish in German and English on psychometry and psychophysics. He also translated into German J. M. Keynes's *A Treatise on Probability* (1926) and J. L. Coolidge's *An Introduction to Mathematical Probability* (1927). He and his Jewish wife stayed in Brünn throughout the Second World War, suffering Hitler's invasion, the bombing of their house, the Russian 'liberation', and the expulsion of German speakers by the returning Czech regime. He was put in a concentration camp first by the Russians and then by the Czechs. He was released owing to pleas from abroad, but was forced to leave Czechoslovakia in 1948. He and his wife went to join their elder daughter first in Norway, and then in Brazil (1949–52), where he lectured on **factor analysis** at São Paulo University. In 1952, they moved in with their younger daughter in France, living in Toulon and, finally, Paris.

Urban's many contributions to statistics are discussed in [2]. The Müller–Urban weights are a refinement to the least squares solution for the psychometric function, when P values are transformed to z values. Müller argued that the proportions near 0.5 should be weighted most, while Urban argued that those with the smallest mean-square error should be weighted most. The combined weightings were laborious to calculate, and made little difference to the final threshold values. The advent of computers and new statistical procedures has relegated these weights to history.

References

- [1] Ertle, J.E., Bushong, R.C. & Hillix, W.A. (1977). The real F.M. Urban, *Journal of the History of the Behavioral Sciences* **13**, 379–383.
- [2] Guilford, J.P. (1954). *Psychometric Methods*, 2nd Edition, McGraw-Hill, New York.
- [3] Urban, F.M. (1908). *The Application of Statistical Methods to Problems of Psychophysics*, Psychological Clinic Press, Philadelphia.
- [4] Urban, F.M. (1909). Die psychophysischen Massmethoden als Grundlagen empirischer Messungen, *Archiv für die Gesamte Psychologie* **15**, 261–355; **16**, 168–227.

HELEN ROSS

Utility Theory

DANIEL READ AND MARA AIROLDI

Volume 4, pp. 2098–2101

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Utility Theory

The term utility is commonly used in two ways. First, following the usage of the utilitarian philosophers, it refers to the total amount of pleasure (or pain) that an action or a good can bring to those affected by it. According to this view, utility is a function of experience. If an apple has more utility than an orange, this means that it brings more pleasure to the person eating it. A second view of utility is that it is a way of representing behaviors, such as choices or expressed preferences, without any reference to the experience arising from what is chosen. For instance, if someone is offered either an apple or an orange, and they choose an apple, we would describe this preference by assigning a higher utility number to the apple than to the orange. This number would say nothing about whether that person would gain more or less pleasure from the apple, but only that they chose the apple. The first use of the term is most closely associated with utilitarian philosophers, pre-twentieth-century economists, and psychologists. The second use is associated with twentieth-century economists.

Daniel Bernoulli is the father of utility theory. He developed the theory to solve the St Petersburg paradox. The paradox revolves around the value placed on the following wager: A fair coin is to be tossed until it comes up heads, at which point those who have paid to play the game will receive 2^n ducats, where n is the number of the toss when heads came up. For instance, if the first head comes up on the third toss the player will receive 8 ducats. Prior to Bernoulli, mathematicians held the view that a gamble was worth its *expected monetary value*, meaning the sum of all possible outcomes multiplied by their probability (*see Expectation*). Yet although the expected monetary value of the St Petersburg wager is infinite:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} 2^n = 1 + 1 + 1 + \dots = \infty, \quad (1)$$

nobody would be willing to pay more than a few ducats for it.

Bernoulli argued that people value this wager according to its expected *utility* rather than its expected *monetary value*, and proposed a specific relationship between utility and wealth: the ‘utility

resulting from any small increase in wealth will be inversely proportional to the quantity of goods previously possessed’ [3]. This implies that the utility of wealth increases logarithmically with additional increments: $u(w) = \log(w)$. With this *utility function*, the St Petersburg paradox was resolved because:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \log(2^n) = 2 \log 2 < \infty. \quad (2)$$

As Menger showed [8], this solution does not hold for all versions of the St Petersburg gamble. For instance, if the payoff is e^{2^n} , the expected utility of the wager is still infinite even with a logarithmic utility function. The paradox can be completely solved only with a bounded utility function. Nonetheless, through this analysis Bernoulli introduced the three major themes of utility theory: first, the same outcome has different utility for different people; second, the relationship between wealth and utility can be described mathematically; third, utility is marginally diminishing, so that the increase in utility from each additional ducat is less than that from the one before.

It is clear that Bernoulli viewed utility as an index of the *benefit* or *good* that a person would get from their income. This view was central to the utilitarians, such as Godwin and Bentham, who considered utility to be a quantity reflecting the disposition to bring pleasure or pain. These philosophers, who wrote in the eighteenth and nineteenth centuries, maintained that the goal of social and personal action is to maximize the sum of the utility of all members of society. Many utilitarian philosophers (and the economists who followed them, such as Alfred Marshall) pointed out that if utility were a logarithmic function of wealth, then transferring wealth from the rich to the poor would maximize the total utility of society. This argument was based on two assumptions that have been difficult to maintain: utility is measurable, and interpersonal comparisons of utility are possible [1, 4].

At the turn of the twentieth century, economists realized that these assumptions were not needed for economic analysis, and that they could get by with a strictly *ordinal* utility function [5, 9]. The idea is that through their choices consumers show which of the many possible allocations of their income they prefer (*see Demand Characteristics*). In general, for any pair of allocations, a consumer will either be indifferent between them, or prefer one to the other.

2 Utility Theory

By obtaining binary choices between allocations, it is possible to draw *indifference curves*. To predict such choices, we just need to know which allocations are on the highest indifference curve. This change in thinking was not only a change in mathematics from interval to ordinal measurement (see **Scales of Measurement; Measurement: Overview**) but also a conceptual revolution. Utility, as used by economists, lost its connection with ‘pleasure and pain’ or other measures of benefit.

One limitation of ordinal utility was that it dealt only with choices under certainty – problems such as the St Petersburg paradox could not be discussed using ordinal utility language. In the mid-twentieth century, Von Neumann and Morgenstern [11] reintroduced the concept of choice under uncertainty with expected utility theory. They showed that just as indifference maps can be drawn from consistent choices between outcomes, so cardinal utility functions (i.e., those measurable on an interval scale) can be derived from consistent choices between gambles or lotteries. Von Neumann and Morgenstern, however, did not reinstate the link between utility and psychology. They did not view utility as a *cause*, but a *reflection*, of behavior: we do not choose an apple over an orange because it has more utility, but it has more utility because we chose it. They showed that if choices between lotteries meet certain formal requirements then preferences can be described by a cardinal utility function. The following assumptions summarize these requirements:

Assumption 1 *Ordering of alternatives.* For each pair of prizes or lotteries an agent will either be indifferent ($\sim$) between them or prefer one to the other ($>$). That is, either $A > B$, $B > A$ or $A \sim B$.

Assumption 2 *Transitivity.* If $A > B$ and $B > C$, then $A > C$.

Assumption 3 *Continuity.* If an agent prefers A to B to C , then there is some probability p which makes a lottery offering A with probability p , and C with probability $(1 - p)$ equal in utility to B . That is $B \sim (p, A; 1 - p, C)$.

Assumption 4 *Substitutability.* Prizes in lotteries can be replaced with lotteries having the same value. Using the outcomes from Assumption 3, the lottery $(q, X; 1 - q, B)$ is equivalent to $(q, X; 1 - q, (p, A; 1 - p, C))$.

Assumption 5 *Independence of irrelevant alternatives.* If two lotteries offer the same outcome with identical probabilities, then the preferences between the lotteries will depend only on the other (unshared) outcomes. If $A > B$, then $(p, A; 1 - p, C) > (p, B; 1 - p, C)$.

Assumption 6 *Reduction of compound lotteries.* A compound lottery is a lottery over lotteries. The reduction of compound lotteries condition means that such a lottery is equivalent to one that has been reduced using the standard rules of probability. For instance, consider the two lotteries, Z and Z' , in Figure 1: Assumption 6 states that $Z \sim Z'$.

If these assumptions hold then a cardinal utility function, unique up to a linear transformation, can be derived from preferences over lotteries. Researchers view these assumptions as ‘axioms of rationality’ because they seem, intuitively, to be how a rational person would behave [10].

One way of doing so is via the *certainty equivalent method* of calculating utilities [12]. In this method, the decision maker first ranks all relevant prizes or outcomes from best to worst. An arbitrary utility value, usually 0 and 1, is given to the worst (W) and best (B) outcome. Then for each intermediate outcome X the decision maker specifies the probability p that would make them indifferent between X for

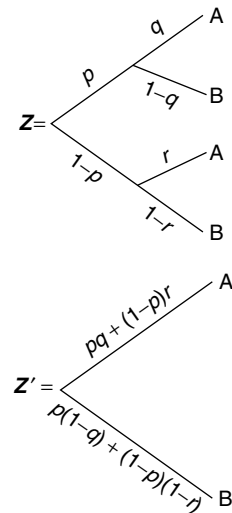


Figure 1 Assumption 6 implies indifference between Z and Z' ($Z \sim Z'$)

sure and the lottery $(p, B; 1 - p, W)$. The utility of X is then found by replacing $u(W) = 0$ and $u(B) = 1$ in the calculation:

$$u(X) = pu(B) + (1 - p) u(W) = p. \quad (3)$$

If different values had been assigned to $u(W)$ and $u(B)$, the resultant utility scale would be a linear transformation of this one.

A further development of utility theory is subjective expected utility theory, which incorporates subjective probabilities. Ramsey, **de Finetti** and **Savage** [10] showed that probabilities as well as utilities could be axiomatized, with choices that reflect probabilities being used to derive a subjective utility function.

Current Thinking About Utility

Neither expected utility nor subjective expected utility theory has proved to be a good *descriptive* theory of choice. An early challenge came from Maurice Allais [2] who showed that independence of irrelevant alternatives could be routinely violated. The challenge, known as ‘Allais Paradox’, shows that the *variance* of the probability distribution of a gamble affects preferences. In particular there is strong empirical evidence that subjects attach additional psychological value or utility to an outcome if it has *zero* variance (*certainty effect* [6]). Allais rejected the idea that behaviors inconsistent with the standard axioms are irrational and claimed that ‘rationality can be defined experimentally by observing the actions of people who can be regarded as acting in a rational manner’ [2]. One consequence of this has been a number of *non-expected utility* theories that either relax or drop some of the assumptions, or else substitute them with empirically derived generalizations. The most influential of these is *prospect theory* [6], which is based on observations of how decision makers actually behave. It differs from expected utility theory in that the probability function is replaced by a nonlinear *weighting function*, and the utility function with a *value function*. The weighting function puts too much weight on low probabilities, too little on moderate to high probabilities, and has a discontinuity for changes from certainty to uncertainty. The value function is defined over deviations from current wealth. It is concave for increasing gains, and convex for losses. This leads to the *reflection effect*, which

means that if a preference pattern is observed for gains, the opposite pattern will be found for losses. For example, people are generally risk averse for gains (\$10 for sure is preferred to a 50/50 chance of \$20 or nothing), but risk seeking for losses (a 50/50 chance of losing \$20 is preferred to a sure loss of \$10).

One issue that is currently the focus of much attention is whether it is possible to find a measure of utility that reflects happiness or pleasure, as originally envisaged by the utilitarians. People often choose options that appear objectively ‘bad’ for them, such as smoking or procrastinating, yet if utility is derived from choice behavior it means the utility of these bad options is greater than that of options that seem objectively better. If we are interested in questions of welfare, there is a practical need for a measure of utility that would permit us to say that ‘ X is better than Y , because it yields more utility’. One suggestion comes from Daniel Kahneman, who argues for a distinction between ‘experienced utility’ (the pain or pleasure from an outcome, the definition adopted by utilitarians), and ‘decision utility’ (utility as reflected in decisions, the definition adopted by modern economists) [7]. Through such refinements in the definition of utility, researchers like Kahneman aspire to reintroduce Bernoulli and Bentham’s perspective to the scientific understanding of utility.

Acknowledgment

The writing of this article was assisted by funding from the Leverhulme Foundation Grant F/07004/Y.

References

- [1] Alchian, A. (1953). The meaning of utility measurement, *The American Economic Review* **43**, 26–50.
- [2] Allais, M. (1953). Le comportement de l’homme rationnel devant le risqué, critique des postulats et axiomes de l’école Américaine, *Econometrica* **21**, 503–546; English edition: Allais, M. & Hagen O. eds. (1979). *Expected Utility Hypothesis and the Allais’ Paradox*, Reidel Publishing, Dordrecht, 27–145.
- [3] Bernoulli, D. (1738). *Specimen Theoriae Novae de Mensura Sortis*, in *Commentarii of Sciences in Petersburg*, Vol. V, 175–192; English edition (1954). Exposition of a new theory on the measurement of risk, *Econometrica* **22**(1), 23–36.

4 Utility Theory

- [4] Ellsberg, D. (1954). Classic and current notions of “measurable utility”, *The Economic Journal* **64**, 528–556.
- [5] Fisher, I. (1892). Mathematical investigations in the theory of value and prices, *Transaction of the Connecticut Academy of Art and Science* **IX**(1), 1–124.
- [6] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–291.
- [7] Kahneman, D., Wakker, P. & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility, *The Quarterly Journal of Economics* **112**, 375–406.
- [8] Menger, K. (1934). Das unsicherheitsmoment in der wertlehre. Betrachtungen im an-schluß an das sogenannte Petersburgerspiel, *Zeitschrift Für Nationalökonomie* **5**, 459–485.
- [9] Pareto, V. (1909). *Manuale di Economia Politica: con una Introduzione Alla Scienza Sociale*, Società editrice libraria, Milan; English edition (1972). *Manual of Political Economy*, Macmillan, London.
- [10] Savage, L.J. (1954). *The Foundations of Statistics*, Wiley, New York.
- [11] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.
- [12] von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

DANIEL READ AND MARA AIROLDI

Validity Theory and Applications

STEPHEN G. SIRECI

Volume 4, pp. 2103–2107

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Validity Theory and Applications

Measurement in the behavioral sciences typically refers to measuring characteristics of people such as attitudes, knowledge, ability, or psychological functioning. Such characteristics, called *constructs*, are hard to measure because unlike objects measured in the physical sciences (such as weight and distance), they are not directly observable. Therefore, the validity of measures taken in the behavioral sciences is always suspect and must be defended from both theoretical and empirical perspectives.

In this entry, I discuss validity theory and describe current methods for validating assessments used in the behavioral sciences. I begin with a description of a construct, followed by descriptions of construct validity and other validity nomenclature. Subsequently, the practice of test validation is described within the context of Kane's [8] argument-based approach to validity and the current *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education, [1]).

Constructs, Validity, and Construct Validity

The term *construct* has an important meaning in testing and measurement because it emphasizes the fact that we are not measuring tangible attributes (*see Latent Variable*). Assessments attempt to measure unobservable attributes such as attitudes, beliefs, feelings, knowledge, skills, and abilities. Given this endeavor, it must be assumed that (a) such attributes exist within people and (b) they are measurable. Since we do not know for sure if such intangible attributes or proficiencies really exist, we admit they are 'constructs'; they are hypothesized attributes we believe exist within people. Hence, we 'construct' these attributes from educational and psychological theories. To measure these constructs, we typically define them operationally through the use of test specifications and other elements of the assessment process [6].

Cronbach and Meehl [5] formally defined 'construct' as 'some postulated attribute of people, assumed to be reflected in test performance' (p. 283). The current version of the *Standards for Educational and Psychological Testing* [1] defines a construct as 'the concept or characteristic that a test is designed to measure' (p. 173). Given that a construct is invoked whenever behavioral measurement exists [9, 10], the *validity* of behavioral measures are often defined within the framework of *construct validity*. In fact, many validity theorists describe *construct validity* as equivalent to validity in general. The *Standards* borrow from Messick [10] and other validity theorists to underscore the notion that validity refers to inferences about constructs that are made on the basis of test scores.

According to the *Standards*, construct validity is:

A term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test. (AERA et al. [1], p. 174)

The notion that construct validity is validity in general asserts that all other validity modifiers, such as content or criterion-related validity, are merely different ways of 'cutting validity evidence' ([10], p. 16). As Messick put it, 'Construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores' (p. 17).

Before the most recent version of the *Standards*, validity was most often discussed using three categories – content validity, criterion-related validity, and construct validity [11]. Although many theorists today believe terms such as content and criterion-related validity are misnomers, an understanding of these traditional ways of describing the complex nature of validity is important for understanding contemporary validity theory and how to validate inferences derived from test scores. We turn to a brief description of these traditional categories. Following this description, we summarize the key aspects of contemporary validity theory, and then describe current test validation practices.

2 Validity Theory and Applications

Traditional Validity Categories: Content, Criterion-related, and Construct Validity

Prior to the 1980s, most discussions of validity described it as comprising the three aforementioned types or aspects – content, criterion-related, and construct. *Content validity* refers to the degree to which an assessment represents the content domain it is designed to measure. There are four components of content validity – domain definition, domain relevance, domain representation, and appropriate test construction procedures [11]. The domain definition aspect of content validity refers to how well the description of the test content, particularly the test specifications, is regarded as adequately describing what is measured and the degree to which it is consistent with other theories and pragmatic descriptions of the targeted construct. Domain relevance refers to the relevance of the items (tasks) on a test to the domain being measured. Domain representation refers to the degree to which a test adequately represents the domain being measured. Finally, appropriate test construction procedures refer to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material.

The term *content domain* is often used in describing content validity, but how this domain differs from the construct is often unclear, which is one reason why many theorists believe content validity is merely an aspect of construct validity. Another argument against the use of the term content validity is that it refers to properties of a test (e.g., how well do these test items represent a hypothetical domain?) rather than to the inferences derived from test scores. Regardless of the legitimacy of the term content validity, it is important to bear in mind that the constructs measured in the behavioral sciences must be defined clearly and unambiguously and evaluations of how well the test represents the construct must be made. Validating test content is necessary to establish an upper limit on the validity of inferences derived from test scores because if the content of a test cannot be defended as relevant to the construct measured, then the utility of the test scores cannot be trusted.

The second traditional validity category is *criterion-related validity*, which refers to the degree to which test scores correlate with external criteria that are considered to be manifestations of the construct of

interest. There are two subtypes of criterion validity – *predictive validity* and *concurrent validity*. When the external criterion data are gathered long after a test is administered, such as when subsequent college grades are correlated with earlier college admissions test scores, the validity information is of the predictive variety. When the external criterion data are gathered about the same time as the test data, such as when examinees take two different test forms, the criterion-related validity is of the concurrent variety. The notion of criterion-related validity has received much attention in the validity literature, including the importance of looking at correlations between test scores and external criteria *irrelevant* to the construct measured (i.e., discriminant validity) as well as external criteria commensurate with the construct measured (i.e., convergent validity, [2]).

The last category of validity is construct validity, and as mentioned earlier, it is considered to be the most comprehensive form of validity. Traditionally, construct validity referred to the degree to which test scores are indicative of a person's standing on a construct. After introducing the term, Cronbach and Meehl [5] stated 'Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured' (p 282). Given that a finite universe of content rarely exists and valid external criteria are extremely elusive, it is easy to infer that construct validity is involved whenever validity is investigated. Therefore, contemporary formulations of validity tend to describe content and criterion-related validity not as types of validity, but as ways of accumulating construct validity *evidence*. Using this perspective, evidence that the content of a test is congruent with the test specifications and evidence that test scores correlate with relevant criteria are taken as evidence that the construct is being measured. But construct validity is more than content and criterion-related validity. As described in the excerpt from the *Standards* given earlier, studies that evaluate the internal structure of a test (e.g., **factor analysis** studies), **differential item functioning** (item bias), cultural group differences (test bias), and unintended and intended consequences of a testing program all provide information regarding construct validity.

Fundamental Characteristics of Validity

The preceding section gave an historical perspective on validity and stressed the importance of construct

validity. We now summarize fundamental characteristics of contemporary validity theory, borrowing from Sireci [13]. First, validity is *not* an intrinsic property of a test. A test may be valid for one purpose, but not for another, and so what we seek to validate in judging the worth of a test is the inferences derived from the test scores, not the test itself. Therefore, an evaluation of test validity starts with an identification of the specific purposes for which test scores are being used. When considering inferences derived from test scores, the validator must ask ‘For what purposes are tests being used?’ and ‘How are the scores being interpreted?’

Another important characteristic of contemporary validity theory is that evaluating inferences derived from test scores involves several different types of qualitative and quantitative evidence. There is not one study or one statistical test that can validate a particular test for a particular purpose. Test validation is continuous, with older studies paving the way for additional research and newer studies building on the information learned in prior studies.

Finally, it should be noted that although test developers must provide evidence to support the validity of the interpretations that are likely to be made from test scores, ultimately, it is the responsibility of the *users* of a test to evaluate this evidence to ensure that the test is appropriate for the purpose(s) for which it is being used.

Messick succinctly summarized the fundamental characteristics of validity by defining validity as ‘an integrated evaluative judgment of the degree to which evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment’ (p. 13). By describing validity as ‘integrative’, he championed the notion that validity is a unitary concept centered on construct validity. He also paved the way for the *argument-based approach* to validity that was articulated by Kane [8], which is congruent with the *Standards*. We turn now to a discussion of test validation from this perspective.

Test Validation: Validity Theory Applied in Practice

To make the task of validating inferences derived from test scores both scientifically sound and manageable, Kane [8] suggested developing a defensible validity ‘argument’. In this approach, the validator builds an argument on the basis of empirical

evidence to support the use of a test for a particular purpose. Although this validation framework acknowledges that validity can never be established absolutely, it requires evidence that (a) the test measures what it claims to measure, (b) the test scores display adequate reliability, and (c) test scores display relationships with other variables in a manner congruent with its predicted properties. Kane’s practical perspective is congruent with the *Standards*, which provide detailed guidance regarding the types of evidence that should be brought forward to support the use of a test for a particular purpose. For example, the *Standards* state that

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses... Ultimately, the validity of an intended interpretation... relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees... (AERA et al. [1], p. 17)

Two factors guiding test validation are evaluating *construct underrepresentation* and *construct-irrelevant variance*. As Messick [10] put it, ‘Tests are imperfect measures of constructs because they either leave out something that should be included... or else include something that should be left out, or both’ ([10], p. 34). Construct underrepresentation refers to the situation in which a test measures only a portion of the intended construct (or content domain) and leaves important knowledge, skills, and abilities untested. Construct-irrelevant variance refers to the situation in which the test measures proficiencies irrelevant to the intended construct. Examples of construct-irrelevant variance undermining test score interpretations are when computer proficiency affects performance on a computerized mathematics test, or when familiarity with a particular item format (e.g., multiple-choice items) affects performance on a reading test.

To evaluate construct underrepresentation, a test evaluator searches for content validity evidence. A preliminary question to answer is ‘How is the content domain defined?’ In employment, licensure, and certification testing, job analyses often help determine the content domain to be tested. Another important

question to answer is ‘How well does the test represent the content domain?’ In educational testing, subject matter experts (SMEs) are used to evaluate test items and judge their congruence with test specifications and their relevance to the constructs measured (see [4] or [12] for descriptions of methods for evaluating content representativeness). Thus, traditional studies of content validity remain important in contemporary test validation efforts (see [14] for an example of content validation of psychological assessments).

Evaluating construct-irrelevant variance involves ruling out extraneous behaviors measured by a test. An example of construct-irrelevant variance is ‘method bias’, where test scores are contaminated by the mode of assessment. Campbell and Fiske [2] proposed a **multitrait-multimethod** framework for studying construct representation (e.g., convergent validity) and construct-irrelevant variance due to method bias.

Investigation of differential item functioning (DIF) is another popular method for evaluating construct-irrelevant variance. DIF refers to a situation in which test takers who are considered to be of equal proficiency on the construct measured, but who come from different groups, have different probabilities of earning a particular score on a test item. DIF is a statistical observation that involves matching test takers from different groups on the characteristic measured and then looking for performance differences on an item. Test takers of equal proficiency who belong to different groups should respond similarly to a given test item. If they do not, the item is said to function differently across groups and is classified as a DIF item (see [3], or [7] for more complete descriptions of DIF theory and methodology). Item bias is present when an item has been statistically flagged for DIF and the reason for the DIF is traced to a factor irrelevant to the construct the test is intended to measure. Therefore, for item bias to exist, a characteristic of the item that is unfair to one or more groups must be identified. Thus, a determination of item bias requires subjective judgment that a statistical observation (i.e., DIF) is due to some aspect of an item that is irrelevant to the construct measured. That is, difference observed across groups in performance on an item is due to something unfair about the item.

Another important area of evaluation in contemporary validation efforts is the analysis of the fairness of a test with respect to consistent measurement across

identifiable subgroups of examinees. One popular method for looking at such consistency is analysis of *differential predictive validity*. These analyses are relevant to tests that have a predictive purpose such as admissions tests used for college, graduate schools, and professional schools. In differential predictive validity analyses, the predictive relationships across test scores and criterion variables are evaluated for consistency across different groups of examinees. The typical groups investigated are males, females, and ethnic minority groups. Most analyses use **multiple linear regression** to evaluate whether the regression slopes and intercepts are constant across groups [15].

Summary

In sum, contemporary test validation is a complex endeavor involving a variety of studies aimed toward demonstrating that a test is measuring what it claims to measure and that potential sources of invalidity are ruled out. Such studies include dimensionality analyses to ensure the structure of item response data is congruent with the intended test structure, differential item functioning analyses to rule out item bias, content validity studies to ensure the relevance and appropriateness of test content, criterion-related validity studies to evaluate hypothesized relationships among test scores and external variables, and surveys of invested stakeholders such as test takers and test administrators. Relatively recent additions to test validation are studies focusing on social considerations associated with a testing program including unintended consequences such as narrowing the curriculum to improve students’ scores on educational tests. It should also be noted that evidence of adequate test score reliability is a prerequisite for supporting the use of a test for a particular purpose since inconsistency in measurement due to content sampling, task specificity, ambiguous scoring rubrics, the passage of time, and other factors adds construct-irrelevant variance (i.e., error variance) to test scores.

References

- [1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.

-
- [2] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.
 - [3] Clauser, B.E. & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items, *Educational Measurement: Issues and Practice* **17**(1), 31–44.
 - [4] Crocker, L.M., Miller, D. & Franks, E.A. (1989). Quantitative methods for assessing the fit between test and curriculum, *Applied Measurement in Education* **2**, 179–194.
 - [5] Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin* **52**, 281–302.
 - [6] Downing, S.M. & Haladyna, T.M., eds (2005). *Handbook of Testing*, Lawrence Erlbaum, Mahwah.
 - [7] Holland, P.W. & Wainer, H. eds (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale.
 - [8] Kane, M.T. (1992). An argument based approach to validity, *Psychological Bulletin* **112**, 527–535.
 - [9] Loevinger, J. (1957). Objective tests as instruments of psychological theory, *Psychological Reports* **3**, (Monograph Supplement 9), 635–694.
 - [10] Messick, S. (1989). Validity, in *Educational Measurement*, 3rd Edition, R. Linn, ed., American Council on Education, Washington, pp. 13–100.
 - [11] Sireci, S.G. (1998a). The construct of content validity, *Social Indicators Research* **45**, 83–117.
 - [12] Sireci, S.G. (1998b). Gathering and analyzing content validity data, *Educational Assessment* **5**, 299–321.
 - [13] Sireci, S.G. (2003). Validity, *Encyclopedia of Psychological Assessment*, Sage Publications, London, pp. 1067–1069.
 - [14] Vogt, D.S., King, D.W. & King, L.A. (2004). Focus groups in psychological assessment: enhancing content validity by consulting members of the target population, *Psychological Assessment*, **16**, 231–243.
 - [15] Wainer, H. & Sireci, S.G. (2005). Item and test bias, in *Encyclopedia of Social Measurement Volume 2*, Elsevier, San Diego, pp. 365–371.

STEPHEN G. SIRECI

Variable Selection

STANLEY A. MULAİK

Volume 4, pp. 2107–2110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Variable Selection

In selecting variables for study, it is necessary to have a clear idea of what a variable is. For mathematicians, it is a quantity that may assume any one of a set of values, such as the set of integers, the set of real numbers, the set of positive numbers, and so on. But when mathematics is used to model the world, the schema of a mathematical variable must be put in correspondence to something in the world. This is possible because the schema of an object is that an object comes bearing attributes or properties. For example, a particular person will have a certain color for the eyes, another for the hair, a certain weight, a certain height, a certain age: ‘John has brown eyes, blonde hair, is 180 cm tall, weighs 82 kg, and is 45 years old’. Eye color, hair color, height, weight, and age are all variables (in a more general sense). For example, eye color pertains to the set (blue, brown, pink, green, hazel, gray, black). No person’s eye color can assume more than one of the members of this set in any one eye at any one time. Eye color is not a quantity, but it varies from person to person and no person’s eye can be more than one of the members of this set. One might map eye colors to integers, but such a mapping might not represent anything particularly interesting other than a way to distinguish individuals’ eye color with numerical names. On the other hand, weight is a quantity and pertains to the set of weights in some unit of measurement represented by positive real numbers. How a number gets assigned to the person to be his or her weight involves measurement, and this is a complex topic of its own (see **Measurement: Overview** and [2]). In behavioral and social sciences, variables may be identified with the responses an individual makes to items of a questionnaire. These may be scaled to represent measurements of some attribute that persons have. We will assume that the variables to be studied are measurements.

A frequent failing of researchers who have not mastered working with quantitative concepts is that in their theorizing they often fail to think of their theoretical constructs as variables. But then they seek to study them statistically using methods that require that they work with variables. A common mistake that may mislead one’s thinking is not to think of them as quantities or by names of quantities.

For example, a researcher may hypothesize a causal relationship between ‘leader–follower exchange’ and ‘productivity’, or between ‘academic self-concept’ and ‘achievement’. These do not refer to quantities, except indirectly. ‘Leader–follower exchange’ does not indicate which aspect of an exchange between leader and follower is being measured and how it is a quantity. Is it ‘the *degree* to which the leader distrusts the follower to carry out orders’? Is it ‘the *frequency* with which the follower seeks advice from the leader’? Is it ‘the *extent* to which the leader allows the follower to make decisions on his own’? And what is ‘productivity’? What is produced? Can it be quantified? As for ‘academic self-concept’, there are numerous variables by which one may describe oneself. So, which is it? Is it ‘the *degree* to which the individual feels confident with students of the opposite sex’? Is it ‘the *frequency* with which the individual reports partying with friends during a school year’? Is it ‘the number of hours the individual reports he/she studies each week’? And is ‘achievement’ measured by ‘GPA’ or ‘final exam score’, or ‘postgraduate income’? When one names variables, one should name them in a way that accurately describes what is measured. Forcing one to use words like ‘degree’, ‘extent’, ‘frequency’, ‘score’, ‘number of’ in defining variables will help in thinking concretely about what quantities have influence on other quantities, and improve the design of studies.

Selecting variables for regression. Suppose Y is a criterion (dependent) variable to be predicted and X_i is a predictor (independent variable) in a set of p predictors. It is known that the regression coefficient between variable X_i and variable Y is

$$\beta_{Yi} = \rho_{Yi|12\cdots(i)\cdots p} \frac{\sigma_{Y|12\cdots(i)\cdots p}}{\sigma_{i|12\cdots(i)\cdots p}}, \quad (1)$$

where $\rho_{Yi|12\cdots(i)\cdots p}$ is the partial correlation between the criterion Y and predictor X_i , with the predicted effects of all other predictors subtracted out of them. (Note ‘ $12\cdots(i)\cdots p$ ’ means ‘the variables X_1, X_2 through to X_p with variable X_i not included’.) $\sigma_{Y|12\cdots(i)\cdots p}$ is the conditional standard deviation of the dependent variable Y , with the predicted effects of all predictor variables except variable X_i subtracted out. $\sigma_{i|12\cdots(i)\cdots p}$ is the conditional standard deviation of X_i , with the predicted effects of all other predictors subtracted from it. So, the regression weight represents the degree to which the part of X_i that

2 Variable Selection

is relatively unique with respect to the other predictors is still related to a part of the criterion not predicted by the other variables. This means that an ideal predictor is one that has little in common with the other predictors but much in common with the criterion Y . In an ideal extreme case, the predictors have zero correlations among themselves, so that they have nothing in common among them, but each of the predictors has a strong correlation with the criterion.

There are several posthoc procedures for selecting variables as predictors in regression analysis. The method of *simultaneous regression* estimates the regression coefficients of all of the potential predictor variables given for consideration simultaneously. Thus, the regression weight of any predictor is relative to the other predictors included, because it concerns the relationship of the predictor to the criterion variable, with the predicted effects of other predictors subtracted out. The success of the set of predictors is given by the multiple correlation coefficient R (see **Multiple Linear Regression**). This quantity varies between 0 and 1, with one indicating perfect prediction by the predictors. The squared multiple correlation R^2 gives the proportion of the variance of the criterion variable accounted for by the full set of predictors. One way of assessing the importance of a variable to prediction is to compute the relative gain in the proportion of variance accounted for by adding the variable to the other variables in the prediction set. Let W be the full set of p predictors including X_j . Let V be the set of $p - 1$ predictors $W - \{X_j\}$. Then $r_{yj \cdot V}^2 = (R_{y \cdot W}^2 - R_{y \cdot V}^2) / (1 - R_{y \cdot V}^2)$ is the relative gain in proportion of variance accounted for due to including variable X_j with the predictors in set V . Here, $r_{yj \cdot V}^2$ is also the squared partial correlation of variable X_j with criterion variable Y , holding constant the variables in V . $R_{y \cdot W}^2$ is the squared multiple correlation for predicting Y from the full set of predictors W . $R_{y \cdot V}^2$ is the squared multiple correlation for predicting Y from the reduced set V . $r_{yj \cdot V}^2$ can be computed for each variable and relative importance compared among the predictors. It is possible also to test whether the absolute incremental gain in proportion of variance accounted for due to variable X_j is significantly different from zero. This is given by the formula $F = (R_{y \cdot W}^2 - R_{y \cdot V}^2) / [(1 - R_{y \cdot W}^2)(N - p - 1)]$. F is distributed as chi squared with 1 and $(N - p - 1)$ degrees of freedom. (Source: [1], pp. 719–720).

Another method is the *hierarchical* method. The predictor variables are arranged in a specific rational order based on the research question. We begin by entering the first variable in the order and determine the degree to which it explains variance in the dependent variable. We then examine how much the next variable in the order adds to the proportion of variance accounted for beyond the first, then how much the next variable afterwards adds to the proportion of variance accounted for beyond the first two, and so on, to include finally the p th predictor beyond the first $p - 1$ predictors. In the end, the R^2 for the full set will be the same as in the simultaneous method, and the regression weights will be the same. But this method can suggest at what point what variables might be dropped from consideration (Source: [1], pp. 731–732).

In addition to this procedure based on rational ordering, entry can be based on empirical ordering. The method of *stepwise regression* begins with no preset order to the variables. Usually, one begins with the variable that has the largest squared correlation with the criterion. Then one pairs the first variable with each of the other variables and computes a squared multiple correlation for each pair with the criterion. One adds the variable from the remaining variables that produces the largest squared multiple correlation among the pairs. One also computes the relative gain in proportion of variance accounted for. Then one seeks a third variable from the remaining predictors, which when added to the first two, produces the largest squared multiple correlation with the criterion. Again one also computes the relative gain in proportion of variance accounted for. And one keeps on until either one has selected all of the variables or comes to a point where no additional variable adds a meaningful or significant increment in proportion of variance accounted for. If one accompanies this with significance tests for the gain, it can involve many tests which are not statistically independent (Source: [1], pp. 732–735).

One can also proceed in a *backward* or *step-down* direction, beginning with all the variables and eliminating, one at a time, variables that successively account for the least decrement in the squared multiple correlation at each step until one gets to a step where any additional elimination of variables would seriously decrease the squared multiple correlation at that point.

An important disadvantage of empirical entry/selection is that it capitalizes on chance. This is particularly important when predictors are highly related to one another. In a given data set, any of the empirical entry procedures might lead to a certain order of entry, whereas in another data set, with values drawn from the same populations of values, the order might be quite different.

References

- [1] Hays, W.L. (1994). *Statistics*. Fort Worth, TX: Harcourt Brace.
- [2] Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*, Lawrence Erlbaum, Hillsdale.

STANLEY A. MULAİK

Variance

DAVID CLARK-CARTER

Volume 4, pp. 2110–2110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Variance

The variance is defined as the mean of the squared deviations of a set of numbers about their mean. If a population consists of the set of values $X_1, X_2, \dots, X_n (i = 1, 2, \dots, n)$, then the population variance, usually denoted by σ^2 , is

$$\sigma^2 = \frac{\sum_i (X_i - M)^2}{n}, \quad (1)$$

where M is the population **mean**.

This formula is also appropriate for finding the variance of a sample in the unlikely event that we know the population mean. However, if we are dealing with a sample and we are also obliged to calculate the sample mean, $\bar{X}$, then dividing the sum of squared deviations by a value fewer than the number of scores in the set ($n - 1$) produces the best estimate of the variance in the population from which the sample was obtained. This form, often denoted by S^2 , is called the *sample variance* and is defined as

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n - 1}. \quad (2)$$

It is this version of the variance that is usually reported by statistical software. Notice, though, that as the sample size increases, there will be less and less difference between the values of the variance obtained using n or $(n - 1)$. Even for small samples, if the variance is required only for descriptive purposes, it is usually immaterial as to which divisor is used.

Table 1 shows the number of words recalled by a hypothetical sample of six participants in a study of

Table 1 Variance calculations for word recall data

Participant	Words recalled	Deviation from mean	Squared deviation
1	5	−2.5	6.25
2	7	−0.5	0.25
3	6	−1.5	2.25
4	9	1.5	2.25
5	11	3.5	12.25
6	7	−0.5	0.25

short-term memory. The mean recall for the sample is 7.5 words. The sum of the squared deviations is 23.5.

The version of the variance treating the data as a population (that is dividing this sum of the squared deviations by 6) is 3.92, while the version of the variance that *estimates* the population variance (dividing by 5) is 4.7.

Calculating variances using the above method can be tedious, so it is worth noting that there are easier computational forms (see, for example, [1]).

As with other summary statistics that rely equally on all the numbers in the set, the variance can be severely affected by extreme scores. Nonetheless, the variance underlies inferential procedures such as the **analysis of variance**, while other methods, such as standardizing a set of data, draw on its close relative, the **standard deviation**.

Reference

- [1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.

DAVID CLARK-CARTER

Variance Components

NICHOLAS T. LONGFORD

Volume 4, pp. 2110–2113

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Variance Components

Variability is a ubiquitous feature of characteristics and attributes studied in human populations. Its prominence is a result of our interest in the details of behavioral, attitudinal, economic, and medical outcomes, aspects in which human subjects differ widely and with limited predictability. Without variability, there is not much to study about a population, because one or a few of its members then inform about the population completely. On the other hand, explanation of variability, in the form of a mechanism describing its causes, is the ultimate and often unachievable research goal. Such goals are pursued by various regression models and their extensions, and the commonly used terminology in connection with imperfect model fit, such as ‘unexplained variation’, implies a failure of the research effort. Such a pessimistic view is often poorly supported. In many situations, we have to resign ourselves to a description of variability less complete than by regression. *Variance components* are a key device for such a description, and they are associated with factors or *contexts* involved in the analyzed study. The contexts may be introduced deliberately, or be an unavoidable nuisance feature. For example, when essays (examination papers) are graded by raters, the raters are such a context. Their training and instruction aims to reduce the differences among the raters, but eradicating them completely is not possible [4].

In this example, the elementary observations (essays) are clustered within raters; the essays graded by a rater form clusters or level-2 units (*see Clustered Data*). Even if the essays are assigned to raters uninformatively (at random, or with no regard for any factor associated with their quality), the scores assigned to them have an element of similarity brought on by being graded by the same rater. Inferences are desired for each essay and an ‘average’ or typical rater; we wish to *generalize* our findings from the incidental to the universal. Hence the term *generalizability theory* introduced by [1] (*see Generalizability*).

In practice, we come across several contexts ‘interfering’ with our observations simultaneously. For example, intellectual or academic abilities can be studied only indirectly, by observing individuals’ performances on tasks that represent the domain of

abilities, and assessing the performances by raters using scales constructed with simplicity in mind. Here, the assessment (testing) instrument, or the individual tasks, the setting (occasion) of the text or examination, and the raters are factors (or contexts) associated with variation. That is, if a different testing instrument were used (and the settings of all other contexts held fixed), the assessments of the (same) subjects would be different. If a different rater graded a particular response, the score assigned may be different. Ability, with the appropriate qualification, is regarded as a characteristic of an individual, whereas the individual’s performances vary around the ability, depending on the momentary disposition, the balance of questions, tasks, or the like, in the assessment instrument and the rater’s judgment, which may be inconsistent over replications and may differ among the raters. Thus, apart from the variation in the ability, that is of key inferential interest, several other sources contribute to the variation of the outcomes (scores) – in a hypothetical replication of the study, different scores would be recorded. These *nuisance* contexts are unavoidable, or have to be introduced, because assessment (measurement) cannot be conducted in a contextual vacuum. The levels (or settings) of a context can be regarded as a *population*, thus introducing the sampling-design issue of good representation of a context in the study. This clarifies the scope of the inference – for what range of settings the conclusions are intended and are appropriate.

For the case of a single context, let y_{ij} be the outcome for elementary-level unit (say, subject) i in setting j (say, group, cluster, or unit at level 2). The simplest nontrivial model that describes y_{ij} is

$$y_{ij} = \mu + \delta_j + \varepsilon_{ij}, \quad (1)$$

where μ is the overall mean that corresponds to (population-related) averaging over settings and elements, δ_j is the deviation specific to setting j , $j = 1, \dots, J$, and ε_{ij} represents the deviation of subject i , $i = 1, \dots, n_j$, from the setting-specific mean $\mu + \delta_j$. The deviations δ_j and ε_{ij} are mutually independent random samples from centered distributions with respective variances σ_1^2 and σ_2^2 . Usually, these distributions are assumed to be normal, often as a matter of expedience because the normal distribution is closed with respect to addition: that is, if a and b are normally distributed random variables, then so is their total $a + b$, with mean $E(a) + E(b)$ and variance $\text{var}(a) + \text{var}(b) + 2\text{cov}(a, b)$.

2 Variance Components

As $\text{var}(y_{ij}) = \sigma_1^2 + \sigma_2^2$, it is meaningful to call σ_1^2 and σ_2^2 the variance components associated with subjects and settings, respectively. The deviations δ_j cause the observations y_{ij} to be correlated – observations within settings are more similar than observations in general. We have

$$\text{cor}(y_{ij}, y_{i'j}) = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (2)$$

for any pair of observations $i \neq i'$ in the same setting. If a single setting j were studied, this correlation could not be recognized (identified); δ_j and μ would be confounded:

$$y_{ij} = \mu_j + \varepsilon_{ij} \quad (3)$$

and, as μ_j is a constant (unchanged in replications), the outcomes are mutually independent, each with variance σ_1^2 . That is, $\text{var}(y_{ij}|j) = \sigma_1^2$. (Instead of conditioning on setting j we can condition on the expectation μ_j .) If the setting-specific expectations μ_j were the observations, σ_2^2 would be identified as their variance. It is essential to distinguish between μ_j and μ , even when the study involves a single setting j of the context. Otherwise, an unjustified generalization is made that all the settings of the context are identical. Also, the parameter μ should be qualified by the population (class) of contexts; different classes of a context are associated with different values of μ .

When several contexts are present each of them is represented by a variance (component). The contexts may be nested, as with individuals within households, streets, towns, and countries, or crossed, as with examination papers rated separately by two raters, each drawn from the same pool (*see Cross-classified and Multiple Membership Models*). In general, nested contexts are much easier to handle analytically, although estimation and other forms of inference simplify substantially when the numbers of observations within the combinations of levels of the contexts are equal (*balanced design*).

When the context-specific deviations are additive, as in

$$y_{hij} = \mu + \delta_h + \gamma_i + \varepsilon_{hij} \quad (4)$$

(contexts h and i are crossed), either a suitable notation has to be introduced, or the model supplemented with a description of how the contexts appear in the data and in the relevant populations. For example, each level h of one context can occur with each level

i of the other context. The covariance of two observations is equal to the total of the variance components for the contexts shared by the two observations. For example, $\text{cov}(y_{hij}, y_{hi'j'}) = \sigma_\delta^2$.

Variance component models can be combined with regression. The model in (1) has the extension

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j + \varepsilon_{ij}, \quad (5)$$

where $\mathbf{x}_{ij}$ are the covariates associated with the element ij . Each covariate may be defined for the elements ij or for the setting j (of the context); a more detailed model formulation is

$$y_{ij} = \mathbf{x}_{1,ij}\boldsymbol{\beta}_1 + \mathbf{x}_{2,j}\boldsymbol{\beta}_2 + \delta_j + \varepsilon_{ij}, \quad (6)$$

in which the covariates $\mathbf{x}_{1,ij}$ are defined for the elements and $\mathbf{x}_{2,j}$ for the settings. A variable defined for the elements may be constant within settings; $x_{ij} \equiv x_j^*$; this may be the case only in the sample (it would not be the case in each replication), or in the relevant population. Conversely, each context-level variable x_j can be regarded as an elementary-level variable by defining $x_{ij}^\dagger \equiv x_j$.

In most research problems, the role of the covariates is to reduce the variance components $\text{var}(\varepsilon_{ij}) = \sigma_1^2$ and $\text{var}(\delta_j) = \sigma_2^2$. Adding a context-level covariate to $\mathbf{x}$ can reduce only the context-level variance σ_2^2 , whereas the addition of an elementary-level variable to $\mathbf{x}$ can reduce both σ_1^2 and σ_2^2 . A variable with the same distribution at each setting of the context reduces only the within-context variation. Variables with both within- and between-setting components of variation can reduce both variance components. Counterintuitive examples in which variance components are increased when the model is expanded by one or several variables are given in [6].

The model in (5) involves several restrictive assumptions. First, linearity is often adopted as a matter of analytical convenience. It can be dispensed with by replacing the predictor $\mathbf{x}\boldsymbol{\beta}$ with a general function $f(\mathbf{x}; \boldsymbol{\theta})$ which would, nevertheless, involve a functional form of f . Next, the variance components are constant. Heteroscedasticity of the context can be introduced by assuming a functional form for $\text{var}(\delta_j)$, dependent on some variables (*see Heteroscedasticity and Complex Variation*). No generality is lost by assuming that these variables are a subset of $\mathbf{x}$. The simplest nontrivial example assumes that the levels of the context belong to a small number of categories (subpopulations), each with its own

variance. In another example, $\text{var}(\delta_j) = \exp(m_j)$, where m_j is a function of the population size of context j .

The deviation of the context from the average regression $\mathbf{x}\boldsymbol{\beta}$ need not be constant. A natural way of introducing this feature is by random coefficients, or by a more general pattern of between-context variation:

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\delta_j + \varepsilon_{ij}, \quad (7)$$

where $\mathbf{z}$ is a subset of the variables in $\mathbf{x}$. Now $\text{var}(\mathbf{z}_{ij}\delta_j) = \mathbf{z}_{ij}\boldsymbol{\Sigma}\mathbf{z}_{ij}^\top$, where $\boldsymbol{\Sigma} = \text{var}(\delta_j)$. If all variables in $\mathbf{z}$ are defined for elements the model in (7) has an interpretation in terms of varying within-context regressions. But the *variation part* of the model, $\mathbf{z}_{ij}\delta_j$, may, in principle, involve any variables. Including a context-level variable in $\mathbf{z}$ is meaningful only when the context has many levels (categories); otherwise estimation of the elements of the variance matrix $\boldsymbol{\Sigma}$ is an ill-conditioned problem.

Finally, the deviations of the contexts and elements within contexts, after an appropriate regression adjustment, need not be additive. The random deviations are multiplicative in the model

$$y_{ij} = f(\mathbf{x}; \boldsymbol{\beta}) \delta_j \varepsilon_{ij}, \quad (8)$$

where δ_j and ε_{ij} are random samples from nonnegative distributions with unit means. This model reduces to additive deviations (and variance components) by the log-transformation. Although nontrivial nonadditive models are easy to specify, they are applied in practice infrequently, because additive random terms are much more tractable and correspond to addition of their variances. Moreover, when the deviations are normally distributed, so are their linear combinations, including totals.

Generalized linear models provide a vehicle for substantial extension of variance component models within the realm of linearity and normality of the deviations. Thus a model

$$g\{E(y_{ij})\} = \mathbf{x}_{ij}\boldsymbol{\beta}, \quad (9)$$

with a suitable link function g and a distributional assumption for y (say, logit function g and binary y) is extended to

$$g\{E(y_{ij} | \delta_j)\} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j, \quad (10)$$

with the obvious further extensions to varying within-setting regressions (see **Generalized Linear Mixed Models**).

A related way of defining more general variance component models is by assuming that an ordinal or dichotomous observed variable y is the *manifest* version of a normally distributed *latent* variable y^* that satisfies an additive variance component model. A generic approach to fitting such models is by the EM algorithm (see **History of Intelligence Measurement**) [2], regarding the latent outcomes as the missing information. See [5].

A suitable process for the conversion of latent values to their manifest versions has to be defined. As many key variables in behavioral research are ordinal categorical, a coarsening process [3] can be posited in many settings. It specifies that there is a small number of cut-points $-\infty = c_0 < c_1 < \dots < c_{K-1} < c_K = +\infty$ and all latent values that fall into the interval (c_{k-1}, c_k) convert to manifest value k . Note that the decomposition of the variance to its context-related components is meaningful only for the latent outcomes. The pattern of dependence or the covariance structure among the values of the observed outcome y is usually not related in any straightforward way to the covariance structure on the latent scale. The connection can usually be explored only by simulations.

References

- [1] Cronbach, L.J. Gleser, G.C. Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements. The Theory of Generalizability of Scores and Profiles*, Wiley and Sons, New York.
- [2] Dempster, A.P. Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B* **39**, 1–38.
- [3] Heitjan, D.F. & Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association* **85**, 304–314.
- [4] Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*, Springer-Verlag, New York.
- [5] McKelvey, R. & Zavoina, W. (1975). A statistical model for the analysis of ordinal dependent variables, *Journal of Mathematical Sociology* **4**, 103–120.
- [6] Snijders, T.A.B. & Bosker, R. (1999). *Multilevel Analysis*, Sage Publications, London.

(See also **Linear Multilevel Models**)

NICHOLAS T. LONGFORD

Walsh Averages

CLIFFORD E. LUNNEBORG

Volume 4, pp. 2115–2116

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Walsh Averages

Let $(x_1, x_2, \dots, x_n)$ be a random sample from a symmetric distribution with unknown median θ . The set of Walsh averages [4] is the collection of $n(n + 1)/2$ pairwise averages, each of the form $(x_i + x_j)/2$ and computed for all $i = 1, 2, \dots, n$ and for all $j = 1, 2, \dots, n$.

The sample (2, 5, 7, 11), for example, gives rise to the 10 Walsh averages shown in Table 1.

The median of the set of Walsh averages is the one-sample **Hodges-Lehmann estimator**, an efficient point estimator of the median of the sampled distribution. For our example, with 10 Walsh averages, the median estimate is the average of the fifth and sixth smallest Walsh averages, $(6 + 6.5)/2 = 6.25$.

Walsh averages also can be used to find a **confidence interval** (CI) for the median. This usage follows the logic of the Wilcoxon **signed-rank test**; that is, the resulting $(1 - \alpha)100\%$ CI includes those values of θ that would not be rejected at the α level by the Wilcoxon signed-rank test.

One approach to the median CI is described in [1]. We illustrate here the mechanics, using our $n = 4$ example. Let $L_{\alpha/2}$ be a signed-rank sum such that, for a sample of size n , the probability is $\alpha/2$ of a signed-rank sum that size or smaller under the null

hypothesis. Such values are tabulated and widely available. Large-sample approximations have been developed as well, for example, [1] and [2].

The lower and upper limits to the $(1 - \alpha)100\%$ CI are given by the $L_{\alpha/2}$ th smallest and the $L_{\alpha/2}$ th largest Walsh average, respectively.

For $n = 4$, the tabled [3] probability is 0.125 that the signed-rank sum will be 1 or smaller. Thus, a 75% CI for the median is bounded by the smallest and largest Walsh averages. For our example, this yields a 75% CI bounded below by 2 and above by 11. We have 75% confidence that the interval from 2 to 11 contains the population median, θ . A slightly different algorithm for the CI is given in [2], where it is referred to as a Tukey CI.

References

- [1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.
- [2] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition Wiley, New York.
- [3] Larsen, R.J. & Marx, M.L. (2001). *An Introduction to Mathematical Statistics and its Applications*, 3rd Edition, Prentice-Hall, Upper Saddle River.
- [4] Walsh, J.E. (1949). Some significance tests for the median which are valid under very general conditions, *SAnnals of Mathematical Statistics* **20**, 64–81.

CLIFFORD E. LUNNEBORG

Table 1 Computation of Walsh averages

	2	5	7	11
2	$(2 + 2)/2 = 2$	$(2 + 5)/2 = 3.5$	$(2 + 7)/2 = 4.5$	$(2 + 11)/2 = 6.5$
5		$(5 + 5)/2 = 5$	$(5 + 7)/2 = 6$	$(5 + 11)/2 = 8$
7			$(7 + 7)/2 = 7$	$(7 + 11)/2 = 9$
11				$(11 + 11)/2 = 11$

Wiener, Norbert

BRIAN S. EVERITT

Volume 4, pp. 2116–2117

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Wiener, Norbert

Born: November 26, 1894, in Columbia, USA.

Died: March 18, 1964, in Stockholm, Sweden.

The son of Russian émigré, Leo Wiener, Norbert Wiener was a child prodigy who entered Tufts College at the age of 11, graduating three years later. He then entered Harvard to begin graduate studies at the age of 14. Beginning with zoology, Wiener soon changed to mathematical philosophy, receiving his Ph.D. from Harvard at the age of 18, with a dissertation on mathematical logic. From Harvard, the 18-year-old Wiener travelled first to Cambridge, England, to study under Russell and Hardy, and then on to Gottingen to work on differential equations under Hilbert. At the end of World War I, Wiener took up a mathematics post at the Massachusetts Institute of Technology (MIT), where he was eventually to become professor in 1932, a post he held until 1960.

Wiener's mathematical work included functions of a real variable, mathematical logic, relativity, quantum theory, Brownian motion, and the Fourier integral and many of its applications. During World War II, he worked on guided missiles. It was during this

period that Wiener studied the handling of information by electronic devices, based on the feedback principle, phenomena that were later compared with human mental processes in his most famous book, *Cybernetics*, first published in 1947 [1]. Cybernetics was generally defined as 'the science of control and communication in the animal and the machine', with 'animal' very definitely including human beings. Essentially, Wiener was making an analogy between man as a self-regulating system, receiving sensory data and pursuing certain objectives, and mechanical or electrical servomechanisms.

Wiener spent his last years working on the applications of cybernetics to human thinking and pointing out its social dangers and the need to reconstruct society to allow for social evolution. Wiener's cybernetics was the forerunner of Artificial Intelligence. Norbert Wiener was one of twentieth century's true polymaths, and the breadth of his work remains apparent throughout mathematics and probability.

Reference

- [1] Wiener, N. (1948). *Cybernetics*, Wiley, New York.

BRIAN S. EVERITT

Wilcoxon, Frank

DAVID C. HOWELL

Volume 4, pp. 2117–2118

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Wilcoxon, Frank

Born: September 2, 1892, in County Cork, Ireland.

Died: November 18, 1965, in Tallahassee, Florida.

Frank Wilcoxon was born to a wealthy family in 1892 and grew up in the Hudson River Valley. He received his early education at home, and a BS degree from Pennsylvania Military College in 1917. He received his MS degree in chemistry from Rutgers in 1921 and a PhD degree in inorganic chemistry from Cornell in 1924.

Most of Wilcoxon's professional life was spent in industry, first with the Boyce Thompson Institute for Plant Research, and then with American Cyanamid. The major focus of his chemical research dealt with insecticides and fungicides, though later he headed up the statistics group at the Lederle Division of American Cyanamid.

Wilcoxon's interest in statistics began with a study group, of which W. J. Youden was also a part. The group studied **Fisher's** (1925) *Statistical Methods for Research Workers*. It was 20 years before Wilcoxon published anything in statistics, but he was always interested in the applications of statistical methods to chemical research.

Wilcoxon's most important paper [10] appeared in the first volume of what is now *Biometrics* in 1945, and concerned the application of ranking methods for testing differences in location. Both his matched-pairs **signed-ranks** test and his rank-sum test (see **Wilcoxon–Mann–Whitney Test**) were presented in that paper. This paper was important because it led to a growing interest in rank methods (see **Rank Based Inference**), and the development of similar methods for other designs.

Wilcoxon's approach relies on two statistical methods, the use of ranks and the use of permutation procedures. Ranked data had been used for a very long time, at least back to **Galton**. However, according to Kruskal and Wallis [7], the earliest treatment of them as a nonparametric statistical tool would appear to be a paper on rank correlation by Hotelling and Pabst [5] in 1936. Wilcoxon cited a paper by Friedman [4] on the use of ranks to avoid assumptions of normality but, interestingly, his own paper says very little about that issue, even though it is one of the major strengths of his approach. Permutation as

a test procedure was considerably more recent and was first used by Fisher [3] and by Pitman [9] in the 1930s. The permutation tests (see **Permutation Based Inference**) produced by Fisher and Pitman (see **Pitman Test**) were unwieldy, requiring lengthy calculations on all possible (or all extreme) permutations. Wilcoxon, however, hit upon the idea of replacing observations with ranks and permuting the ranks. The first thing this did was to simplify the calculations, which his paper seems to emphasize as the goal. Since, for a given sample size, ranks are constant from one experiment to another, it was possible for Wilcoxon to establish tables of extreme results, thereby standardizing the process. More importantly, using rank substitution allowed statistics to move away from normality assumptions that had underlain nonpermutation tests to that time. Initially, Wilcoxon said very little about what assumptions remained.

Interestingly, Leon Festinger [2] independently developed the same test, retaining the possibility of unequal sample sizes, and published in *Psychometrika* the next year. Mann and Whitney [8] published a very similar idea the next year, and the two-sample test is now frequently referred to as the *Wilcoxon–Mann–Whitney test*. Over the next several years, Wilcoxon, later in conjunction with Roberta Wilcox, published extensive table for his tests [11].

Wilcoxon's tests went on to form the core of a whole set of rank-permutation tests and remain some of the most powerful nonparametric tests. Kruskal [6] provides historical notes on the development of the two-sample test prior to Wilcoxon's time. Bradley [1] provides a summary of his life.

Wilcoxon retired in 1957 but joined the faculty of Florida State University in 1960, where he remained until his death. His later work dealt with sequential ranking methods.

References

- [1] Bradley, R.A. (1966). Obituary: Frank Wilcoxon, *Biometrics* **22**, 192–194.
- [2] Festinger, L. (1946). The significance of differences between means without reference to the frequency distribution function, *Psychometrika* **11**, 97–105.
- [3] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.
- [4] Friedman, M. (1937). The use of ranks to avoid the assumption of normality, *Journal of the American Statistical Association* **32**, 675–701.

2 Wilcoxon, Frank

- [5] Hotelling, H. & Pabst, M.R. (1936). Rank correlation and tests of significance involving no assumption of normality, *Annals of Mathematical Statistics* **7**, 29–43.
- [6] Kruskal, W.H. (1958). Historical notes on the Wilcoxon unpaired two-sample test, *Journal of the American Statistical Association* **52**, 356–360.
- [7] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.
- [8] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.
- [9] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations, *Supplement to the Journal of the Royal Statistical Society* **4**, 119–130.
- [10] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin (now Biometrics)* **1**, 80–83.
- [11] Wilcoxon, F. (1949). *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, New York.

DAVID C. HOWELL

Wilcoxon–Mann–Whitney Test

VENITA DEPUY, VANCE W. BERGER AND YANYAN ZHOU

Volume 4, pp. 2118–2121

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Wilcoxon–Mann–Whitney Test

The Wilcoxon-Mann-Whitney test, also known as the Mann-Whitney U -test or the Wilcoxon rank sum test, is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to location (the median). The alternative may be either directional or nondirectional. For example, in one study [9] the investigators wished to determine whether men and women supported traditional gender roles in differing amounts. After evaluating each person's level of support, the scores were grouped together and ranked from smallest to largest. The sum of the ranks for each gender were then compared. If the two rank sums differed significantly, then the conclusion would be that there is a significant gender difference.

One common approach to testing this hypothesis is by means of a parametric analysis, which assumes normality; typically a two-sample Student's t Test (see **Catalogue of Parametric Tests**). A problem with this approach is that the normality assumption is rarely warranted, and when it is not, the normal-based P value will differ from the exact P value by an amount that cannot be determined without actually computing each one, and then comparing them [1, 2]. But if one were to go through the effort to compute the exact P value, then why would one use it to validate the approximation rather than as the P value to be reported? How can an approximation be preferred to the quantity it is trying to approximate, if that gold standard quantity is already in hand? The alternative nonparametric approach does not assume normality and hence offers more robust results.

There are three primary assumptions of the nonparametric Wilcoxon-Mann-Whitney test:

1. Each sample is randomly selected from the specific population and the observations within each sample are independent and identically distributed.
2. The two samples are independent of each other (otherwise, consider the Wilcoxon **signed rank test**).
3. The populations may differ in their location (mean or median), but not in their distributional

shape or spread (if this assumption is questionable, then consider the **Smirnov test** [11]).

Let $x_1, \dots, x_m$ be a sample from population X and $y_1, \dots, y_n$ be a sample from population Y . Let $F_X(t)$ and $G_Y(t)$ be the cumulative distribution functions for the two populations. The location shift model is that $G_Y(t) = F_X(t + \Delta)$, for all values of t . The null hypothesis, that the X and Y variables have the same probability distribution, can then be stated as: $H_0: \Delta = 0$. This makes sense only if the distributions are continuous, but the Wilcoxon-Mann-Whitney test can be applied even to ordered categorical data.

Let $N = m + n$. To compute the Wilcoxon test statistic, W , first order the combined sample of N values from least to greatest and assign ranks, 1 through N , to the observations (average ranks can be used in case of ties). Next, let $S_1, S_2, \dots, S_n$ denote the ranks assigned to $y_1, \dots, y_n$, the sample from population Y . Then W is the sum of the ranks assigned to the Y sample [14],

$$W = \sum_{j=1}^n S_j. \quad (1)$$

Interpretation of the Wilcoxon test statistic depends on the substantive hypothesis about the two population medians; hence the alternative hypothesis for the location shift parameter Δ may be either one-sided or two-sided:

- a. $[Mdn(Y) > Mdn(X)]$ and $H_a: \Delta > 0$. H_0 is rejected at the α level of significance if $W \geq W_\alpha$, where W_α is chosen to make the type I error probability equal to (or no greater than) α .
- b. $[Mdn(Y) < Mdn(X)]$ and $H_a: \Delta < 0$. H_0 is rejected at the α level of significance if $W \leq n(m + n + 1) - W_\alpha$, where W_α is chosen to make the type I error probability equal to (or no greater than) α .
- c. $[Mdn(Y) \neq Mdn(X)]$ and $H_a: \Delta \neq 0$. H_0 is rejected at the α level of significance if $W \geq W_{\alpha/2}$ or $W \leq n(m + n + 1) - W_{\alpha/2}$, where $W_{\alpha/2}$ is chosen to make the overall type I error probability equal to (or no greater than) α .

The Wilcoxon test is a permutation test (see **Permutation Based Inference**); that is, under the null hypothesis, the set of ranks may be permuted, any m of them being assigned to the X population, with the remaining n being assigned to the Y population

regardless of true group membership. Thus, the null distribution for W consists of the set of Wilcoxon statistics for all $[N!/(m! \times n!)]$ permutations of the ranks. W_α is the $(1 - \alpha)100\%$ quantile of this reference distribution. Because ranks, rather than raw data, are used in the computation of W , it is not necessary to carry out the permutation test for each new data set. For sample sizes, n or m , up to 10, the critical value of W_α may be obtained from tables such as in [7]. It should be noted that the tabled values are exact only if there are no ties among the $(m + n)$ observations.

For larger sample sizes, an approximation based on the asymptotic normality of W is often used [7]. For this purpose, W is standardized:

$$W^* = \frac{W - \{n(m + n + 1)/2\}}{\sqrt{mn(m + n + 1)/12}}. \quad (2)$$

The null hypothesis is rejected in favor of $H_a: \Delta > 0$ if $W^* \geq -z_\alpha$ where z_α is the $(\alpha \times 100)\%$ quantile of the standard normal distribution, for example, -1.96 is the 5% quantile. Where the alternative hypothesis is that $\Delta > 0$, reject H_0 if $W^* \leq z_\alpha$ and where the alternative hypothesis is nondirectional, reject the null hypothesis if $|W^*| \geq -z_{\alpha/2}$.

Both the exact permutation test and approximate calculations are available from a variety of commercial statistical packages (see **Software for Statistical Analyses**).

Correction for Ties

The use of this normal approximation may be unreliable when the sampled distributions are sparse, skewed, or heavily tied. In fact, when ties are present in the sample, the test statistic given above may be too conservative. The following revision decreases the denominator slightly, rendering the outcome less conservative and yielding smaller P values.

$$W^* = \frac{W - \{n(m + n + 1)/2\}}{\sqrt{\frac{mn}{12} \left[m + n + 1 - \frac{\sum_{j=1}^g (t_j - 1)t_j(t_j + 1)}{(m + n)(m + n - 1)} \right]}}, \quad (3)$$

where g is the number of tied groups and t_j is the size of the j th tied group.

Continuity Correction

When the large sample approximation is used, a continuity correction may be desirable to allow for the fact that the test statistic W has a discrete distribution, whereas the normal distribution is continuous. This correction decreases the numerator and renders the outcome more conservative [12].

The U -test Formulation

The Mann–Whitney U -test has a completely different derivation, and is based on the set of pair-wise comparisons of x values to y values. There are $n \times m$ such pair-wise comparisons. For each comparison, the Mann–Whitney U statistic is incremented by 0 if $x_i > y_j$, by 1/2 if $x_i = y_j$, and by 1 if $x_i < y_j$. The resulting sum is related to the Wilcoxon W ,

$$W = U + \left(\frac{1}{2}\right)[n(n + 1)], \quad (4)$$

with the result that tests based on U are equivalent to tests based on W [10]. The U -test formulation allows a more natural handling of data that are only partially ordered [4].

Computation

Except where the sample sizes are small, the Wilcoxon–Mann–Whitney test usually is evaluated within a statistical computing package. Comparisons of the Wilcoxon–Mann–Whitney test in 11 commercial statistical packages is presented in [5]. Many packages offer corrections for ties and continuity as well as exact computation, although only SAS was found to have all three options in this study.

The Wilcoxon Test as a Linear Rank Test

The Wilcoxon test is a linear rank test. That is, the W statistic is a weighted sum of ranks. The regular spacing of the ranks, as integers, contributes to

tied values in the permutation distribution of the W statistic. So too does the use of a small number of mid ranks (averages of tied ranks) when calculating the statistic from ordered categorical data. This discreteness in the null distribution results in a conservative and less powerful test [8]. Eliminating or reducing the number of tied outcomes could result in a more powerful test [13]. Several approaches have been proposed for assigning different values to some permutations that would, otherwise, have equal W statistics [3, 8, 11]. One of these assures that even the exact permutation version of the test becomes uniformly more powerful [3].

Asymptotic Properties

The **asymptotic relative efficiency** of the Wilcoxon–Mann–Whitney test, against the t Test, makes it a strong candidate for testing for differences in location. Where the distributions sampled are, in fact, normal, the Wilcoxon–Mann–Whitney test has an asymptotic relative efficiency of 0.955. In no case is the asymptotic relative efficiency of the Wilcoxon test lower than 0.864 [6]. And, the Wilcoxon test can offer much greater efficiency than the t Test for some types of distributions. The worst-case potential loss of efficiency of $13\% - 0.136 = (1.000 - 0.864)$ – might be regarded as a relatively small insurance premium to be paid in case one of these distributions arises, and renders the t Test inefficient.

Example The following excerpt from [9] (see Table 1) shows the ranks of mens’ and womens’ scores after being tested regarding the strength of their endorsement of traditional sex ascriptions.

To test the null hypothesis that there is no difference in endorsement between genders against the alternative hypothesis that a difference exists, we compute the sums of the ranks for each gender.

To compute the large-sample approximation we note that

The sampling mean is $17 \cdot (17 + 17 + 1)/2 = 297.5$. The sampling variance is $17 \cdot 17 \cdot (17 + 17 + 1)/12$, or a standard error of 29.033.

These result in a standardized test statistic of $W^* = (378 - 297.5)/29.033 = 2.77$.

We reject the null hypothesis at $\alpha = 0.05$ as $W^* \geq z_{\alpha/2} = 1.96$, and we see that a significant difference exists between the two groups. The P value of the normal approximation can be calculated as 0.0056.

References

- [1] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [2] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**(1), 74–82.
- [3] Berger, V. & Sackrowitz, H. (1997). Improving tests for superior treatment in contingency tables, *Journal of the American Statistical Association* **92**(438), 700–705.
- [4] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**, 48–55.
- [5] Bergmann, R., Ludbrook, J. & Spooren, W. (2000). Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages, *The American Statistician* **54**, 72–77.
- [6] Hodges, J.L. & Lehmann, E.L. (1956). The efficiency of some non-parametric competitors of the t -test, *Annals of Mathematical Statistics* **27**, 324–335.
- [7] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, John Wiley, New York.
- [8] Ivanova, A. & Berger, V.W. (2001). Drawbacks to integer scoring for ordered categorical data, *Biometrics* **57**, 567–570.
- [9] Kando, T.M. (1972). Role strain: a comparison of males, females, and transsexuals, *Journal of Marriage and the Family* **34**, 459–464.
- [10] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.
- [11] Permutt, T. & Berger, V.W. (2000). A new look at rank tests in ordered 2xk contingency tables, *Communications in Statistics – Theory and Methods* **29**(5 and 6), 989–1003.

Table 1 Ranks of scores for men and women

	Ranks	Sum
Men	1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16.5, 18, 19, 27.5, 34	217
Women	2, 3, 13, 16.5, 20.5, 20.5, 22, 23, 24, 25, 26, 27.5, 29, 30, 31, 32, 33	378

4 Wilcoxon–Mann–Whitney Test

- [12] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw-Hill, New York.
- [13] Streitberg, B. & Roehmel, J. (1990). On tests that are uniformly more powerful than the Wilcoxon-Mann-Whitney test, *Biometrics* **46**, 481–484.
- [14] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(See also **Hodges–Lehman Estimator; Kruskal–Wallis Test; Randomized Block Design: Nonparametric Analyses; Wilcoxon, Frank**)

VENITA DEPUY, VANCE W. BERGER AND
YANYAN ZHOU

Winsorized Robust Measures

RAND R. WILCOX

Volume 4, pp. 2121–2122

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Winsorized Robust Measures

Consider any n observations and let g be $0.1n$, rounded down to the nearest integer. Then, 10% trimming refers to removing the g smallest, as well as the g largest values (see **Trimmed Means**). Winsorizing the values means that the g smallest values are reset to the smallest value not trimmed, and the g largest are set to the largest value not trimmed. As an illustration, consider the eleven values 6, 2, 10, 14, 9, 8, 22, 15, 13, 82, and 11. Then $g = 1$ and Winsorizing these values by 10% refers to replacing the smallest value, 2, with the next smallest, 6. Simultaneously, the largest value, 82, is replaced by the next largest, 22. So the 10% Winsorized values are 6, 6, 10, 14, 9, 8, 22, 15, 13, 22, and 11. The 20% Winsorized values are obtained in a similar fashion; only, now g is $0.2n$, rounded down to the nearest integer. The average of the Winsorized values is called a *Winsorized mean*, and the variance of the Winsorized values is called a *Winsorized variance*.

Winsorized means can be used to compare groups; under nonnormality, a Winsorized mean can have a substantially lower standard error than the usual sample mean, which can result in higher **power**. But, usually other robust estimators are used. Instead,

Winsorization is typically used to obtain a theoretically correct estimate of the standard error of a trimmed mean, which has certain practical advantages over comparing groups with a Winsorized mean. For details, see [1–5]. Winsorization also plays a role when searching for robust alternatives to Pearson's correlation (see **Pearson Product Moment Correlation**) [4, 5]. The so-called Winsorized correlation guards against outliers among the marginal distributions, which can help detect associations that would be missed when using Pearson's correlation. A criticism, however, is that the Winsorized correlation does not take into account the overall structure of the data when dealing with **outliers**. For example, only two unusual values can mask an association that would be detected by other correlation coefficients [5].

References

- [1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.
- [2] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.
- [3] Staudte, R.C. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.
- [4] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.
- [5] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

RAND R. WILCOX

Within Case Designs: Distribution Free Methods

LISA M. LIX AND H.J. KESELMAN

Volume 4, pp. 2122–2126

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Within Case Designs: Distribution Free Methods

In a within-case or repeated measures (RM) design (see **Repeated Measures Analysis of Variance**), subjects provide data at K successive points in time or for each of K experimental conditions. Data collected from different subjects are assumed to be independent, while data from the same subject are correlated. Tests of within-case main or **interaction effects** may be conducted using univariate or multivariate parametric or nonparametric procedures; the valid use of any one of these approaches depends on the data conforming to its underlying derivational assumptions.

The **analysis of variance (ANOVA)** F test, the usual parametric test for RM designs, rests on the assumption of **sphericity**,

$$\mathbf{C}\Sigma\mathbf{C}' = \sigma^2\mathbf{I}, \quad (1)$$

where $\mathbf{C}$, of dimension $(K - 1) \times K$, defines a set of orthonormalized contrasts on the repeated measurements, Σ is the covariance matrix, and $\mathbf{I}$ is an identity matrix of dimension $(K - 1)$. The multivariate parametric approach makes no assumptions about the structure of the covariance matrix of the repeated measurements. Both univariate and multivariate parametric procedures assume a K -variate normal distribution.

Type I **error rates** of parametric procedures for testing within-case effects are relatively robust (i.e., insensitive) to departures from normality, although skewed distributions may be associated with inflated error rates [12]. However, nonnormality can result in a substantial loss of statistical **power** to detect the presence of within-case effects [20]. When the data are highly skewed or have heavy tails due to the presence of outliers, nonparametric procedures, which make no assumptions regarding the distribution of the data, may result in more powerful tests of within-case effects. In this section, we select a number of different nonparametric procedures that may be applied to RM data, and illustrate their application.

Procedures Based on Rank Scores

Rank tests for within-case designs include procedures that are applied to inter-case ranks, where all NK

scores are ranked without regard to subject membership. They also include procedures that are applied to intra-case ranks, where the scores for each of the N subjects are arranged in increasing order of magnitude, and then ranked from 1 to K . (see **Rank Based Inference**.)

For the simplest within-case design that contains a single group of subjects and a single within-case factor, let y_{ik} represent the score for the i th subject ($i = 1, \dots, N$) for the k th treatment or time period ($k = 1, \dots, K$), and let r_{ik} represent the intra-case rank of y_{ik} . Midranks are assigned for ties. **Friedman's test** [7], a well-known procedure that has been used to test within-case effects with intra-case ranks, is defined as

$$FR = \frac{K(K+1)}{12N} \sum_{k=1}^K \left(\bar{r}_{.k} - \frac{K+1}{2} \right)^2, \quad (2)$$

where $\bar{r}_{.k}$ is the k th treatment mean rank. FR is asymptotically distributed as $\chi^2[\alpha; K - 1]$ (see [10] for approximations to the distribution of FR).

Friedman's procedure [7] tests the exchangeability hypothesis, $G_i(\mathbf{y}) = G(y_1 \dots y_K)$ for $\mathbf{y}_i = [y_{i1} \dots y_{iK}]$ and any permutation of $[1 \dots K]$, where $G_i(\mathbf{y})$ denotes the distribution function of $\mathbf{y}_i$. In other words, under the null hypothesis, the joint distribution of the observations is assumed to be invariant for any permutation of the intra-subject ranks. Accordingly, Friedman's test assumes a common correlation between pairs of observations [1]. Thus, while Friedman's procedure may be insensitive to the shape of the underlying distribution, it is known to be sensitive to departures from sphericity.

For designs that contain both within-case and between-case (i.e., grouping) factors, procedures based on intra-case ranks include extensions of Friedman's test [4], Hollander and Sethuraman's [8] two-group test, and extensions of Hollander and Sethuraman's test to multi-group designs [4, 17]. These procedures are used to test within-case interactions, which can be expressed as tests of discordance or inconsistency of intra-case ranks across independent groups of subjects.

Procedures based on inter-case ranks include rank transform tests, in which standard parametric tests, such as the analysis of variance (ANOVA) F test or a multivariate test, are applied to ranked data. The ANOVA F , which tests the exchangeability

2 Within Case Designs: Distribution Free Methods

hypothesis, is given by

$$F_{RT} = \frac{MS_K}{MS_{S \times K}} = \frac{\sum_{k=1}^K (\bar{r}_{.k}^* - \bar{r}_{..}^*)^2}{\sum_{k=1}^K \sum_{i=1}^N (r_{ik}^* - \bar{r}_{i.}^* - \bar{r}_{.k}^* + \bar{r}_{..}^*)^2}, \quad (3)$$

where r_{ik}^* , $\bar{r}_{i.}^*$, $\bar{r}_{.k}^*$, and $\bar{r}_{..}^*$ respectively represent the rank for the i th subject at the k th level of the within-case factor, the mean rank for the i th subject, the mean rank for the k th within-case factor level, and the grand mean rank. F_{RT} is approximately distributed as $F[\alpha; K-1, (N-1)(K-1)]$. For one-group within-case designs, the multivariate rank transform test is Hotelling's T^2 [9], $T_{RT} = N(\mathbf{CR})(\mathbf{CS}_r\mathbf{C})^{-1}(\mathbf{CR})$, where $\mathbf{C}$ defines a set of $(K-1)$ contrasts for the repeated measurements, $\bar{\mathbf{R}} = [\bar{r}_{.1}^* \dots \bar{r}_{.K}^*]'$, and $\mathbf{S}_r$ is the covariance matrix of the ranks. The statistic $F_T = [(N/(K-1))T_{RT}/(N-1)]$ is approximately distributed as $F[\alpha; K-1, N-K+1]$ [1]. Hotelling's T^2 for rank transform data is used to test the hypothesis of equality of the marginal distributions of the repeated measurements, that is, $G_1(\mathbf{y}) = G_2(\mathbf{y}) = \dots = G_K(\mathbf{y})$.

Rank transform tests are appealing to researchers because they can be easily applied with standard statistical software package. One limitation is that they cannot be applied to tests of within-case interactions. The ranks are not a linear function of the original observations, therefore, ranking the data may introduce additional effects into the statistical model. Moreover, ranking may alter the pattern of the correlations among repeated measurements. Accordingly, rank transform tests, while insensitive to departures from normality, must be used with caution in multi-factor designs [1, 2, 3, 6, 18].

Nonparametric Procedures Based on Resampling

When the assumption of multivariate normality is in doubt (see **Multivariate Normality Tests**), within-case effects may be tested using statistical procedures based on the resampling technique of bootstrapping (see **Bootstrap Inference; Permutation Based Inference**) [13, 19]. Under this

approach, the usual univariate or multivariate parametric test statistic is computed on the original data, but statistical significance of within-case effects is assessed using the empirical distribution of the test statistic rather than the theoretical distribution.

To illustrate, let F denote the conventional ANOVA F statistic for testing the within-case effect in a single group design. An iterative process is used to obtain the empirical distribution of the test statistic as follows: A bootstrap data set is generated by randomly sampling with replacement the K -variate vectors of repeated measurements. Let $\mathbf{y}_i^*$ represent the i th resampled vector. Each $\mathbf{y}_i^*$ is centered on the sample mean vector, $\bar{\mathbf{y}} = [\bar{y}_{.1} \dots \bar{y}_{.K}]^T$, where $\bar{y}_{.k} = \sum_{i=1}^n y_{ik}$, so that $\mathbf{y}_i^{*C} = \mathbf{y}_i^* - \bar{\mathbf{y}}$. The test statistic, F^* , is computed on the centered bootstrapped data set. This process is repeated B times. Let $F_{(1)}^* \leq F_{(2)}^* \leq \dots \leq F_{(B)}^*$ denote the B bootstrapped test statistics arranged in ascending order, and let $m = (1 - \alpha)B$. Then, F is referred to the critical value $F_{(m)}^*$. The bootstrapped ANOVA F Test will control the rate of Type I errors to α under departures from both normality and sphericity [5]. The bootstrapped Hotelling's T^2 also performs well under departures from normality.

A Numeric Example

To illustrate these various nonparametric tests based on ranks and the bootstrap, we selected a data set ([14], p. 571) for an experiment in which the length of gaze (in seconds) at a stimuli was obtained for each of 14 infants (see Table 1). Four different stimuli were considered: face, concentric circles, newspaper, and unpatterned white circle. We modified the original data by adding a constant to each of the measurements for the first two subjects in order to produce a skewed distribution. The inter-case ranks for the modified data set are in Table 2. Table 3 contains $\mathbf{S}$ and $\mathbf{S}_r$, the covariance matrix of the raw scores and the ranks, respectively. Both covariance matrices reveal the presence of increasing heterogeneity in the data across the four stimuli. Friedman's test gives $FR = 6.8$ with a P value of $p_{FR} = .08$. Applying the rank transform Hotelling's T^2 to the data gives $T_{RT} = 17.0$ with $p_{RT} < .0001$. The bootstrap P value for these

Table 1 Raw scores for one-group within-case design

Infant	Stimuli			
	Face	Circle	Newspaper	White
1	6.1	6.4	6.7	6.8
2	6.3	6.6	6.7	6.5
3	2.1	1.7	1.2	0.7
4	1.5	0.9	0.6	0.4
5	0.9	0.6	0.9	0.8
6	1.6	1.8	0.6	0.8
7	1.8	1.4	0.8	0.6
8	1.4	1.2	0.7	0.5
9	2.7	2.3	1.2	1.1
10	1.5	1.2	0.7	0.6
11	1.4	0.9	1.0	0.5
12	1.6	1.5	0.9	1.0
13	1.3	1.5	1.4	1.6
14	1.3	0.9	1.2	1.4

Table 2 Inter-case ranks for one-group within-case design

Infant	Stimuli			
	Face	Circle	Newspaper	White
1	49	51	54.5	56
2	50	53	54.5	52
3	46	43	26	10
4	37.5	17.5	6	1
5	17.5	6	17.5	13
6	41	44.5	6	13
7	44.5	33	13	6
8	33	26	10	2.5
9	48	47	26	23
10	37.5	26	10	6
11	33	17.5	21.5	2.5
12	41	37.5	17.5	21.5
13	29.5	37.5	33	41
14	29.5	17.5	26	33

Table 3 Variance-covariance matrix for raw scores and inter-case ranks

Raw scores				
2.98	3.30	3.52	3.53	
	3.72	3.96	4.01	
		4.44	4.49	
			4.58	
Inter-case ranks				
84.8	113.7	56.6	51.4	
	210.7	128.0	155.7	
		245.4	262.9	
			346.5	

data, p_B , which was based on 1000 replications (see [19]), is also $<.0001$. Applying the ANOVA F

test to the original observations gives $F = 8.0$ with $p = .0003$. The bootstrap P value for this test is also $<.0001$.

Concluding Remarks

Behavioral scientists may be reluctant to bypass conventional parametric approaches for the analysis of within-case effects in favor of nonparametric tests based on ranking or resampling methods. This reluctance may stem, in part, from the belief that parametric procedures are robust to departures from normality. While Type I error rates of parametric procedures may be relatively robust to the presence of nonnormal distributions, power rates can be substantially affected, particularly when the data are skewed. Researchers may also be reluctant to adopt nonparametric procedures because they are unfamiliar with test for multi-factor and multivariate designs, or with methods for testing linear contrasts on the within-case effects. Recent research has focused on the development of nonparametric tests for a variety of univariate and multivariate repeated measures designs [15, 16]. Procedures based on the bootstrap can be readily applied to a variety of complex univariate and multivariate designs to test hypotheses on omnibus effects as well as linear contrasts of within-case effects [19].

Finally, we note that alternative parametric procedures have been proposed for testing within-case effects when the data are nonnormal. For designs that contain both within-case and between-case factors, Keselman, Kowalchuk, Algina, Lix, and Wilcox [11] examined approximate degrees of freedom procedures that assume neither equality (i.e., homogeneity) of group covariances nor sphericity of the common covariance of the repeated measurements. These procedures were extended to the case of nonnormality by substituting the usual (i.e., least-squares) estimators with robust estimators based on **trimmed means**. Trimmed means are obtained by removing the most extreme observations from the tails of the data distribution prior to computing the average score. These approximate degrees of freedom tests based on trimmed estimators were shown to be insensitive to the presence of both skewed and heavy-tailed distributions. The tests were also examined when critical values were generated via the bootstrap. As expected, the bootstrapped tests were

also robust to nonnormality, although the Type I error rates of the two approaches were not appreciably different.

References

- [1] Agresti, A. & Pendergast, J. (1986). Computing mean ranks for repeated measures data, *Communications in Statistics, Theory and Methods* **15**, 1417–1433.
- [2] Akritas, M.G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part I, *Journal of the American Statistical Association* **86**, 457–460.
- [3] Akritas, M.G. (1993). Limitations of the rank transform procedure: a study of repeated measures designs, part II, *Statistics & Probability Letters* **17**, 149–156.
- [4] Beasley, T.M. (2000). Non-parametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs, *Journal of Educational and Behavioral Statistics* **25**, 20–59.
- [5] Berkovits, I., Hancock, G.R. & Nevitt, J. (2000). Bootstrapping resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations, *Educational and Psychological Measurement* **60**, 877–892.
- [6] Blair, R.C., Sawilowsky, S.S. & Higgins, J.J. (1987). Limitations of the rank transform statistic in test for interactions, *Communications in Statistics: Simulation and Computation* **16**, 1133–1145.
- [7] Friedman, M. (1937). The used of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.
- [8] Hollander, M. & Sethuraman, J. (1978). Testing for agreement between two groups of judges, *Biometrika* **65**, 403–411.
- [9] Hotelling, H. (1931). The generalization of student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.
- [10] Inman, R.L. & Davenport, J.M. (1980). Approximations of the critical region of the Friedman statistic, *Communications in Statistics, Theory and Methods* **A9**, 571–595.
- [11] Keselman, H.J., Kowalchuk, R.K., Algina, J., Lix, L.M. & Wilcox, R.R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping, *British Journal of Mathematical and Statistical Psychology* **53**, 175–191.
- [12] Keselman, J.C., Lix, L.M. & Keselman, H.J. (1996). The analysis of repeated measurements: A quantitative research synthesis, *British Journal of Mathematical and Statistical Psychology* **49**, 275–298.
- [13] Lunnenborg, C.E. & Tousignant, J.P. (1985). Efron's bootstrapping with application to the repeated measures design, *Multivariate Behavioral Research* **20**, 161–178.
- [14] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing data: A Model Comparison Perspective*, 2nd Edition, Lawrence Erlbaum, Mahwah.
- [15] Munzel, U. & Brunner, E. (2000). Nonparametric methods in multivariate factorial designs, *Journal of Statistical Planning and Inference* **88**, 117–132.
- [16] Rahman, M.M. & Islam, M.N. (1998). Nonparametric analysis of multifactor repeated measures experiments, *Journal of the Indian Society of Agricultural Statistics* **51**, 81–93.
- [17] Rasmussen, J.L. (1989). Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity, *British Journal of Mathematical and Statistical Psychology* **42**, 91–102.
- [18] Sawilowsky, S.S., Blair, R.C. & Higgins, J.J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA, *Journal of Educational Statistics* **14**, 255–267.
- [19] Wasserman, S. & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology, *Psychophysiology* **26**, 208–221.
- [20] Wilcox, R.R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size?, *Review of Educational Research* **65**, 51–77.

LISA M. LIX AND H.J. KESELMAN

Yates' Correction

CATALINA STEFANESCU, VANCE W. BERGER AND SCOTT L. HERSHBERGER

Volume 4, pp. 2127–2129

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Yates' Correction

Yates' correction [15] is used as an approximation in the analysis of 2×1 and 2×2 **contingency tables**. A 2×2 contingency table shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables, in a sample of size N . A schematic form of such a table is shown in Table 1.

A research question of interest is often whether the variables summarized in a contingency table are independent of each other. The test to determine if this is so depends on which, if any, of the margins are fixed, either by design or for the purposes of the analysis. For example, in a randomized trial in which the number of subjects to be randomized to each treatment group has been specified, the row margins would be fixed but the column margins would not (it is customary to use rows for treatments and columns for outcomes). In a matched study, however, in which one might sample 100 cases (smokers, say) and 1000 controls (non-smokers), and then test each of these 1100 subjects for the presence or absence of some exposure that may have predicted their own smoking status (perhaps a parent who smoked), it would be the column margins that are fixed. In a random and unstratified sample, in which each subject sampled is then cross-classified by two attributes (say smoking status and gender), neither margin would be fixed. Finally, in **Fisher's** famous tea-tasting experiment [13], in which a lady was to guess whether the milk or the tea infusion was first added to the cup by dividing eight cups into two sets of four, both the row and the column margins would be fixed by the design. Yet, in the first case mentioned, that of a randomized trial with fixed row margins but not fixed column margins, the column margins may be treated as fixed for the purposes of the analysis, so as to ensure exactness [2].

When the row and column margins are fixed, either by design or for the analysis, independence

can be tested using Fisher's exact test [4] (*see* **Exact Methods for Categorical Data**). This test is based on the hypergeometric distribution (*see* **Catalogue of Probability Density Functions**), and it is computationally intensive, especially in large samples. Therefore, Fisher advocated the use of Pearson's statistic,

$$X^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}, \quad (1)$$

which, under the null hypothesis, has a χ^2 distribution with one degree of freedom. Yates [15] argued that the χ^2_1 distribution gives only approximate estimates of the discrete probabilities associated with frequency data, and, thus, the P values based on Pearson's X^2 statistic will generally underestimate the true P values. In general, when a statistic takes discrete values $a < b < c$, the P value corresponding to b is estimated by the tail of the continuous function defined by the point $(a + b)/2$. Therefore, the tail of the continuous function computed at b will underestimate the P value. In this context, Yates suggested that X^2 should be corrected for continuity and proposed the corrected test statistic

$$\frac{N(|AD - BC| - \frac{1}{2}N)^2}{(A + B)(C + D)(A + C)(B + D)}.$$

Although Yates' correction is best known for its use in the analysis of 2×2 contingency tables, it is also applicable to the analysis of 2×1 contingency tables. A 2×1 contingency table displays the frequencies of occurrence of two categories in a random sample of size N , drawn from a population in which the proportions of cases within the two categories are p and $1 - p$. The research question is usually whether the observed numbers of cases x and $N - x$ in the two categories have been sampled from a population with some prespecified value of p . This can be tested using Pearson's statistic,

$$X^2 = \frac{(x - Np)^2}{Np(1 - p)}, \quad (2)$$

which asymptotically has a χ^2_1 distribution under the null hypothesis. Yates showed that, in this case as well, the use of Pearson's X^2 results in P values that systematically underestimate the true P values based on the binomial distribution. Therefore, he suggested the corrected statistic

$$\frac{(|x - Np| - \frac{1}{2})^2}{Np(1 - p)}. \quad (3)$$

Table 1 A 2×2 contingency table

Row variable	Column variable		Totals
	1	2	
1	A	B	A + B
2	C	D	C + D
Totals	A + C	B + D	N

Kendall and Stuart [7] remarked that Yates' procedure is a special case of a general concept of a continuity correction, while **Pearson** [10] noted that Yates' correction derives naturally from the Euler-Maclaurin theorem used to approximate binomial and hypergeometric distributions. Subsequently, the use of Yates' correction to Pearson's X^2 has been widely emphasized for the analysis of contingency tables [14]. There are, however, several issues related to Yates' correction, and we shall discuss some of these in turn.

Firstly, in the analysis of 2×1 contingency tables, the P values associated with the corrected statistic (3) tend to overestimate the true P values in the tails of the distribution and to underestimate them towards the center. This is illustrated in Table 2, which displays the two-tailed P values in a contingency table with $N = 10$ and $p = 0.5$, obtained with Pearson's X^2 statistic and Yates' correction. The table reports as well the true binomial P values, which are the gold standard. It should also be noted [15] that the P values obtained with the continuity correction are much less accurate when the binomial probability p is substantially different from 0.5.

Secondly, Yates' correction is appropriate only for one-sided tests, as it is based on a comparison between the observed contingency and the next strongest contingency in the same direction ([6, 8]). For two-sided tests, the statistic involves an overcorrection. Along the same lines, it can be proven analytically that Yates' correction is systematically conservative when carrying out two-sided tests [9].

Thirdly, a more important issue related to Yates' correction is its applicability to the analysis of contingency tables arising from different research designs. Many researchers have argued that Yates's correction is based upon comparisons among contingency

tables with fixed row and column marginal totals, particularly since Yates is specifically concerned with approximating the hypergeometric distribution from Fisher's exact test. However, Yates' method has also been recommended for the analysis of 2×2 contingency tables arising from sampling schemes, where one or both sets of marginal totals are free to vary, and are, thus, subject to sampling errors. It should be noted that such sampling schemes are the ones most frequently found in actual research context. While Yates [16] argues along the lines of Fisher's reasoning that the analysis of 2×2 contingency tables should always be performed conditional on the observed marginal totals, this approach is still subject to debate [12]. On the other hand, when the marginal totals are not fixed, Yates' procedure involves an additional overcorrection, and the test statistic is conservative. This has been investigated through Monte Carlo simulations ([5, 11]), and confirmed analytically ([3, 6]). In particular, Grizzle [5] notes that for contingency tables with nonfixed marginal totals, Yates's procedure 'produces a test that is so conservative as to be almost useless'.

Finally, Yates's correction originated as a device of eliminating the discrepancies that arose when approximating the hypergeometric distribution in Fisher's exact test. The approximation using Pearson's X^2 was necessary 'for the comparative simplicity of the calculations' ([4], p. 99), because the exact analysis of 2×2 contingency tables with the limited computing power available at the time was prohibitive in many cases. This is no longer the case today. Indeed, Agresti [1] notes that Yates' correction is not necessary anymore since current software makes Fisher's exact test computationally feasible even when the sample sizes are large.

References

- [1] Agresti, A. (2002). *Categorical Data Analysis*, Wiley, New York.
- [2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.
- [3] Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables, *Journal of the American Statistical Association* **69**, 374–376.
- [4] Fisher, R.A. (1934). *Statistical Methods for Research Workers*, 5th Edition, Oliver and Boyd, Edinburgh.
- [5] Grizzle, J.E. (1967). Continuity correction in the χ^2 test for 2×2 tables, *The American Statistician* **21**, 28–32.

Table 2 Binomial distribution for $N = 10$ and $p = 0.5$, and two-tailed P values. (Adapted from [12])

x	$p(x)$	P values		
		Pearson	Yates	Binomial
0, 10	0.0010	0.0016	0.0044	0.0020
1, 9	0.0098	0.0114	0.0268	0.0215
2, 8	0.0439	0.0580	0.1138	0.1094
3, 7	0.1172	0.2060	0.3428	0.3437
4, 6	0.2051	0.5270	0.7518	0.7539
5, 5	0.2461	1.0000	1.0000	1.0000

-
- [6] Haber, M. (1982). The continuity correction and statistical testing, *International Statistical Review* **50**, 135–144.
 - [7] Kendall, M.G. & Stuart, A. (1967). *The Advanced Theory of Statistics*, Vol. 2, 2nd Edition, Griffin, London.
 - [8] Mantel, N. (1976). The continuity correction, *The American Statistician* **30**, 103–104.
 - [9] Maxwell, E.A. (1976). Analysis of contingency tables and further reasons for not using Yates correction in 2×2 tables, *Canadian Journal of Statistics* **4**, 277–290.
 - [10] Pearson, E.S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table, *Biometrika* **34**, 139–167.
 - [11] Richardson, J.T.E. (1990). Variants of chi-square for 2×2 contingency tables, *British Journal of Mathematical and Statistical Psychology* **43**, 309–326.
 - [12] Richardson, J.T.E. (1994). The analysis of 2×1 and 2×2 contingency tables: an historical review, *Statistical Methods in Medical Research* **3**, 107–133.
 - [13] Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, Owl Books, London.
 - [14] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill, New York.
 - [15] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* **1**(Suppl. 1), 217–235.
 - [16] Yates, F. (1984). Tests of significance for 2×2 contingency tables, *Journal of the Royal Statistical Society Series A* **147**, 426–463.

Further Reading

- Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on 2×2 tables, *Journal of the American Statistical Association* **75**, 510–515.
- Plackett, R.L. (1964). The continuity correction in 2×2 tables, *Biometrika* **51**, 327–337.

CATALINA STEFANESCU, VANCE W. BERGER
AND SCOTT L. HERSHBERGER

Yates, Frank

BRIAN S. EVERITT

Volume 4, pp. 2129–2130

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Yates, Frank

Born: 12 May, 1902 in Manchester, England.

Died: 17 June, 1994 in Harpenden, England.

Born in Manchester, Yates read mathematics at St. John's College, Cambridge and received a first-class honors degree in 1924. After a period working in the Gold Coast (now Ghana) on a geodetic survey as a mathematical advisor, he obtained a post as assistant statistician at Rothamsted Experimental Station in 1931, where he worked under **R.A. Fisher**. Two years later when Fisher left to take up a chair at University College, London, Yates became Head of Statistics at Rothamsted, a post he held until his retirement in 1968.

Yates continued the work on design of experiment, replication, **randomization** and blocking (*see* **Block Random Assignment**) (to reduce error), topics introduced to Rothamsted by Fisher (*see* [2] for a selection of work in this area). All these ideas were originally applied to agriculture but spread rapidly to many other disciplines. Yates extended and clarified the ideas of orthogonality, confounding, and balance, and suggested the use of split-plot designs. During World War II, Yates studied food supplies and applications of fertilizers to improve crops, and applied experimental design techniques to a wide range of problems such as control of pests. But despite all his important contributions to designing studies, Yates is most widely remembered for his continuity correction in **contingency tables** (**Yates' correction**); ironically

this correction has been made almost obsolete by the development of software for applying exact methods [1].

Yates was quick to realize the possibilities for statistics and statisticians provided by the development of electronic computers in the 1950s. And in 1954, the first British computer equipped with effective magnetic storage, the Elliot 401, was installed at Rothamsted. Using only machine code, Yates and other members of the statistics department produced programs both for the **analysis of variance** and to analyze survey data. Yates helped establish the British Computer Society, of which he was made President in 1960–1961. In 1948, Yates was made a Fellow of the Royal Society, and in 1960 he was awarded the Royal Statistical Society's Guy Medal in Gold. In 1963, he was awarded the CBE.

Despite retiring from Rothamsted in 1968, Yates kept a room there and never lost touch with agriculture. Just before his death in Harpenden in 1994, Yates completed, in 1993, 60 years of work at Rothamsted.

References

- [1] Mehta, C.R. & Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *Journal of the American Statistical Society* **78**, 427–434.
- [2] Yates, F. (1970). *Experimental Design: Selected Papers of Frank Yates*, Griffin Publishing, London.

BRIAN S. EVERITT

Yule, George Udny

MICHAEL COWLES

Volume 4, pp. 2130–2131

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

Yule, George Udny

Born: February 18, 1871, in Morham, Scotland.

Died: June 26, 1951, in Cambridge, England.

Scottish by birth, Yule was educated in England and spent all his working life there. After school, the young Yule studied engineering at University College, London. Engineering and physics never captured his interest though he was greatly influenced by his Professor of Applied Mathematics, **Karl Pearson**. He attended Pearson's Gresham Lectures in 1891 and corresponded with him. He spent a year in Bonn studying and researching under Hertz. His work there produced his first published papers. On his return to England, Pearson offered him the post of Assistant Professor. Though poorly paid, it provided Yule with the impetus and experience that led to his adopting the study and teaching of statistics as his lifelong career, apart from a spell as a civil servant during World War I.

Yule was Pearson's workhorse and he often lectured on his behalf when Pearson was indisposed. Their personal interactions were at that time cordial. They worked closely and took holidays together. The exigencies of his financial situation forced Yule to apply for other employment and in 1899 he became Secretary to the Examining Board of the Department of Technology of the London City and Guilds Institute. He married in the same year, a union that proved to be unhappy and led to a separation. Yule's tenure of the Newmarch Lectureship in Statistics at University College from 1902 to 1909 produced the material that became his famous textbook, *An Introduction to the Theory of Statistics*, the earliest of the major texts, running to 14 editions during Yule's lifetime, the final 4 editions being written in collaboration with **Maurice Kendall** [3].

In 1912, the University of Cambridge offered Yule a newly established lectureship in statistics and later he became Reader. He took up residence at St John's College for the rest of his life until his health forced him into a nursing home.

Yule's contributions to statistics consist largely of his clarification, interpretation, and expansion

of Pearson's work and the laying down of the groundwork for the future contributions of others. He demonstrated the least squares approach to the method of correlation and regression and showed that Pearson's coefficient and the regression model can be derived without assuming a bivariate normal distribution of the variables of interest. This not only indicated that r could be used as a descriptive statistic but gave us a readily understandable theoretical approach that led to the same outcome as Pearson's original **maximum likelihood method** [1]. He worked on measures of association in 2×2 tables and devised mathematical procedures that reduced the sometimes dense Pearsonian algebra of partial correlation [2]. His approaches became accepted by those who were interested in the theoretical bases of the methods.

Yule was elected a Fellow of the Royal Statistical Society in 1895 and remained so for almost 60 years. He was its Honorary Secretary for 12 years and the Society awarded him its highest honor, the Guy Medal in gold. He became a Fellow of the Royal Society in 1922.

In later life, he turned his considerable powers to the study of literary vocabulary, showing how statistical techniques might be used to determine authorship and to compare authors.

Yule was a determined, cheerful, and kindly man, articulate and well-read. He had a gift, as his correspondence with his great friend Major Greenwood shows, for irony and parody. In later years he regretted the transition from what he termed 'Karlovingian' (Pearson) statistics to the 'piscatorial' approach of Fisher and felt that he himself had little more to offer, though he was gratified to see his own contributions recognized. Even today, there are teachers and practitioners of statistics who use his text and benefit from it.

References

- [1] Yule, G.U. (1897). On the theory of correlation, *Journal of the Royal Statistical Society* **60**, 812–854.
- [2] Yule, G.U. (1900). On the association of attributes in statistics, *Philosophical Transactions, A* **194**, 257–319.
- [3] Yule, G.U. (1911 to 1973. Later editions with Maurice Kendall). *An Introduction to the Theory of Statistics*, Griffin Publishing, London.

MICHAEL COWLES

z Scores

DAVID CLARK-CARTER

Volume 4, pp. 2131–2132

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9

ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

© John Wiley & Sons, Ltd, Chichester, 2005

z Scores

A *z*-score is a form of standardized score. That is, it is a linear transformation of a raw score using the mean and the standard deviation (SD) of the sample or, if known, the population mean and SD. It is defined as

$$z = \frac{\text{score} - \text{mean}}{\text{standard deviation}}. \quad (1)$$

A set of scores that has been transformed to *z*-scores has a mean of 0 and an SD of 1. Note that a sample statistic, such as a mean, can also be transformed to a *z*-score using the mean and SD of the sampling distribution of the statistic.

Suppose that a sample of students had mathematics test scores with a mean of 100 and a SD of 10, then a person whose mathematics score was 120 would have a *z*-score of 2. Here, we would say that this person's score was two SDs above the mean. The advantage of standardizing scores is that it allows comparison across different tests. For example, if the same sample of students had also taken a language test with a mean of 50 and an SD of 4, and this same person had scored 51 on the test, then the standardized score would be 0.25. This tells us that, whereas

the student was two SDs above the mean in mathematics, he or she was only quarter of an SD above the mean for language.

Furthermore, when a sample comes from a normal distribution with a known mean and SD, we can use standard normal tables to find the **percentile** point for a given score. Our hypothetical person is in the top 2.28% for mathematics but only in the top 40.1% for language. The *z*-statistic employed in common hypothesis tests, for example, about means from normal populations is based on the *z*-score transformation.

z-scores can also be used to identify scores that could be **outliers**. An often quoted value for what might constitute an outlier is when the absolute (unsigned) value of the *z*-score is greater than or equal to 3 [1].

Reference

- [1] Lovie, P. (1986). Identifying outliers, in *New Developments in Statistics for Psychology and the Social Sciences*, A.D. Lovie, ed., The British Psychological Society and Methuen, London, (pp 44–69).

DAVID CLARK-CARTER